ISSN 1407-5806

COMPUTER MODELLING AND NEW TECHNOLOGIES

Volume 17
No 4

2013

Transporta un sakaru institūts
(Transport and Telecommunication Institute)

Computer Modelling and New Technologies

Volume 17, No.4 – 2013

ISSN 1407-5806
ISSN 1407-5814
(On-line: www.tsi.lv)

Riga – 2013

Host Organization:
Transport and Telecommunication Institute

Supporting Organizations:
Latvian Transport Development and Education Association
Latvian Academy of Sciences
Latvian Operations Research Society

THE JOURNAL IS DESIGNED FOR PUBLISHING PAPERS CONCERNING THE FOLLOWING FIELDS OF RESEARCH:

- mathematical and computer modelling
- mathematical methods in natural and engineering sciences
- physical and technical sciences
- computer sciences and technologies
- material science, solid state physics and chemistry
- nanomaterials, nanodevices and nanotechnologies
- aviation and aerospace technologies
- electronics and telecommunication
- transport and logistics
- economics and management
- innovative education
- social sciences

In journal articles can be presented in English. All articles are reviewed.

EDITORIAL CORRESPONDENCE
Transporta un sakaru institūts (Transport and Telecommunication Institute)
Lomonosova iela 1, LV-1019, Riga, Latvia. Phone: (+371) 67100593. Fax: (+371) 67100535
E-mail: cm&nt@tsi.lv, www.tsi.lv

COMPUTER MODELLING AND NEW TECHNOLOGIES, 2013, Vol. 17, No.4
ISSN 1407-5806, **ISSN** 1407-5814 (on-line: www.tsi.lv)
Scientific and research journal of Transport and Telecommunication Institute (Riga, Latvia)
The journal is being published since 1996.

The papers published in Journal "Computer Modelling and New Technologies" are included in **INSPEC** (since 2010), **VINITI** (since 2011), **CAS Database**
www.theiet.org/resources/inspec/
http://www2.viniti.ru/
http://www.cas.org/

Computer Modelling and New Technologies, 2013, Volume 17, No.4

CONTENTS

Computer Modelling and New Technologies, 2013, Volume 17, No.4

Editors' Remarks

****

On the Grasshopper and Cricket

by John Keats

The poetry of earth is never dead:
When all the birds are faint with the hot sun,
And hide in cooling trees, a voice will run
From hedge to hedge about the new-mown mead;
That is the Grasshopper's - he takes the lead
In summer luxury,- he has never done
With his delights; for when tired out with fun
He rests at ease beneath some pleasant weed.
The poetry of earth is ceasing never:
On a lone winter evening, when the frost
Has wrought a silence, from the stove there shrills
The Cricket's song, in warmth increasing ever,
And seems to one in drowsiness half lost,
The Grasshopper's among some grassy hills.

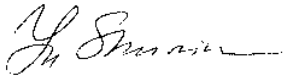
December 30, 1816

****

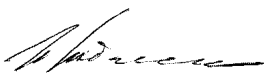
John Keats (1795-1821) [♣]

This 17th volume No.4 presents actual papers on main topics of Journal specialization, namely, Operation Research and Decision Making, Information and Computer Technologies and Mathematical and Computer Modelling. This is issue demonstrates the scope of research activities of scientific institutions in the People's Republic of China and reflects the spectrum of particular scientific problems in the field of computer modelling. Our journal policy is directed on the fundamental and applied sciences researches, which are the basement of a full-scale modelling in practice. This edition is the continuation of our publishing activities. We hope our journal will be interesting for research community, and we are open for collaboration both in research and publishing. We hope that Journal's contributors will consider the collaboration with the Editorial Board as useful and constructive.

EDITORS



Yu.N. Shunin



I.V. Kabashkin

[♣] **John Keats,** 31 October 1795 – 23 February 1821, was an English Romantic poet. He was one of the main figures of the second generation of Romantic poets along with Lord Byron and Percy Bysshe Shelley, despite his work only having been in publication for four years before his death. He had a significant influence on a diverse range of poets and writers. The poetry of Keats is characterised by sensual imagery, most notably in the series of odes. Today his poems and letters are some of the most popular and most analysed in English literature.

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 9-16
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

A NEW LEARNING STRUCTURE OF ADAPTIVE ALGORITHM FOR DIGITAL PREDISTORTION

Zh. Zeng[1], Q. Ma[2]

[1]Communication University of China, Beijing, 100024, China
[2] Beijing Institute of Information Technology, Beijing, 100094, China
[*]Corresponding author E-mail address: zhbzeng@cuc.edu.cn

Adaptive digital predistortion is a dominant technique for the linearization of power amplifier. The learning structure of adaptive algorithm is a major determinant of linearization performance. Indirect structure is widely used in due to its easy implementation and relatively good performance. In this paper, a new learning structure combined with both indirect and direct is proposed. Simulations show that this joint method can avoid feedback and system noise, hence achieve a significant improvement on linearization performance over indirect structure.

Keywords: Power Amplifier; Predistotion; Learning Structure; Linearization performance.

1. Introduction

It is well known that power amplifier (PA) is one of the key parts for wireless communication system. Signals, especially with high peak-to-average power ratio (PAPR), will always produce nonlinearity, thus causing spectral broadening as well as in-band distortion when passing through PA. Linearization technology, therefore, is critical to suppress spectral regrowth, especially in-band distortion.

Among all the linearization techniques, digital predistortion is one of the most cost effective due to its stability, adaptivity and ease of implementation. On the other hand, to adapt to the varying nonlinear characteristics of power amplifiers, adaptive algorithm is indispensable and plays the core role in digital predistortion system.

Adaptive algorithm is deployed to calculate coefficients of PA model with feedback signals from PA and original signals i.e. parameter estimation and then adjust the so-called predistorter's parameters to cancel the nonlinearity of PA. Least Mean Square (LMS) and Recursive Least Square (RLS) are conventional adaptive algorithm. Both of them can be used as adaptive algorithm and have the approximate performance. Therefore, the learning structure is a major factor that determines the linearization performance. There are two conventional learning structures of adaptive algorithm currently, named direct and indirect learning structure. The former structure has the advantage of simplity and anti-noise, but the classical adaptive learning algorithm cannot be used directly for that the power model is always unknown. Indirect learning structure is easy for implementation, while it is vulnerable to noise.

Therefore, this paper concentrates on the new learning structure of adaptive algorithm [1] that is based on these two learning structures. Simulations show that compared with the indirect learning structure, this proposed joint structure can not only reduce the noise from feedback path, but also achieve a desirable linearization performance.

2. Predistortion Theory

The nonlinear characteristics of PA can be presented by their amplitude modulation to amplitude modulation conversion (AM/AM) and amplitude modulation to phase modulation conversion (AM/PM).The predistorter, whose coefficients are adaptively derived from the adaptive algorithm, is in charge of producing anti-phase and anti-amplitude curves to the PA model, i.e. inverse to the AM/AM and AM/PM characteristics of PA, thus compensating the nonlinearity [2].

As seen in Figure 1, in a transmitting system, predistorter with characteristic function $F\left(|V_i|\right)$ is placed before the nonlinear power amplifier. The characteristic function $G\left(|V_d|\right)$ represents the nonlinear characteristics of power amplifier. Observe that both $F\left(|V_i|\right)$ and $G\left(|V_d|\right)$ depend only on the power (or

amplitude) of input signals, but not on their input phases. By appropriately choosing the coefficients of $F(|V_i|)$ through adaptive algorithm, making $F(|V_i|)$ and $G(|V_d|)$ mutually inverse, then an equitant linear amplification will be achieved as follows:

$$F(|V_i|)*G(|V_d|)=K, \tag{1}$$

where K is constant of system gain. The output signal $V_0=KV_i$ is then be linearly amplified when passing both predistorter and the power.

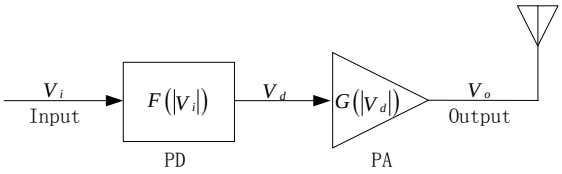


Figure 1. Simplified predistorter schematic

3. Predistortion Learning Structure

The architectures of adaptive algorithm for digital predistortion are divided into two types. One is the direct learning architecture [3], which must identify nonlinear characteristics of PA and then find the inverse characteristics of PA, which is of high computational complexity. Another type is the indirect learning architecture [4], which does not need to find inverse nonlinear characteristics of PA. Therefore, the indirect learning architecture is more attractive in the adaptive algorithm.

3.1. Direct Learning Structure

A direct adaptive predistortion block diagram is shown in Figure 2. In this structure, $x(n)$ and $y(n)$ represent input and output signal, respectively. At the same time, supposed the desired response of system is $d(n)$ and linear magnification equals to G. If $e(n)=d(n)-y(n)$ in the algorithm converges, i.e. $e(n)=0$, then the output of the amplifier is linear to input signal, and $y(n)=G\cdot x(n)$.

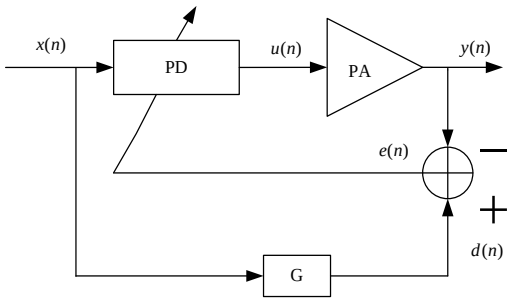


Figure 2. Direct learning structure of predistortion

Direct learning structure has the advantages of simple structure, after the algorithm, converging can achieve better predistortion effects, and predistortion coefficients are not affected without noise of amplifier nonlinear system. However, its drawback is cannot directly use the classical adaptive learning algorithm.

3.2. Indirect Learning Structure

If the coefficients of the predistortion are adaptive recursive estimation directly, the amplifier model is obtained first, so in order to avoid getting the power amplifier model, the indirect learning structure predistorter is used as shown in Figure 3. The benefits of indirect learning structure are we need not to identify the amplifier model first, predistorter model can be identified directly, and the structure is simple. G is the linear magnification of amplifier in Figure 3, the amplifier output $y(n)$ is the input to the predistorter training network after attenuated G times, the actual predistortion module is copy the

parameters of training network module. In the ideal case, $y(n) = G \cdot x(n)$ is desirable, which $z(n) = \hat{z}(n)$ is required, the target of estimation algorithm is to calculate the parameters of estimation module and then pass to the predistortion module.

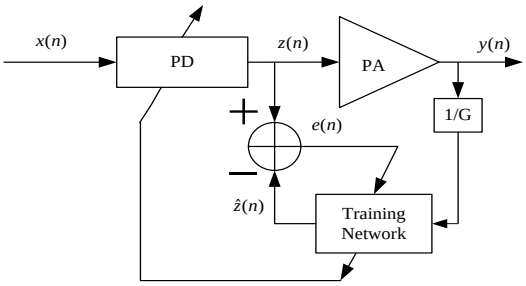


Figure 3. Indirect learning structure of predistortion

Indirect learning structure can use the classical adaptive learning algorithm and is easy to implement in engineering. However, its drawback is the addition noise of amplifier nonlinear system can make the parameters deviate from the optimum value [5], so that the predistortion linearization effects become poor.

3.3. New Predistortion Learning Structure

Indirect learning structure is simple and easy for implementation, but its drawback is susceptible to additive noise in transmitting system, while direct learning structure has an excellent performance of anti-noise. It is natural to suppose that if indirect structure is combined with direct, this new joint structure could contribute to improvement on linearization performance.

This new learning structure proposed is shown as Figure 4. Firstly, the switch *S* is set to the A-side, and then indirect learning structure of adaptive algorithm 1 in the dotted line frame will work. Since indirect learning structure is susceptible to additive noise, adaptive algorithm in this structure cannot achieve optimal parameters of PA model, although it is possible close to the best parameters. Since direct learning structure has the advantage of anti-noise, it can be used to amend the parameters based on indirect learning structure. When the adaptive algorithm 1 converges, switch *S* is set to the B-side and the algorithm 2 takes effect using convergent parameters of algorithm 2 as the initial value. Because the parameters derived from indirect learning structure is not within the optimal parameter area, the adaptive algorithm 2 can directly use the classical adaptive algorithm. The parameters will be more close to the optimal parameters of PA model through algorithm 2.

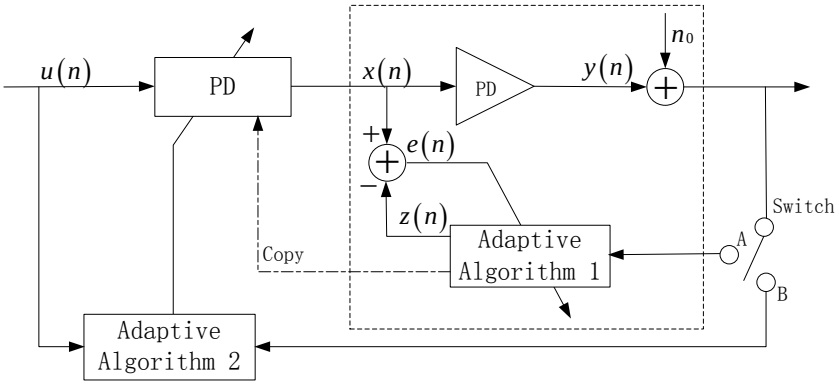


Figure 4. New predistortion learning structure

4. Adaptive Predistortion Algorithm

The adaptive algorithm is a key technology of predistortion. As for adaptive linearization technique, the core issue is how to get and automatic adjustment function parameters. The study of

adaptive part, this can be adaptive abstracting and using numerical analysis method to solve the problem in different ways to achieve the three elements of adaptive algorithm: the minimization algorithm, the objective function form and the error function, the final purpose is to make the objective function minimize which is based on the error function $e(n)$, in a manner such that the output signal approximates the expected output signal.

Matrix least square (LS) linear equation is given by as follows:

$$Z = UA, \tag{2}$$

where Z and A represents $[z(n), z(n-1), \ldots, z(n-L+1)]^T$ and $[a_{10}, \ldots, a_{(2k+1)0}, a_{11}, \ldots, a_{(2k+1)Q}]^T$ respectively.

For memory polynomial model, U is a coefficient matrix A with corresponding equations as follows:

$$\begin{aligned} U &= (Y_{10}, \ldots, Y_{(2k+1)0}, Y_{11}, \ldots, Y_{(2k+1)Q}) \\ Y_{ij} &= [r_{ij}(n), r_{ij}(n-1), \ldots, r_{ij}(n-L+1)]^T . \\ r_{ij} &= y(n-j) \cdot |y(n-j)|^{2(i-1)} \end{aligned} \tag{3}$$

The solution of Equation (3) can be obtained by the following steps:

$$U^H A = U^H U A . \tag{4}$$

Set $V = U^H Z, W = U^H U$, then

$$V = WA . \tag{5}$$

$$A = W^{-1} V . \tag{6}$$

So A is the best estimate LS under the length L of the sample data in Equation (6).

Since the Equation (6) involves complex matrix inversion problem, large amount of calculation and difficulty to implementation, the matrix U must be decomposed into the matrix composition, which are relatively simple.

Singular value decomposition (SVD) of the matrix is used to complete the least square algorithm. SVD is an important matrix factorization of linear algebra and is widely deployed in many fields such as signal processing and statistical. Eigen value decomposition is an efficient method to extract features of matrix, but its drawback is applicable only for the square. SVD method, however, can be applied to an arbitrary matrix, regardless of the value of the matrix being a real number or a complex number.

It is obvious that if the order of a matrix U is $M \times N$, then U can be derived from two unitary matrices W and V which are orthogonal matrices with Σ which is a diagonal matrix and has the same dimensions of matrix U, i.e., U can be written as:

$$U = W \sum V^T , \tag{7}$$

which leads to:

$$\begin{aligned} \|Ua - z\|_2 &= \|W \sum V^T a - z\|_2 \\ &= \|W(\sum V^T a) - W(W^T z)\|_2 . \\ &= \|W(\sum y - c)\|_2 \end{aligned} \tag{8}$$

For W is an orthogonal matrix, we have that $\|Ua - z\|_2 = \|\sum y - c\|_2$. Next step is to seek the smallest y, which makes $\|\Sigma y - c\|_2$ minimum. It can be simply achieved in that Σ is a diagonal matrix [6].

Assuming the rank of matrix U is r, then the corresponding equation is yielded as follows:

$$\Sigma y = \begin{bmatrix} \sigma_1 y_1 \\ \vdots \\ \sigma_r y_r \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ \Sigma y - c = \begin{bmatrix} \sigma_1 y_1 - c_1 \\ \vdots \\ \sigma_r y_r - c_r \\ -c_{r+1} \\ -c_{r+2} \\ \vdots \\ -c_m \end{bmatrix}. \tag{9}$$

Thus, $y_i = \frac{c_i}{\sigma_i}, (i = 1, 2, \cdots, r)$ makes $\Sigma y - c$ reaching its minimum length $\left[\sum_{i=r+1}^{m} c_i^2 \right]^{1/2}$. When $r = m$ and the column of the matrix U is m, the least squares problem can be solved without error.

The matrix transposing Σ and getting the inverse of the nonzero diagonal elements of matrix is defined as Σ^+, then the r elements of $y = \Sigma^+ c$ will be equal to $\frac{c_i}{\sigma_i}, (i = 1, 2, \cdots, r)$, and the rest of elements to zero. At the same time, since $y = V^T x$, $c = W^T z$, x can be computed from the following Equation:

$$x = V \Sigma^+ W^T z . \tag{10}$$

As a result, the minimum norm of LS algorithm can be obtained using Equation 10.

5. Predistorter Framework

Adaptive digital predistortion structure has two main categories based on lookup table LUT [7] or polynomial digital adaptive predistortion method.

Predistorter that is based on lookup table structure is using one or more lookup tables to represent the nonlinear relationship of predistorter. The basic principle is that the appropriate predistortion coefficients are stored in terms of table. Input signals are processed according to the index of the table to find the corresponding predistortion coefficients for computing, and the results are output to the successor circuits to achieve the purpose of correction of nonlinear power amplifier. At the same time, through a feedback path, the contents of the lookup table are updated dynamically by adaptive algorithm.

The method based on LUT is easy to implementation, but the algorithm converges relatively slowly, and a lot of memory cells are required, which also increases the resource cost. Polynomial linearization method is better than the lookup table. It has many advantages: fast convergence, requires less storage unit. The design of the predistorter structure is also use the method of the memory polynomial. Predistortion based on polynomial model mainly deploys some form of function, calculating the numerical, which is the original input signal through predistorter. It needs only to adjust the number of the polynomial coefficients, and bears a faster convergence rate. This method concentrates more on the issue of saving RAM resources in a lookup table.

Predistortion has three common memory polynomial models: Memory Polynomial/ Wiener model (MP/W), Memory Polynomial (MP) and Generalized Memory Polynomial (GMP).

5.1. MP Model

Expression of MP model is given by:

$$z_{MP}(n) = \sum_{k=0}^{k-1} \sum_{m=0}^{M-1} a_{km} x(n-m) \left| x(n-m) \right|^k . \tag{11}$$

Memory polynomial predistortion is significantly effective for the actual power amplifier model [8]. The essence of the polynomial predistortion is the nonlinear approximation problems of power amplifier model, and the core of polynomial nonlinear approximation depends on the structure used. The existing structure is based on integer powers, without taking into consideration fractional power. Hence, an improved model of fractional order proposed to extend polynomial design and improve predistortion performance [9].

Fractional order memory polynomial can be expressed as:

$$z_{FMP}(n) = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} a_{km} x(n-m) \left| x(n-m) \right|^{k + \frac{\mathrm{mod}(k,2)}{2}} . \tag{12}$$

5.2. MP/W Model

MP/W model with cross terms is given below:

$$\begin{aligned} z_{MPW}(\mathrm{n}) = & \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} a_{km} x(n-m) \left| x(n-m) \right|^k \\ & + \sum_{k=1}^{K-1} b_k x(n) \left| \sum_{m=0}^{M-1} c_m \left| x(n-m) \right| \right|^k . \end{aligned} \tag{13}$$

Coefficients a_{km} and b_k are estimated using the least squares method, and the coefficients c_m obtained through the Newton iterative method [10]. Its convergence, however, is unstable in that it is susceptible to initial coefficient values. Therefore, the cross terms of polynomial introduction although theoretically can improve the predistortion effect, but the coefficients estimation is not robust enough.

5.3. GMP Model

Expression of GMP model is given by [9]:

$$\begin{aligned} z_{GMP}(n) = & \sum_{k\in K_a}\sum_{l\in L_a} a_{kl}x(n-l)\left|x(n-l)\right|^k \\ & + \sum_{k\in K_b}\sum_{l\in L_b}\sum_{m\in M_b} b_{klm}x(n-l)\left|x(n-l-m)\right|^k \\ & + \sum_{k\in K_c}\sum_{l\in L_c}\sum_{m\in M_c} c_{klm}x(n-l)\left|x(n-l+m)\right|^k \end{aligned} . \tag{14}$$

It can be seen that GMP model expression is so cumbersome that huge computation and complex algorithm are the limitations of coefficient estimation.

In summary, the MP model is most commonly applied due to its easy implementation and good approximation to practical PA model.

6. Simulation and Analysis

To verify the linearization performance of the new joint learning structure compared with the indirect, simulations are carried out in terms of the improvement on power spectrum density (PSD).

In this simulation, original signals use orthogonal frequency division multiplexing (OFDM) signals with bandwidth of 8MHz. The PA model is based on memory polynomial. Expression of a frequently-used memory polynomial is given by:

$$y(n) = \sum_{\substack{k=1 \\ odd}}^{K}\sum_{l=0}^{L} a_{kl}x(n-l)\left|x(n-l)\right|^{k-1} . \tag{15}$$

Here:

$$\begin{aligned} a_{10} &= \ \ 1.0513+0.0904j & a_{30} &= -0.0542-0.2900j \\ a_{50} &= -0.9657-0.7028j & a_{11} &= -0.0680-0.0023j \\ a_{31} &= \ \ 0.2234+0.2317j & a_{51} &= -0.2451-0.3735j \\ a_{12} &= \ \ 0.0289-0.0054j & a_{32} &= -0.0621-0.0932j \\ a_{52} &= \ \ 0.1229+0.1508j \end{aligned}$$

The comparisons between joint and indirect structure on linearization performance of PSD are portrayed in Figure 5 and Figure 6.

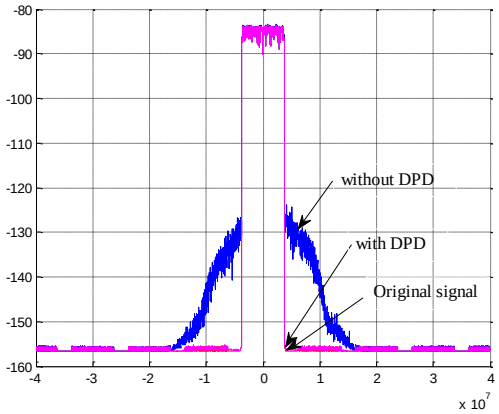


Figure 5. Indirect structure without noise

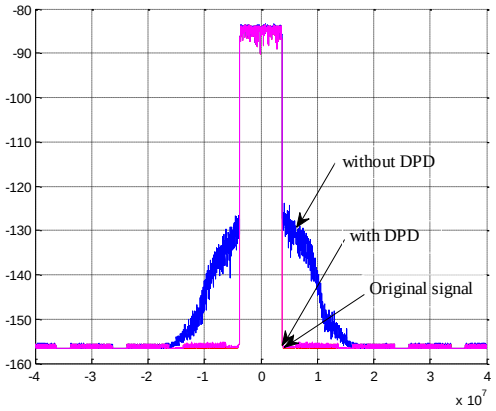


Figure 6. New structure without noise

It can be seen from Figure 5 and Figure 6 that both indirect learning structure and the new structure of the feedback signal have similar linearization performance on PSD in case of noise-free condition.

At the same time, in the case of a certain extent of noise, as seen in Figure 7 and Figure 8, the performance between joint and indirect structure show a difference on improvement on PSD. The former has a significant advantage over the latter. This suggests that the new joint learning structure bears a strong anti-noise characteristic, and hence, better linearization performance for PA than the indirect learning structure.

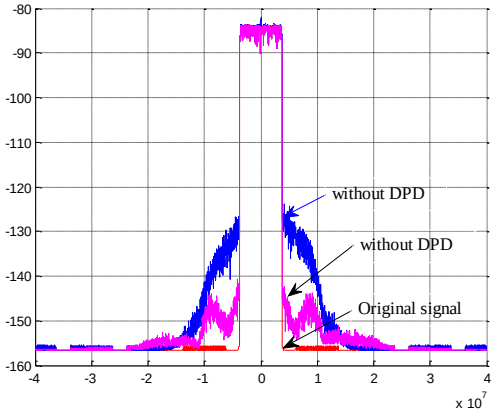


Figure 7. Indirect structure with noise

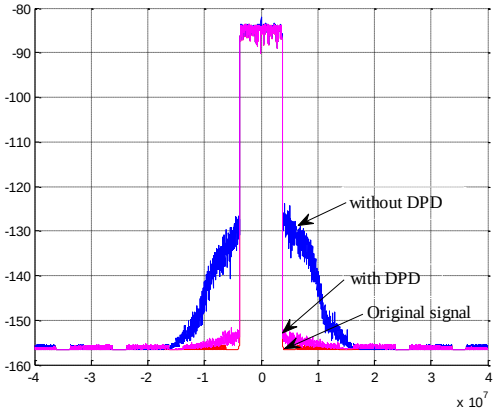


Figure 8. New structure with noise

7. Conclusions

In this paper, a joint structure which uses indirect with direct for linearization of PA is presented. Simulations show that in the case of other condition being the same, this new structure of adaptive algorithm can avoid the noise interference, and therefore, obtaining a significant improvement over indirect learning structure.

Acknowledgements

Sponsor by Research and Development on Command and Scheduling System for Intensive Reports of Major News (2012BAH17B01).

References

1. Qian, Y. (2006) High efficient adaptive predistortion structure for RF power amplifier linearization. *Journal on Communications*, 27, 35-40.
2. Presti, C. D., Metzger, A.G., Asbeck, H. M. (2010) Efficiency improvement of a handset WCDMA PA module using adaptive digital predistortion. IEEE MTT-S Int. Microwave Symp. Dig., June 2010 (pp. 804-807)
3. Li, Zh., Kuang, J., Wu, N. (2012) Direct Learning Predistorter with a New Loop Delay Compensation Algorithm, Vehicular Technology Conference (VTC Spring), IEEE 75th (pp.1-5)
4. Bo, A., Zhi-xing, Y., Chang-Yong, P. (2008) A novel scheme for power amplifiers based on indirect learning architecture, *Wireless Personal Communications*, 46, 523-530
5. Guan, L., Zhu, A. (2011) Dual-loop model extraction for digital predistortion of wideband RF power amplifiers. *IEEE Microwave and Wireless Components letters*, 21(9), 501-503.
6. Bergqvist, G., Larsson, E.G. (2010) The Higher-Order Singular Value Decomposition: Theory and an Application, *IEEE Signal Processing*, 27(3), 151-154
7. Zhao, Y., Qin, Z., Gaofeng, X., Jiong, L. (2012) PA linearization using multi-stage look-up-table predistorter with optimal linear weighted delay. IEEE Conference on Signal Processing, October 21-25, 2012 (pp. 47-51)
8. Zhou, H., Wan, G., Chen, L. (2010) A Nonlinear Memory Power Amplifier Behavior Modeling and Identification Based on Memory Polynomial Model in Soft-Defined Shortwave Transmitter. IEEE Conference on Wireless Communications Networking and Mobile Computing, September 23-25, 2010 (pp. 1-4)
9. Ai, B., Zhong, Z.-D., Jiang, T., Li, B. (2011) Novel Pre-Distortion of Power Amplifier with Proposed Fractional Order Memory Polynomial. IEEE Conference on Global Telecommunications, December 5-9, 2011 (pp. 1-6)
10. Ai, B., Zhong, Z. D., Zhu, G., Xu, R. T., Li, Z. Q. (2008) Two-dimensional indexing polynomial-based pre-distorter for power amplifiers with memory effects. *IET Communications*, 2(10), 1263-1271

Received on the 3rd of November 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 17-20
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

STUDY ON MAHALANOBIS DISCRIMINANT ANALYSIS OF EEG DATA

Yuan Shi, Linlin Yu, Li Zou

[1]Dalian Institute of Science and Technology
Dalian Jiao Tong University
BinGang Str.999-26, DaLian, China
Phone: (+0086)-13664247491. E-mail: 20088041@qq.com

Objective In this paper, we have done Mahalanobis Discriminant analysis to EEG data of experiment objects that are recorded impersonally come up with a relatively accurate method used in feature extraction and classification decisions. Methods In accordance with the strength of α wave, the head electrodes are divided into four species. In use of part of 21 electrodes EEG data of 63 people, we have done Mahalanobis Discriminant analysis to EEG data of six objects. Results In use of part of EEG data of 63 people, we have done Mahalanobis Discriminant analysis, the electrode classification accuracy rates is 64.4%. Conclusions Mahalanobis Discriminant has higher prediction accuracy, EEG features (mainly α wave) extract more accurate. Mahalanobis Discriminant would be better applied to the feature extraction and classification decisions of EEG data.

Keywords: Electroencephalogram, Mahalanobis Discriminant, α hythm

1. Introduction

The purpose of routine brain wave inspection is to evaluate whether the brain wave is normal or not and provide help to diagnose the brain disorders which is also known as brain wave interpretation. The traditional brain wave interpretation is realized through reading the multi-channel electroencephalogram on the recording paper by experts, which is to understand and evaluate electroencephalogram (EEG) with the method of visual inspection. The essence of this method based on expertise is that experts utilize experience to wipe out the disturbance and artefact of signals, conduct feature extraction to the EEG according to the frequency, range, phase position and other information, and carry out the category description for the extracted features with the recognized experience to analyse and evaluate the EEG [1]. Up to now, this method is widely applied to the clinic. The visual inspection, to some extent, can catch the pathological waveform or even confirm the position of the brain focus. However, due to the strong non-stationary and nonlinear characteristics of EEG, with the addition of the great dependence of visual inspection on knowledge-level and experience of EEG analysis personnel, the new method must be explored to realize the breakthrough of EEG research [2].

Mahalanobis Discriminant analysis has been introduced into the research of EEG, which will actively promote the extraction and classification of EEG data to assist the inspection and quantitative analysis of EEG and provide the effective analysis means for the EEG examination.

2. Objects and Methods

2.1. Object of Study

We take 28 men and 35 women as the research objects, whose age is ranging from 20 to 60, and the average age is 36.7. All the subjects are enjoying good health without serious nerve system diseases and history of taking psychotropic drugs, and they are selected from the normal population.

2.2. Build the Selection of Mathematical Modelling EEG Data

The sampling frequency of experiment recording of EEG is 100Hz which is recording 21 electrode data according to the lead location in international 10-20 system: C3, CZ, C4, FP1, FPZ, FP2, F7, F8, FZ, F3, F4, O1, OZ, O2, P3, PZ, P4, T5, T6, T3, T4. A block (indicating a short time period) of EEG data is acquired at every turn, and the number of sampling points for each block is 512 with the recording time of 5.12s. The electroencephalogram of normal people is mainly in α rhythm, the strengths of α wave appear in the occiput, and then weakening gradually from back to front. Classify the 21 conducting electrodes into 4 categories in accordance with the intensity differences of the α rhythm in various parts of the head, which is, former head electrode, side head electrode, central electrode, occiput electrode. The specific

classification situation is as follows:

(1) The first category: central electrode (C3, CZ, C4)
(2) The second category: former head electrode (FP1, FPZ, FP2, F7, F8, FZ, F3, F4)
(3) The third category: occiput electrode (O1, OZ, O2, P3, PZ, P4, T5, T6)
(4) The forth category: side head electrode (T3, T4).

2.3. The Computer Processing of EEG Data

The electroencephalogram dedicated toolbox EEG Toolbox is designed with the MATLAB programming language in order to facilitate analysing the original data of the electroencephalogram. In EEG Toolbox, after the original data was introduced, it was saved in the matrix, and line represents the timing of experiment recording (that is sampling point) while column indicates the electrode. All the data of every subject were introduced before analysis and the electroencephalogram should be shown intuitively, and a block of EEG data should be displayed on each page [3].

The 4-population Mahalanobis distance discriminance classifies the sample data into four categories based on the electrode classification method introduced above. Firstly, put the 21 electrode EEG data, which will build the mathematical models into the four matrixes $\overline{X}_i (i = 1, 2...4)$ on the basis of classification. Put the current block of EEG data into the Matrix X, which is 512×21 matrix. The electrode classification results are predicted with Mahalanobis distance discriminance and expressed by putting them into vector.

When the Mahalanobis distance discriminance is adopted, the discriminating data will be classified into the nearest category based on the distance length of population centre from the distinguishing EEG data. The Mahalanobis distance analysis procedure in the research is on the basis of multi-channel EEG data design. Firstly, the mathematical model, namely, discrimination function, should be built, then predict the category of EEG data according to the discrimination rules. The Mahalanobis distance discriminance could be explained by the following mathematical formula:

$$d_j^2(X) = A * B, \tag{1}$$

$$A = (X - \frac{1}{n_j}\sum_{i=1}^{n_j} X_i^{(j)})', \tag{2}$$

$$B = [\frac{1}{n - n_j}\sum_{j=1}^{n_j}(X_i^{(j)} - \overline{X}^{(j)})(X_i^{(j)} - \overline{X}^{(j)})']' A \cdot \tag{3}$$

Formula (1-3) is the Mahalanobis distance discriminance function, and 4-population Mahalanobis distance discriminance will build four discriminance functions. Substitute an unknown classified EEG data X into the four Mahalanobis distance discrimination functions, and acquire the minimum Mahalanobis distance and discriminate it to the corresponding totality.

Each block of EEG data can be predicted and classified by the Mahalanobis distance discrimination function and the predicated classification results as well as actual classifications can be intuitively displayed in the Mahalanobis distance discriminance predicating result chart.

3. EEG Data Analysis Results from Mahalanobis Distance Discriminance

Classify the 21 conducting electrodes into four categories according to the intensity differences of α wave on various parts of the head, and forecast 21 electrodes classification conditions of six subjects in the current block with Mahalanobis distance discriminance: C3, CZ, C4, FP1, FPZ, FP2, F7, F8, FZ, F3, F4, O1, OZ, O2, P3, PZ, P4, T5, T6, T3, T4. Draw the Mahalanobis distance discriminance predicating result chart and 2D pie chart of Mahalanobis distance discrimination accuracy rate with Mahalanobis distance discriminance procedure [4].

Mahalanobis distance discriminant analysis procedure can forecast and classify the EEG data of all blocks for various subjects. Due to the space constraints, only the forecast results of the EEG data in six blocks for three subjects are given in detail, the whole situation can be reflected by showing only classification results of two blocks for each subject, and the forecast classification results in other blocks are similar to these. Analyse the predicated results of one subject in detail. We number the six subjects for the sake of convenient description: 1, 2, 3, 4, 5, 6. First conduct the Mahalanobis distance discriminant analysis for the 12th block of EEG data of Subject 1. The figure of EEG is shown in Fig1.

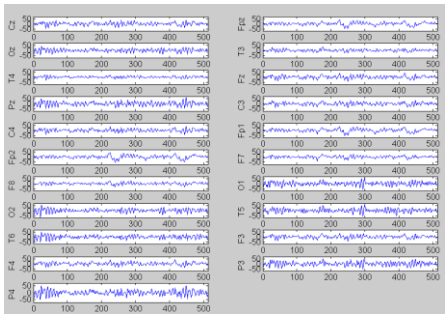


Figure 1. EEG data of the twelfth block of in subject 1
Note 1): The horizontal ordinate is frequency (Unit: Hz), while the vertical coordinate is voltage (Unit: μV), and the electrode parts have been marked on the left side of the data.

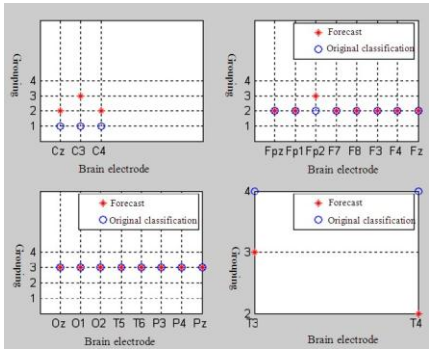


Figure 2. Mahalanobis distance predicating results of the twelfth block in Subject 1
Note 1) The top left corner is the predicating results of the first–category electrode; the top right corner is the predicating results of the second–category electrode; the left bottom is the predicating results of the third–category electrode; and the bottom right corner is the predicating results of the forth–category electrode. 2) The horizontal ordinate is the names of various kinds of electrodes and the category is listed on the vertical coordinate. 3) red* represents the predicated classification situation, blue O indicates the actual category situation. * And O will coincide when the predicated classification is consistent with the actual classification. 4) first–category electrode (C3, CZ, C4), second–category electrode (FZ, F3, F4, FP1, FPZ, FP2, F7, F8), third–category (P3, PZ, P4, O1, OZ, O2, T5, T6), forth–category electrode (T3, T4)

Carry out the Mahalanobis distance discriminant analysis for the six subjects, and randomly extract 10 blocks of EEG data for every subject to predict classification. The Mahalanobis distance predicating results is shown by Fig2. The average accuracy rate is 64.4% (shown in Table 1). Overall, the forecast results of Mahalanobis distance discriminance are better; the extraction of the EEG characteristics (mainly α wave) is relatively accurate. The predicating results can reflect the intensity differences of wave α in various parts of the head. To some extent, the occurrence rate or amount of wave α, namely, the quantity of wave α recorded in the EEG within a certain period of time, has huge differences in individuals. The forecast results are influenced by such situations where wave α constantly appears in some people and sporadically in other people, or when other frequency waves appear. The mis-discriminance can also be caused by the amplitude modulation and right-and-left difference.

Table 1. The average accuracy rate of Mahalanobis distance discriminance predicating EEG classification

Subject	1	2	3	4	5	6	Average Accuracy Rate
Accuracy Rate	66.7%	63.2%	64.2%	63.5%	65.4%	63.1%	64.4%

4. Conclusions

Because Euclidean distance is oversimplified, and the absolute distance and Chebyshevy distance cannot completely express the characteristic differences of the multidimensional data in the high-dimensional space, therefore, we usually analyse the EEG data with Mahalanobis distance discriminance in experiments. Classify 21 conducting electrode into four categories according to the intensity differences of the wave α in every part, build Mahalanobis distance discriminance mathematical model with 21 brain electrode data, and conduct the Mahalanobis distance discriminant analysis to the EEG data of 6 normal subjects. The predicating classification accuracy rate is 64.4%. On the whole, the predicating

results of Mahalanobis distance discriminance is better, the extraction of EEG characteristics (mainly wave α) is more accurate, and the predicating results can reflect the intensity differences of wave α on the various parts of the head. The experiment indicates that Mahalanobis distance discrimination can preferably extract the EEG characteristics of normal people and can be applied to the classification decision of EEG data.

The EEG of normal people presents α rhythm and wave α is the major EEG characteristic of normal people. The predicating classification results of different blocks are not completely equivalent, which reflects that the EEG is a non-stationary random signal and wave α is constantly changing. The amplitude modulation phenomenon, left-and-right difference and individual differences in subjects will exert an influence on predicating classification of EEG data and cause the mis-discriminance. We analyse the reasons for mis-discriminance of the electrodes as follows:

(1) When the Mahalanobis distance classification discriminance is adopted, the discriminating data will be classified into the nearest category based on the distance length of each population centre from the distinguishing EEG data [5]. It is impossible to get infinite EEG data samples, so the limited samples with centralized training are used to estimate every population centre. The waveforms, amplitudes and phase positions of the normal people collected in the experiment are different in the EEG data, which may lead to the inaccuracy of predicating classification.

(2) When Mahalanobis distance discrimination is adopted to analyse the EEG data of normal people, where mis-discrimination exists. The principle of Mahalanobis distance discriminance is to classify the discriminating samples to the nearest category [6]. If the two kinds of the samples are overlapped and the discriminating samples are just in the overlapped area, the samples will probably be classified wrongly. We take the two populations for example in order to give a better explanation. Suppose that the sample x is one-dimension variable, G1 and G2 are two populations, and Figure 3 is the distribution situation of the two populations. For instance, if x belongs to Population G1, but it falls on the left side of $\overline{\mu}$, in accordance to the rules, x is classified to G2; similarly, the point in G2 can be classified to G1 wrongly with Mahalanobis distance discriminance. The error probability is shown in the shadow area of the diagram. If the two populations are near to each other, the rate of mis-discrimination must score high [7].

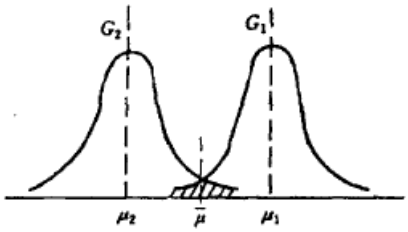


Figure 3. Distribution diagram of two populations

Sponsored by Hi-Tech Res earch and Development Program of China (863) (2012AA040912).

References

1. Trinder, J., A. van Beveren, J., Smith, P., et.al. (2005) Correlation between ventilation and EEG arousal during sleep onset in young subjects [J], *Journal of Applied Physiology*, 83, 2005-2011.
2. Kasper, K., Schuster, H. G. (2001) Easily calculable measure for complexity of spatial temporal pattern. Physical Review Online Archive, 36 (2), 842-848.
3. Viglione, S. S., Ordon V. A., Risch, F. (1999) A methodology for detecting ongoing changes in the EEG prior to clinical seizures. In: Western Institute on Epilepsy, West Huntington Beach, 27-28.
4. Williams, W. J. (1999) Time-Frequency Analysis of Biological Signals, *IEEE Electrical Computer Science*, 12(1), 83-86.
5. Blanco, S., et.al. (2007) Applying time-frequency analysis to seizure EEG activity, *IEEE Engineering in medicine and biology magazine*, 16, 65-71.
6. Williams, W. J., Hitten, P., Zaveri, J. (1995) Time-frequency Analysis of Electrophysiology Signals in Epilesy, *IEEE Transaction on Biolmedical Engineering*, 3(2), 133-142.
7. Richman, J. S., Moorman, J. R. (2000) Physiological time-series analysis using approximate entropy and sample entropy, *AJP Heart and Circulatory Physiology*, 278(6), 2039-2049.

Received on the 17th of June 2013

Mathematical and Computer Modelling

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 21-28
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

AN INVERSE PROBLEM OF FINDING A PARAMETER FOR THE HEAT CONDUCTION EQUATION

***Sufang Yue*[1,*]**

[1] *School of Mathematics and Computing Science, Anqing Normal University, Anqing, 246011, China*
[*] *Corresponding author e-mail address: ysf6510@126.com*

In this paper, the inverse problem of finding a parameter for the heat equation is investigated. The reproducing kernel space is constructed to absorb all the conditions for determining solution. Using the reproducibility of reproducing kernel, we have conducted two iteration sequences and proved their convergence. Numerical experiments are presented and discussed.

Keywords: Inverse problem, Reproducing kernel space, Iteration

1. Introduction

In this paper, we consider the following inverse problem for the heat equation to find a pair (w, f) such that

$$\begin{cases} \rho(t,x)w_t - w_{xx} = f(x)g(t,x) + h(t,x), & 0 \le x \le 1, 0 < t < T \\ w(0,x) = w_0(x), & 0 \le x \le 1 \\ w(t,0) = w(t,1) = 0, & 0 \le t \le T \\ w(T,x) = q(x), & 0 \le x \le 1 \end{cases}, \tag{1}$$

where $\rho(t,x), h(t,x), g(t,x), w_0(x)$ and $q(x)$ are known functions, while the functions $w(t,x), f(x)$ are unknown.

Let $u(t,x) = w(t,x) - w_0(x)$, we get

$$\begin{cases} \rho(t,x)u_t - u_{xx} = f(x)g(t,x) + h(t,x) + (w_0)_{xx}, & 0 \le x \le 1, 0 < t < T \\ u(0,x) = 0, & 0 \le x \le 1 \\ u(t,0) = u(t,1) = 0, & 0 \le t \le T \\ u(T,x) = q(x) - w_0(x), & 0 \le x \le 1 \end{cases}. \tag{2}$$

Thus, we only need to solve (2).

Inverse problem of heat conduction equations is applied widely in project, physics and so on. Since it has ill-posed difficulties in essence and heavy computation in practice, more and more people engage in it, such as H. Han et al.'s element method [1], N.S. Mera et al.'s iterative boundary element method [2], S.M. Kirkup, and M.Wadsworth's operator-splitting method [3], K.Iijima's finite difference method [4]. V.L.Kamynin [5] proved the unique solvability of an inverse problem of determining the right-hand side of heat conduction. However, the numerical experiments were seldom discussed.

In this paper, the inverse problem of finding a parameter for the heat equation is investigated. The reproducing kernel space is constructed to absorb all the conditions for determining solution. Using the reproducibility of reproducing kernel, we have conducted two iteration sequences and proved their convergence. Numerical experiments are presented and discussed. The reproducing kernel method has been proved by many authors to be a powerful mathematical tool for various kinds of linear and nonlinear problems, such as singular two-point boundary value problem [6], nonlinear age-structured population model [7] and inverse problem [8].

2. The Reproducing Kernel Spaces

In this section, we introduce several reproducing kernel spaces needed.

2.1. The Reproducing Kernel Space $W_2^1[0,1]$

The Hilbert space $W_2^1[0,1]$ is defined by $W_2^1[0,1]=\{v|v$ is absolutely continuous real value function, $v,v'\in L^2[0,1]\}$. The inner product is given by

$$(u,v)_{W_2^1[0,1]}=u(0)v(0)+\int_0^1 u'(x)v'(x)dx. \tag{3}$$

The kernel function is

$$K_x(y)=\begin{cases}1+x, y>x\\1+y, y\le x\end{cases}. \tag{4}$$

We usually call it reproducing kernel with the property

$$(v(y),K_x(y))=v(x)..\forall v\in W_2^1[0,1]. \tag{5}$$

2.2. The Reproducing Kernel Space ${}^0W_2^2[0,T]$

The reproducing kernel space ${}^0W_2^2[0,T]$ is given by ${}^0W_2^2[0,T]=\{v|v,v'$ are absolutely continuous functions, $v,v',v''\in L^2[0,T],v(0)=0\}$. The inner product is defined by

$$(u,v)_{{}^0W_2^2[0,T]}=u'(0)v'(0)+\int_0^T u''(x)v''(x)dx, \tag{6}$$

and the corresponding reproducing kernel is

$$K2_x(y)=\begin{cases}\dfrac{1}{2}x(2+x)y-\dfrac{x^3}{6}, y>x\\ xy+\dfrac{xy^2}{2}-\dfrac{y^3}{6}, y\le x\end{cases}. \tag{7}$$

It is easy to verify that this kernel satisfy the reproducing property

$$(v(y),K2_x(y))=v(x),\forall v\in{}^0W_2^2[0,T]. \tag{8}$$

2.3. The Reproducing Space ${}^0W_2^3[0,1]$

Similarly, we define the reproducing space ${}^0W_2^3[0,1]=\{v|v,v',v''$ are absolutely continuous real value functions, $v,v',v'',v'''\in L^2[0,1],v(0)=v(1)=0\}$. The inner product is defined by

$$(u,v)_{{}^0W_2^3[0,1]}=u'(0)v'(0)+u''(0)v''(0)+\int_0^1 u'''(x)v'''(x)dx. \tag{9}$$

For any $y\in[0.1]$ and any $v(x)\in{}^0W_2^3[0,1]$, $K3_x(y)$ must satisfy $(v(y),K3_x(y))=v(x)$. Applying (9), we have

$$(v(x),K3_y(x))_{{}^0W_2^3[0,1]}=v'(0)K3'_y(0)+v''(0)K3''_y(0)+\int_0^1 v'''(x)K3'''_y(x)dx. \tag{10}$$

Since

$$\int_0^1 v'''(x)K3'''_y(x)dx=\sum_{i=0}^{2}(-1)^i v^{(2-i)}(x)\frac{\partial^{3+i}(K3)_y(x)}{\partial x^{3+i}}\Bigg|_{x=0}^{x=1}-\int_0^1 v(x)\frac{\partial^6(K3)_y(x)}{\partial x^6}dx. \tag{11}$$

By variable substitution, one obtains

$$\sum_{i=0}^{2}(-1)^i v^{(2-i)}(x)\frac{\partial^{3+i}(K3)_y(x)}{\partial x^{3+i}}=\sum_{i=0}^{2}(-1)^{2-i}v^{(i)}(x)\frac{\partial^{5-i}(K3)_y(x)}{\partial x^{5-i}}. \tag{12}$$

Moreover,

$$(v(x),K3_y(x))=\sum_{i=0}^{2}v^{(i)}(0)\left|\frac{\partial^i K3_y(0)}{\partial x^i}-(-1)^{2-i}\frac{\partial^{5-i}K3_y(0)}{\partial^{5-i}}\right|+\sum_{i=0}^{2}(-1)^{2-i}v^{(i)}(1)\frac{\partial^{5-i}K3_y(1)}{\partial^{5-i}}-\int_0^1 v(x)\frac{\partial^6 K3_y(0)}{\partial^6}dx. \quad (13)$$

Therefore, $K3_y(x)$ is the solution of the following generalized differential equations:

$$\begin{cases} -\dfrac{\partial^6(K3)_y(x)}{\partial x^6}=\delta(x-y), \\ \dfrac{\partial^i(K3)_y(0)}{\partial x^i}-(-1)^{2-i}\dfrac{\partial^{5-i}(K3)_y(0)}{\partial x^{5-i}}, \\ \dfrac{\partial^{5-i}(K3)_y(1)}{\partial x^{5-i}}=0, i=0,1,2. \end{cases} \quad (14)$$

While $x\neq y$, it is easy to know that $K3_y(x)$ is the solution of the following constant linear homogeneous differential equation with 6 orders, i.e.

$$-\frac{\partial^6(K3)_y(x)}{\partial x^6}=0 \quad (15)$$

with the boundary conditions:

$$\begin{cases} \dfrac{\partial^i(K3)_y(0)}{\partial x^i}-(-1)^{2-i}\dfrac{\partial^{5-i}(K3)_y(0)}{\partial x^{5-i}}=0, \\ \dfrac{\partial^{5-i}(K3)_y(1)}{\partial x^{5-i}}=0, i=0,1,2. \end{cases} \quad (16)$$

We know that equation (15) has characteristic equation $\lambda^6=0$, and the eigenvalue $\lambda=0$ is a root whose multiplicity is 6. Therefore, the general solution of Eq. (14) is

$$K3_x(y)=\begin{cases} LK_y(x)=\sum_{i=0}^{5}a_i(x)y^{(i)}, y\le x \\ RK_y(x)=\sum_{i=0}^{5}b_i(x)y^{(i)}, y>x \end{cases}. \quad (17)$$

Now we are ready to calculate the coefficients $a_i(x)$ and $b_i(x)$, i=1,…,6.

Since

$$-\frac{\partial^6(K3)_x(y)}{\partial y^6}=\delta(y-x) \quad (18)$$

we have

$$\frac{\partial^i LK_x(x)}{\partial y^i}=\frac{\partial^i RK_x(x)}{\partial y^i}, i=0,1,2,3,4 \quad (19)$$

and

$$\frac{\partial 5LK_x(x^+)}{\partial y^5}-\frac{\partial^5 RK_x(x^-)}{\partial y^5}=-1. \quad (20)$$

The above 6 equations in (19) and (20) provided 6 conditions for solving the coefficients $a_i(x)$ and $b_i(x)$, (i=1,…,6) in Eq. (17). Noting that Eq. (16) provided six boundary conditions, we have 12 equations, i.e., (16), (19) and (20). It is easy to know these 12 equations are linear equations with the variables $a_i(x)$ and $b_i(x)$ could be calculated by many methods. As long as the coefficients $a_i(x)$ and $b_i(x)$ are known, the exact expression of the reproducing kernel function $K3_y(x)$ could be calculated from Eq. (17).

2.4. The Reproducing Kernel Space $W_{(2,3)}(\Omega)$

$W_{(2,3)}(\Omega)=\left\{\sum_{i,j=1}^{\infty}a_{i,j}g_i(t)h_j(x)\middle|\sum_{i,j=1}^{\infty}\left|\alpha_{i,j}\right|^2<\infty\right\}$, where $g_k(t),h_l(x)$ are respectively the complete orthonormal systems of $W_2^2[0,T]$ and $W_2^3[0,1]$.

The inner product of $W_{(2,3)}(\Omega)$ is defined by $(u(t,x),v(t,x))=\sum_{i,j=1}^{\infty}\alpha_{i,j}\beta_{i,j}$, where

$$u(t,x)=\sum_{i,j=1}^{\infty}\alpha_{i,j}g_i(t)h_j(x) \tag{21}$$

and

$$v(t,x)=\sum_{i,j=1}^{\infty}\beta_{i,j}g_i(t)h_j(x). \tag{22}$$

Space $W_{(2,3)}(\Omega)$ is a reproducing space, and its reproducing kernel is $K_{(s,x)}(t,y)=K2_s(t)\cdot K3_x(y)$, where $K2_s(t),K3_x(y)$ are given by (7) and (17) respectively.

Besides, we can define the spaces $W_{(1,1)}(\Omega)$, $W_{(2,3)}(\Omega)$ and they are all the reproducing kernel spaces.

3. The Iteration Procedure

If $f(x)$ is known and we define the linear operator L: $Lu=\rho(t,x)u_t-u_{xx}$, $u\in W_{(2,3)}(\Omega)$, where $\Omega=[0,T]\times[0,1]$, then we can get the exact solution of equation (1). Let $M=(t,x)$, $M_i=(t_i,x_i)$, $\varphi_i(M)=K_{M_i}(M)$.

$\psi_i(M)=L^*\varphi_i(M)$, where $K_{M'}(M)$ is the reproducing kernel of $W_{(1,1)}(\Omega)$. L^* is the conjugate operator of L. $\{\overline{\psi_i}(M)\}_{i=1}^{\infty}$ derives from Gram-Schmidt orthonormalization of $\{\psi_i(M)\}_{i=1}^{\infty}$, that is $\overline{\psi_i}(M)=\sum_{k=1}^{\infty}\beta_{ik}\psi_k(M)$, $\beta_{ii}>0, i=1,2,\ldots)$.

Theorem 3.1. If $\{M_i\}_{i=1}^{\infty}$ is dense on Ω, and the equations (1) has unique solution, then the solution can give by

$$u(M)=\sum_{i=1}^{\infty}\sum_{k=1}^{i}\beta_{ik}F(M_k,u(M_k),f(x_k))\overline{\psi_i}(M), \tag{23}$$

where $F(M,u,f(x))=f(x)g(t,x)+h(t,x)+w_0)_{xx}$.

Proof. For each fixed $u(M)\in W_{(2,3)}(\Omega)$, let $u(M),\psi_i(M))=0(i=1,2,\ldots)$, then

$$(u(M),(L^*\varphi_i)(M))=(Lu(\cdot),\varphi_i(\cdot))=(Lu)(M_i)=0,(i=1,2,\ldots). \tag{24}$$

Note that $\{M_i\}_{i=1}^{\infty}$ is dense on Ω, therefore $(Lu)(M)=0$. Then get then $\{\psi_i(P)\}_{i=1}^{\infty}$ is the complete system of $W_{(2,3)}(\Omega)$. When the equations (1) has unique solution, we get

$$\begin{aligned}u(M)&=\sum_{i=1}^{\infty}(u(M),\overline{\psi_i}(M))\overline{\psi_i}(M)=\sum_{i=1}^{\infty}\sum_{k=1}^{i}\beta_{ik}(u(M_k),L^*\varphi_k(M))\overline{\psi_i}(M)\\&=\sum_{i=1}^{\infty}\sum_{k=1}^{i}\beta_{ik}Lu(M),\varphi_k(M))\overline{\psi_i}(M)=\sum_{i=1}^{\infty}\sum_{k=1}^{i}\beta_{ik}F(M_k,u(M_k),f(x_k))\overline{\psi_i}(M)\end{aligned}. \tag{25}$$

Then the proof of the theorem 3.1 is complete.

Let $u_0 = q(x) = w_0(x)$, the iteration algorithm is performed through the following two equations

$$f_n(x) = \frac{\rho(T,x)(u_n)_t(T,x) - q_{xx}(x) - h(T,x) + (w_0)xx}{g(T,x)}, \tag{26}$$

$$u_n(M) = \sum_{i=1}^{n}\sum_{k=1}^{i} \beta_{ik} F(M_k, u_{k-1}(M_k), f_{k-1}(x_k))\overline{\psi_i}(M). \tag{27}$$

Now we show they are convergence.

Theorem 3.2. Suppose that $\|u_n\|$ is bounded in equation (26), (27), $\{M_i\}_{i=1}^{\infty}$ is dense on Ω, then u_n convergence to the exact solution u and we also have $u(M) = \sum_{i=1}^{\infty} B_i \overline{\psi}_i(M)$.

Proof. Set $B_i = \sum_{k=1}^{i} \beta_{ik} F(M_k, u_{k-1}, p_{k-1}(x_k))$, then $u_n(M) = \sum_{i=1}^{n} B_i \overline{\psi}_i(M)$ and $u_{n+1} = u_n + B_{n+1}\overline{\psi}_{n+1}$. Considering the $\{\psi_i(M)\}_{i=1}^{\infty}$ is the orthonormality, we can get $\sum_{i=1}^{\infty}(B_i)^2$ is convergent. So $\|u_n\|$ is a Cauchy sequence, u_n is convergent. Define the limit of it by u. Using the density of $\{M_i\}_{i=1}^{\infty}$, we can get $u(M) = \sum_{i=1}^{\infty} B_i \overline{\psi}_i(M)$ easily.

4. Numerical result

Example. Consider Eq. (1) with T=1 and

$$\rho(t,x) = \frac{1}{10} e^{-t-x}, \tag{28}$$

$$g(t,x) = \frac{1}{5} t(e^{-t-x} + 5\pi^2 t), \tag{29}$$

$$h(t,x) = \frac{1}{10} t(e^{-x} - 10\pi^2 + 10e^t\pi^2)\sin(\pi x), \tag{30}$$

$$w_0(x) = e\sin(\pi x) \tag{31}$$

for which the exact solution is

$$u(t,x) = (e^t + t^2 - 1)\sin(\pi x), \tag{32}$$

$$f(x) = \sin(\pi x). \tag{33}$$

In the process of computation, the numerical computations are performed by using Mathematica 5.0. Choosing 144 points on $\Omega = [0,1]\times[0,1]$, respectively, we obtain the approximate solution u_{144} on $\Omega = [0,1]\times[0,1]$. The numerical results are given in Table 1, 2.

TABLE 1. The Computational results for u_{144}.

(t,x)	True solution of $u(t,x)$	Approximate solution of u_{144}	Absolute error
(0.1,0.1)	0.0355898	0.0357156	1.25793E-04
(0.2,0.2)	0.153649	0.152909	7.39758E-04
(0.3,0.3)	0.355853	0.353200	2.65305E-04
(0.4,0.4)	0.619922	0.615087	4.83468E-03
(0.5,0.5)	0.898721	0.891593	7.12779E-03
(0.6,0.6)	1.12426	1.11549	8.77085E-03
(0.7,0.7)	1.21656	1.20747	9.09148E-03
(0.8,0.8)	1.09654	1.08892	7.62057E-03
(0.9,0.9)	0.701346	0.697265	4.08101E-03

TABLE 2. The Computational results for f_{144} .

x	True solution of $f(x)$	Approximate solution of f_{144}	Absolute error
0.1	0.309017	0.308733	2.84057E-04
0.2	0.587785	0.587262	5.23036E-04
0.3	0.809017	0.808351	6.66212E-04
0.4	0.951057	0.950340	7.16628E-04
0.5	1.00000	0.999314	6.86202E-04
0.6	0.951057	0.950464	5.92481E-04
0.7	0.809017	0.808561	4.56293E-04
0.8	0.587785	0.587468	3.16803E-04
0.9	0.309017	0.308876	1.07698E-04

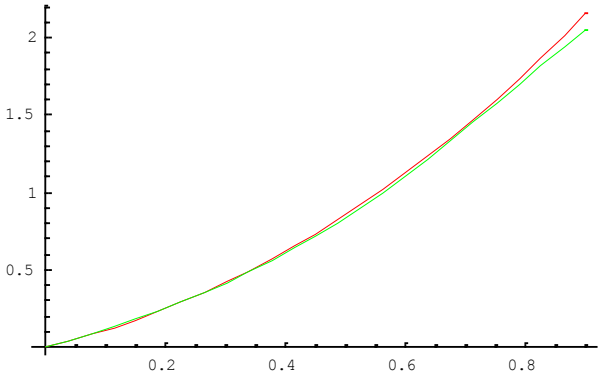


Figure 1. The superimposed image of $u(x,0.4)$ and $u_{25}(x,0.4)$.

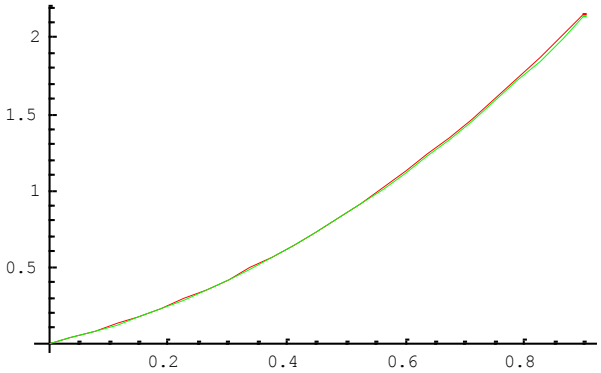


Figure 2. The superimposed image of $u(x,0.4)$ and $u_{144}(x,0.4)$.

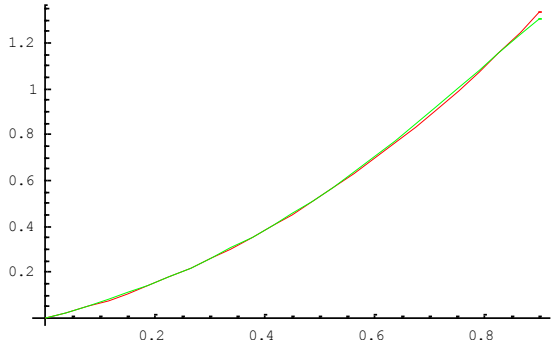


Figure 3. The superimposed image of $u(x,0.8)$ and $u_{25}(x,0.8)$.

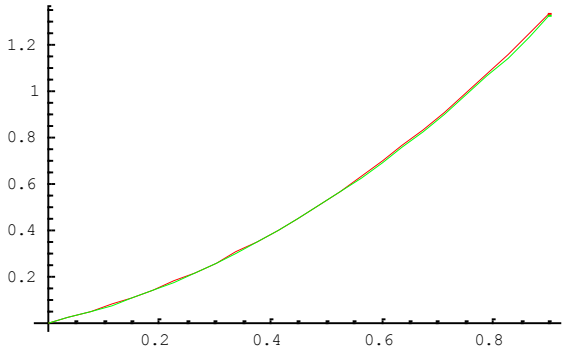


Figure 4. The superimposed image of $u(x,0.8)$ and $u_{144}(x,0.8)$.

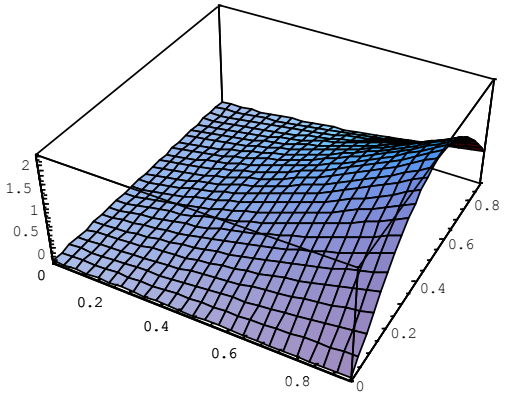


Figure 5. True solution of $u(x,t)$.

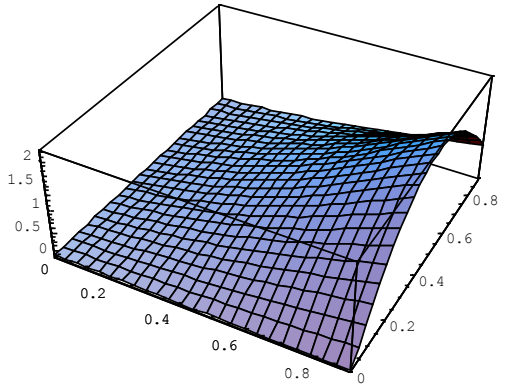


Figure 6. Approximate solution $u_n(x,t)$ of $u(x,t)$.

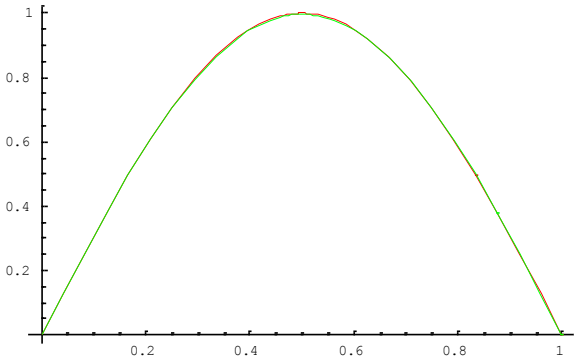


Figure 7. The superimposed image of $f(x)$ and $f_{25}(x)$.

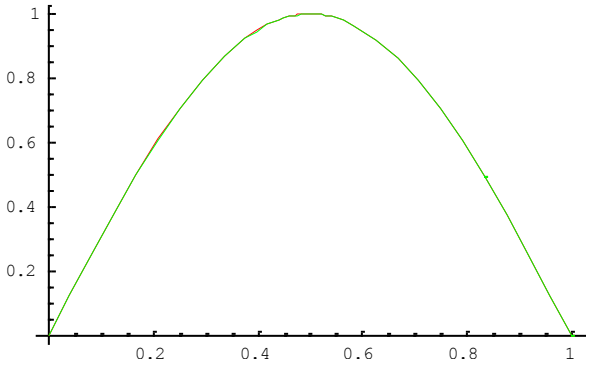


Figure 8. The superimposed image of $f(x)$ and $f_{144}(x)$.

From the results obtained by the method of constructing reproducing kernel spaces that absorb all the conditions for determining solution, we can see the exact solution and the approximate solution are in good agreement with each other.

Acknowledgements

We would like to express our gratitude for financial support from Nature Science Foundation of Anhui province of China (No.KJ2011A197).

References

1. Han, H., Ingham, D.B., Yuan, Y. (1995) The boundary element method for the solution of the backward heat conduction equation. *J. Comp. Phys.*, 116(2), 92-99.
2. Mera, N.S., Elliott, L., Ingham, D.B., Lesnic, D. (2001) An Iterative Boundary Element Method for Solving the One Dimensional Backward Heat Conduction Problem. *International Journal of Heat and Mass Transfer*, 44(10), 1937-1946.
3. Kirkup, S.M., Wadsworth, M. (2002) Solution of Inverse Diffusion Problems by Operator-Splitting Methods. *Applied Mathematical Modelling*, 26(10), 1003-1018.
4. Iijima, K. (2004) Numerical solution of backward heat conduction problems by a high order lattice-free finite difference method. *J. Chin. Inst. Eng.*, 27(4), 611-620.
5. Kamynin, V.L. (2003) On the Unique Solvability of an Inverse Problem for Parabolic Equations under a Final Overdetermination Condition. *Mathematical Notes*, 73(2), 202-211
6. Wang, W.Y., Cui, M.G., Han, B. (2008) A new method for solving a class of singular two-point boundary value problems. *Applied Mathematics and Computation*, 206(2), 721-727.
7. Cui, M.G., Chen, Z. (2007) The exact solution of nonlinear age-structured population model. *Nonlinear Analysis: Real World Applications*, 8(4), 1096-1112.
8. Cui, M.G., Lin, Y.Z. (2007) A new method of solving the coefficient inverse problem of differential equation. *Science in China Series A*, 50(4), 561-572.

Received on the 3rd of November 2013

Mathematical and Computer Modelling
Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 29-36
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

RESEARCH ON A NEW DISTRIBUTED HYDROLOGICAL MODEL (RDG) APPLY IN AGRICULTURE

X.-Y. Meng[1,2,5], X.-N. Ji[3], Y.-C. Guo[1,2], A.-X. Xi[1], Q.-D. Zhao[4], Y. Liu[1,2*]

[1]School of Resources and Environment Science, Xinjiang University, Urumqi, 830046, China;
[2]Key Laboratory of Oasis Ecology Ministry of Education, Xinjiang University, Urumqi, 830046, China;
[3]Jiang An Campus of Sichuan University, College of Architecture and Environment, Chengdu, 610207, China;
[4]Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou,730000, China;
[5] Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, 830011, China;
**Corresponding author: Yang Liu E-mail address: 2587625936 @qq.com*

This research created an integrated distributed hydrologic model (RDG) for quantification of water availability over agricultural basin. The RDG model is a physical- based distributed hydrologic model. This model divided river basin into a raster network which represents the heterogeneous characters of the basin. The model components for the RDG model include: vegetation Interception, evapotranspiration infiltration and groundwater flow, furthermore, this model is combined with a GIS software package to deal with the complex landscape characteristics for agricultural basins. The model performance is verified with observed data and applied to discuss the effects of irrigation strategies, for instance, the reasonable method of pumped water and diversion corresponds to water utilization.

Keywords: Agricultural water management, Hydrologic model, Distributed hydrologic model, GIS

1. Introduction

Recent decades, natural hydrologic cycle are affected by continuous drought impact and anthropogenic disturbance in China's agricultural areas, e.g. the Huang – Huai-Hai (HHH) plain as well as many similar regions in the world These factors have greatly altered the conditions of surface runoff generation and groundwater recharge, which has led to the continuous attenuation of water resources. For the universality of the problems, the study of comprehensive, long-term strategies including legal system and technology methods of water resources development becomes increasingly important. Compared to legal system, the basic study on hydrological cycle and the technology of water resources development are relatively non-sufficient. This work needs to be done so as to anticipating and minimizing potential environmental and human activities impacts [1]. To accomplish this task, the integrated distributed hydrologic models must be developed and they are comprehensive, basin-wide, and continuous in time as well as practical and conceptually clear.

Distributed hydrologic models have been developing rapidly since the first outline of a physically-based distributed model [2] and two well-known physically based distributed catchment models are the Système Hydrologique Européen [3].The Institute of Hydrology Distributed Model [4]. Later, some fully coupled distributed hydrologic models occurred for integrated surface, soil and groundwater processes. For example, a fully coupled distributed model was developed and the simultaneous solution strategy was used to couple surface and groundwater flow model [5].The governing equations in the coupled distributed models are highly nonlinear and require the solution of large systems of equations even for small-scale problems [6]. Therefore, the usefulness of such models for water management of agricultural watershed remains limited.

The practical and operational models applicable to agricultural water management need to not only represent all relevant hydrological processes in a physically meaningful way, but also include the most important factors that are necessary for the majority of management decisions, and allow large-scale basin-wide application. Such practical considerations have led to many integrated distributed hydrologic model systems used in agricultural basin. For example, the well-known SWAT model [7], which is a surface water model in essence, and a simple groundwater model was added to it in recent versions. By incorporating tile drain functions into the framework of TOPMODEL, the model was modified for rainfall-runoff simulation in upland agricultural watersheds equipped with tile drains [8]. In order to forecast the effects of agriculture management on hydrological behavior in small catchments, a distributed model system was applied which consist two parts, a 1D SVAT model and the KINEROS

model [9].

So far distributed hydrologic models have been successfully used in many natural basins [3; 4; 10], but only a few studies have been conducted to take into account the specific characteristics of man-made farmed basins in hydrological modelling [11; 12]. Though there are many agro-hydrological models for irrigation district water demand assessments at field scale, which consider soil texture and crop conditions influenced in the hydrological processes simulation [13-15]. However, such models were not designed for basin scale water cycle simulation. Surface, groundwater flow and interactions between them were usually omitted in these models.

As the natural hydrology, behaviours were greatly changed by human activities in agriculture basins, Human effects such as irrigation strategies, and the arrangements of ditch drains should be integrated in distributed hydrologic model system. In this paper, we developed a hydrologic model for agricultural basin water resources management (RDG) by combining a GIS module. The specific objectives in this paper were to introduce the general structure, data demand, spatio-temporal organization of modules and theoretical background.

2. General Description of RDG Model

The RDG model is a physically-based, integrated distributed hydrologic model. Nearly all the parameters can be determined by field experiments or according to their physical meanings. It is also a continuous simulation model for agricultural water management evaluation.

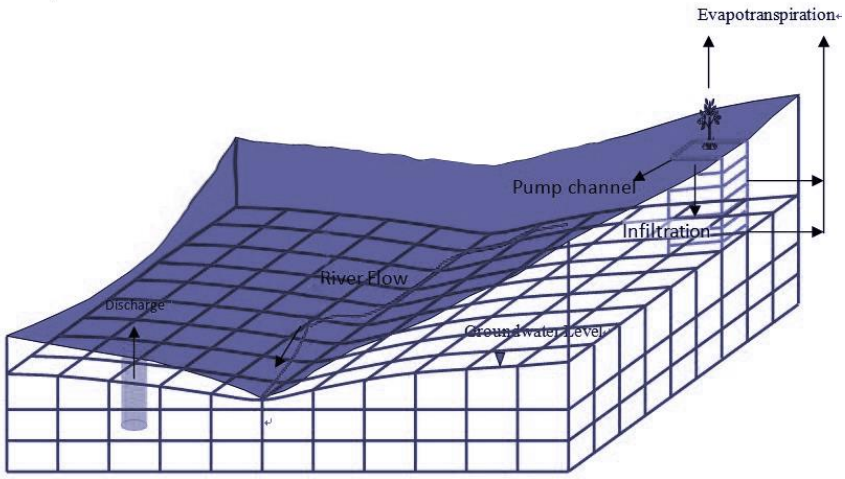


Figure 1. Hydrological processes depiction and basin discrete for RDG model

In Fig.1, the hydrological processes simulated in RDG model, and the scheme for basin discretization was shown. It can be seen from Fig.1 that hydrological processes such as vegetation interception, evapotranspiration, unsaturated soil flow, groundwater flow etc., are considered in the model. As depicted in Fig.1, the basin is subdivided into limited volumes in spatial domain to represent the continuity and heterogeneity of water flow in different medium. In horizontal direction, a basin is subdivided into a ranked grid system with size, e.g. 1 km×1 km for surface flow routing, and the grids are defined as field and channel by a drainage network digitizing method. A simply example for the conceptualization of the field and channel grids over a 19 rows×19 columns DEM were shown in Fig. 2, and there were 12 possible outlets (grids with grey dots) in the raster network. The arrow direction in each grid represents the flow direction. The grids with slight arrows represent fields, grids with light thick lines represent channel segments, and there are some null grids on the boundary of the DEM. Surface flow in fields and channels is routed by a kinematic wave approximation solution. In vertical direction, the vadose zone is subdivided into soil-vegetation column units, SVCUs (see also in Fig. 1), and the column is further divided into several layers according to soil texture in each unit. 1D infiltration and soil flow model is used for soil water flow simulation. Like the surface zone, the shallow saturated zone is also subdivided into grid system according to aquifer characteristics. In each SVCU, recharge between soil water and groundwater is determined by water potential gradient. Especially, recharge between channel water and groundwater should be considered within river grids.

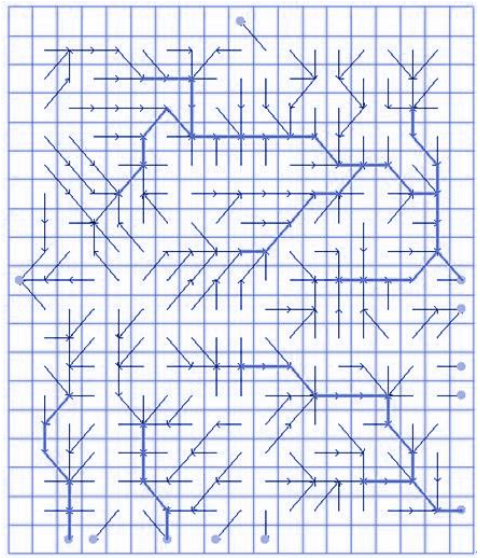


Figure 2. The conceptualization of the field and channel grids

The model contains four different modules, namely, the Input Module, Pre-processed Module, Integrated Distributed Hydrologic Module and the Output Module. Input data of the model includes hydrology, climate, land use, topography, aquifer and agricultural practices, etc. Before the model is running, a pre-processed module is performed and the spatial domain is conceptualized, ranked and subdivided into grid system in horizontal and vertical direction. The Integrated Distributed Hydrologic Module computes the surface and subsurface hydrological processes and yields the flow fluxes to the channel networks, soil layers and the underlying unconfined groundwater aquifer.

The dynamic interaction is also considered at any set point of the watershed. Hydro-meteorological data, i.e. rainfall or temperature are input to drive the model. To link each model components together, we adopted a type of time-lagging scheme. In this scheme, all the components are calculated individually and linked by flow fluxes that are determined according to the initial state values. The internal fluxes are calculated according to the initial state value, and then the flow will be routed in each sub-model. The results will provide initial state value for the next time step and the final output results include stream flow hydrograph, soil moisture, groundwater head, etc. at any set point within a basin.

2.1. Vegetation Interception and Evapotranspiration

Absorption of water by roots will reduce the soil moisture in the soil zone that have crop in it. The transpiration rate of plant t at one moment is the sum of all the Absorption of water by the root zone, like:

$$E_T = \int_0^{z_r} s(z,t)dz \tag{1}$$

In this expression, S(z,t), as the absorption rate of the plant root, is the water absorption of the soil-root in volume per unit time, it changed with the time and soil depth. E_T is the plant transpiration rate and Z_r is the root depth.

Nowadays, the workflow technology is becoming more and more perfect, which is applied to many fields. However, it does not have common and clear. WFMC is defined as the whole or part of process of business to be automatic. During the process of achieving some business goals, the files, messages and tasks will be delivered by ruler to make workmates work together better. That is, workflow is a series of interrelated, automated or semi-automated business activities and processes of implementation [2-3].

2.2. Richard Model

Richard equation is often used to estimate infiltration in hydrological modelling. In the soil zone which have growing plants, the continuity equation of the moisture of soil-root zone is：

$$\frac{\partial\theta}{\partial t}=\frac{\partial}{\partial z}(K\frac{\partial\Phi}{\partial t}-1)-s(z,t) \tag{2}$$

In this expression, θ(cm^3/cm^3) is volumetric water content of soil; ϕ is The soil suction (cm); z is the z-axis, earth surface is origin area and downward is the positive direction; K(cm/min) is the transmissibility coefficient of soil; s is the water absorption of the plant root.

2.3. Quantification of the Water Movement Parameters of Crop

There are some major information in the water movement parameters of crop, such as the transmission coefficient (K), specific water capacity(C) and rate of water diffusion (D).

a) Quantification of the transmission coefficient (K)

Assuming the movement of soil moisture is stable, the expression (2) will turn to

$$0=\frac{\partial}{\partial z}(K\frac{\partial\Phi}{\partial t}-1)-s(z,t) \tag{3}$$

normally, assuming $K=K_0$(K_0 is saturated transmission coefficient, α is empirical coefficient), the expression (2) will turn to (4):

$$\frac{1}{\alpha}\frac{d^2K}{dz^2}-\frac{dK}{dz}=s(z,t). \tag{4}$$

If integrate the both side of the equation in this expression (4)(from 0 to Z_r) , the result will like:

$$(\frac{1}{\alpha}\frac{dK}{dz}-K)\Big|_0^{Z_r}=\int_0^{Z_r}s(z,t)dz. \tag{5}$$

When earth surface is the origin of the coordinate, the expression (2) will turn to (6):

$$(\frac{K}{\alpha}\frac{\partial K}{\partial z}-1)\Big|_{z=0}=E, \tag{6}$$

where E is the water flux density in ground surface, upward is the positive direction, thus

$$\frac{1}{\alpha}\frac{dK}{dz}-K-E=\int_0^{Z_r}s(z,t)dz. \tag{7}$$

Take expression (1) substitute into expression (7), the result is expression (8), as:

$$\frac{1}{\alpha}\frac{dK}{dz}-K=E+E_T. \tag{8}$$

It means the unsaturated transmission coefficient of crops (K) meet with expression (8), The general solution of expression (8) is $K=C^{'}e^{2D}-(E+E_T)$.

While z=D (ground water depth), $K=K_0$ (saturate transmission coefficient), thus:

$$C^{'}=(K_0+E_T+E)e^{-\alpha D}.$$

Therefore, the calculation formula of crop zone transmission coefficient is:

$$K=(K_0+E_T+E)e^{-\alpha(z-D)}-(E+E_T). \tag{9}$$

In the expression (9), K_0 is saturated transmission coefficient, E_T is transpiration rate, α is empirical constant; z is any depth counting start from ground surface; E is surface flux, that is evaporation of soil surface, calculate by the two expression (10) and (11) below:

$$E=(1-veg)\frac{\rho_a}{R_a}(h_a q_{sat}(T_s)-q_a) \tag{10}$$

$$\begin{cases}\Phi_u=\frac{1}{2}[1-\cos(\frac{w_g}{w_{fl}}\pi)], if: w_g<w_{fl}\\ \Phi_u=1, if: w_g\geq w_{fl}\end{cases} \tag{11}$$

In the expression, veg is vegetation coverage; ρ_a is air density; R_a is aerodynamic drag; $q_{sat}(T_S)$ is

the saturation specific humidity at TS temperature; q_a is air specific humidity, is relative humidity of earth's surface; W_g is the average volumetric water content of the surface layer (0-5cm)(cm^3/cm^3), W_{fl} is field capacity.

b) quantification of specific water capacity

The characteristic curve of soil moisture affect by some factors like texture, structure, etc. especial the influence of texture is the most obvious. Normally, the soil clay content is higher, the soil moisture rate is larger and the slope of the curve is more gently in the same condition of water potential.

Specific water capacity can be express as its definition like:

$$C(\theta) = \frac{d\theta}{d\varphi_m} = -\frac{d\theta}{dh} \tag{12}$$

In the expression above, C(θ) is specific water capacity; θ is volumetric water content (cm^3/cm^3); φ_m is soil matric potential， h is suction (negative value of soil matric potential).

The expression of soil moisture characteristic surface is like:

$$\theta_m(\rho_0, h) = a\rho_0^{-m}h^{-n} \tag{13}$$

In the expression (13), h is suction (negative value of soil matric potential); a, m, n are the fitting coefficients; θ_m is mass water content (g/g), ρ_0 is soil capacity (g/cm^3).

Expression (13) is used for describe the characteristic of soil moisture capacity, that is to confirm the relationship of spatial curved surface by three variables.

At present, the relationship between matric potential of soil water and water content cannot fitting to an empirical equation by the measured results.

Due to the curve of soil moisture capacity, $\theta_m(\rho_0,h)$ hydraulic conductivity (K) and specific water capacity C(θ), according to the Richard equation, it is possible the analysis of the law of soil water movement effectively and irrigate the crops in a higher efficiency.

2.4. Recharge Models

With the coordinate system positively downwards and according to the Darcy-Buckingham's flux law, recharge between groundwater flow and unsaturated soil water at the interface of unconfined water table can be derived as

$$UG = -D(\frac{\theta_u + \theta_s}{2})\left(\frac{\partial\theta}{\partial z}\right)_{if} + K(\frac{\theta_u + \theta_s}{2}) \tag{14}$$

and

$$\left(\frac{\partial\theta}{\partial z}\right)_{if} = \frac{\theta_s - \theta_u}{Z_g - Z_u}, \tag{15}$$

where Z_g(cm) is depth of groundwater table; Z_u(cm) is depth of mid-point of the affected layer;
D (cm^2/s) is diffusion rate, K (cm/s) is unsaturated hydraulic conductivity and can be determined by

$$K(\theta) = K_s(\frac{\theta}{\theta_s})^{2b+3}, \tag{16}$$

$$h(\theta) = h_b(\frac{\theta}{\theta_s})^{-b}, \tag{17}$$

where h_b(cm) is air entry value.
And diffusion rate is

$$D(\theta) = \frac{K_s h_b b}{\theta_s} \cdot (\frac{\theta}{\theta_s})^{b+2}, \tag{18}$$

where K_S, h_b, b, θ_s are functions of soil types [16].

The Darcy's law about the conceptualization of the aquifer-channel dynamic interaction in channel grids is used to determine recharge rate between groundwater and channel flow and the equation, as

$$CG = K_r A_r \frac{(h_g - h_r)}{d_r}, \tag{19}$$

Where CG (m^3/s) is recharge rate, the symbol means leakage from channel to groundwater while means leakage from groundwater to channel; $A_r(m^2)$ is wetted area of channel segment; $h_g(m)$ is groundwater table; $h_r(m)$ is channel water level; d_r (m) is thickness of streambed; K_r (m/s) is hydraulic conductivity of streambed.

2.5. Groundwater Flow Model (Darcy Law of the Water Flow)

In the care of density-independent flow in a uniform homogeneous aquifer (20) is considered in this work:

$$C(\theta)\partial_t\phi = \frac{k}{r^{n-1}}\partial_r[r^{n-1}\partial_r\phi] \quad (20)$$

,

where both the specific water capacity C(θ), and hydraulic conductivity K are scalar and constant quantities and n=1,2 or 3 is the radial dimension. To be complete the flowing set of initial and boundary conditions is added.

$$\phi(r,0) = \phi_0, \quad \phi(r,t)\big|_{r\to r_{w0}} = \phi_0, \quad (21)$$

$$Q = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} r_w^{n-1} K d^{3-n} \partial_r \phi(r_{w,}t), \quad (22)$$

where in:

- Q is the discharge rate of a borehole, with radius r_w.
- ϕ_0 is initial saturated water content.
- d the thickness of the aquifer from which the borehole taps water is the initial pierometrie head in the aquifer.
- r_{wo} is the radius of the land for crops. Boundary conditions is the content of crop land to satisfy the initial saturated water, and after a period of time, r_{wo} position reached the same saturated water content.

3. Conclusion Combination of GIS with the RDG Model

Integrated distributed hydrologic models like the RDG divided the basin into a raster network to account for the spatial variability of basin physical properties and allow for the spatial assessment of modelled hydrological variables. The progresses in geographic information systems (GIS), remote sensing technology (RS), and computer science have made a vast amount of spatially distributed data processing and distributed hydrological modelling possible. Because the RDG model parameters were related to the physical properties and attributes of the watershed, it is possible to estimate the geomorphologic, vegetation, soil, aquifer parameters by means of the existing GIS and RS data sets. Moreover, GIS will further enhance the modelling processes and the effective uses of modelling results. In fact, GIS provides a tool for watershed hydrologic model advancing from lumped model to distributed hydrologic model. Compared the basic input/output functions supported by GIS, basin delineation will provide a pre-processed module for RDG model. As distributed hydrological models require detailed definitions of a basin's internal drainage structure, terrain analysis based on DEMs such as flow vector definition, sub-basin delineation and raster slope extraction were always included in distributed hydrologic models. For example, the ANSWERS model uses a slop map and the drainage network computed by the PC RASTER package [17] as inputs. The AGNPS model also uses drainage network as inputs just like ANSWERS model and the GIS-assisted computer program, TOPAZ [18], is used to extract the terrain parameters. The TOPMODEL uses flow directions to compute the topographic index. Other hydrologic models such as SWAT [19] and Hydrotel [20], etc., all include terrain analysis modules within them.

Since the RDG model was developed to apply in agricultural basin, the model has its unique parameters set and basin delineation characteristics. Compared with natural basins, the hydrological processes as well as basin physical attributes have been greatly changed accordingly. Compared with natural drainage networks, water flow in the ditch networks does not necessarily follow the steepest slope of the surface topography. Therefore, specially designed software package, DigitalHydro, for automatic extraction of digital drainage network had been developed to assemble the GIS database, extract channel

network, and delineate the basin boundary from DEM and other map information. Particularly, an algorithm was applied to delineate the digital drainage network for basins with intense human activities by using multiple map information such as river, lake, and dam or watershed boundary other than DEM. The map was digitized and stored by a raster matrix in this method. Two elevation - distance functions were used to adjust the DEM according to the ratter map information. This new approach allows for an accurate fit between the map and the modelled drainage structure. The proposed approach will benefit the distributed hydrological models that require data such as a flow direction matrix, river, lake and sub - watersheds for drainage structure information.

4. Conclusions

Integrated surface- subsurface water development is inevitable to improve water use rate and achieve sustainable development of water resources. To ensure sustainable developments and provide a tool for water resources management of agricultural basins in HHH Plain, an integrated distributed hydrological model termed RDG has been developed.

In the RDG model, a spatial grid structure is used for represent the continuity and heterogeneity to simulate spatially variable water processes in basin with complex landscape characteristics. The model presents some newly - developed approaches to the modelling of hydrological processes in order to quantify and incorporate the effect of agricultural practices. Towards this purpose, basin is subdivided into SVCUs in vertical direction. A layered infiltration model, GALS, is used to account the characteristics of layered soils. In order to consider ditch management effects to flow directions, digital drainage network is extracted based on DEM and other map information in horizontal direction. The continuous model simulates most of the hydrological processes such as interception, evapotranspiration, infiltration, unsaturated soil flow, groundwater flow,, channel flow, recharges between channel flow and groundwater flow, recharges between unsaturated soil flow and groundwater flow and discharge of groundwater pumped for agriculture, etc., as depicted in Fig. 1.

The RDG model is an evolving study and significant further works will be carried out to take important agricultural characteristics into account in the RDG model.

Acknowledgements

This research was funded by the MWR public sector research and special funds - the most stringent in arid zone water resources management key technologies(201301103) and by a grant from the National Natural Science Foundation of China(Grants Nos. 41201025, 70361001, and 40871023). This financial support is gratefully acknowledged and appreciated.

References

1. ASCE (1985) Task committee on quantifying land-use change effects of the watershed management and surface-water committees of the irrigation and drainage division. Evaluation of hydrologic models used to quantify land-use change effects. *Journal of Irrigation and Drainage Engineering*, 111(1), 1-17.
2. Freeze, R.A.L., Harlan, R. (1969) Blueprint for a physically-based digitally-simulated hydrological response model. *Journal of Hydrology*, 9(3), 237-258.
3. Abbott, M.B., Bathurst, J.C. Cunge, J.A. (1986) O'Connell P.E., Rasmussen J. An introduction to the European hydrological system–Système Hydrologique Européen, SHE, 2: structure of a physically-based, distributed modelling system. *Journal of Hydrology*, 87(1), 61-77.
4. Beven, K.J., Calver, A., Morris, E.M. (1987) *The Institute of Hydrology Distributed Model*. Institute of Hydrology Report 98.
5. Gunduz, O., Aral, M.M. (2005) River networks and groundwater flow: a simultaneous solution of a coupled system. *Journal of Hydrology*, 301(1-4), 216-234.
6. Troch, P.A., Paniconi, C., van Loon, E.E. (2003) Hillslope-storage Boussinesq model for subsurface flow and variable source areas along complex hillslopes: 1. Formulation and characteristic response. *Water Resources Research*, 39(11), 13-16.
7. Arnold, J.G., Allen, P. M. (1993) Bernhardt G. A comprehensive surface-groundwater flow model. *Journal of Hydrology*, 142(1), 47-69.

8. Kim, S.H., Delleur, J.W. (1997) Sensitivity analysis of extended TOPMODEL for agricultural watersheds equipped with tile drains. *Hydrological Processes*, 11(9), 1243-1261.
9. Bormann, H., Diekkrüger, B., Hauschild, M. (1999) Impact of landscape management on the hydrological behavior of small agricultural catchments. *Physics and Chemistry of the Earth (B)*, 24(4), 291-296.
10. Grayson, R.B.D., Moore, I.A., Mcmahon, T. (1992) Physically based hydrologic modelling, 1. A terrain-based model for investigative purposes. *Water Resources Research*, 28(10), 2639-2658.
11. Dunn, S.M., Mackay, R. (1996) Modelling the hydrological impacts of open ditch drainage. *Journal of Hydrology*, 179(1), 37-66.
12. Moussa. R., Voltz. M., Andrieux. P. (2002) Effects of the spatial organization of agricultural management on the hydrological behaviour of a farmed catchment during flood events. *Hydrological Processes*, 16(2), 393-412.
13. Wöhling, T.H., Schmitz, G. (2007) Physically based coupled model for simulating 1D surface-2D subsurface flow and plant water uptake in irrigation furrows. 1: model development. *Journal of Irrigation and Drainage Engineering*, 133(6), 538-547.
14. Li, Y. R., White, L., Chen, D.B., Zhang, J.G., Li, B., Zhang, M., Huang, Y.F., Edisa, Y.R. (2007) A spatially referenced water and nitrogen management model (WNMM) for intensive cropping systems in the North China Plain. *Ecological Modelling*, 203(3-4), 395-423.
15. Minacapilli, M., Lovino, M.D. (2008) A distributed agro-hydrological model for irrigation water demand assessment. *Agricultural Water Management*, 95(2), 123-132.
16. Cosby, B.J.M., Hornberger, G.B., Clapp, R.R., Ginn, (1984) T. A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. *Water Resources Research*, 20(6), 682-690.
17. van Deursen,W.P.A., Wesseling, C. (1992) *The PC RASTER package*. Department of Physical Geography: The University of Utrencht, The Netherlands.
18. Martz, L.W., Garbrecht, J. (1992) Numerical definition of drainage network and sub catchment areas from digital elevation models. *Computers & Geosciences*, 18(6), 747-761.
19. Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R. (1998) Large area hydrologic modelling and assessment. Part I: Model development. *Journal of the American Water Resources Association*, 34(1), 73–89.
20. Fortin, J.P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J. (2001) A distributed water-shed model compatible with remote sensing and GIS data, part 1: description of the model. *Journal of Hydrologic Engineering*, ASCE, 6(2), 91-99.

Received on the 3rd of November 2013

Mathematical and Computer Modelling

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 37-43
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

TIBETAN TEXT SIMILARITY COMPUTATION BASED ON ONTOLOGY

G.Xu[1,*], L. Qiu[1], Q. He[1]

[1]College of Information Engineering, Minzu University of China, Beijing, 100081, China
[*] Corresponding author E-mail address: xuguixian2000@sohu.com

Tibetan information processing technology has made some advances. However, it falls behind Chinese and English information processing and still need to be more studied. The construction of Tibetan semantic ontology can enrich the Tibetan digital semantic resource. The semantic study based on the Tibetan Ontology is important and valuable. In this paper, an approach of Tibetan text semantic similarity computation based on Ontology is proposed. The approach is to firstly calculate the semantic relationships between concepts based on Tibetan Ontology and then use the concept semantic similarity and features' weight to compute the text semantic similarity. The experimental results present the proposed approach can effectively measure the similarity between texts. Making full use of ontology semantic relationship will greatly enhance the Tibetan information processing's accuracy.

Keywords: Tibetan Information Processing, Semantic Ontology, Concept Similarity, Text Semantic Similarity.

1. Introduction

As the part of language information processing, the semantic framework has been proposed for a long time in Chinese and English information processing. Tibetan semantic study is fewer than Chinese and English. Tibetan semantic research is meaningful to understand the relationship and connection of Tibetan texts. Tibetan text similarity computation can make more people to achieve the useful Tibetan information.

Concept semantic similarity calculation occupies an important place in today's rapid development of science and technology. It is the premise of the text semantic similarity computation. It has the essential value and significance for different language information processing and the research. It has been widely used in information retrieval, text classification, and natural language processing and many other areas. Many researchers focus on the concept semantic similarity calculation [1-3]. However, the accuracy of concept semantic similarity needs to be improved. The text similarity computation based on understanding ontology has a higher accuracy. Agirre used WordNet to conduct the similarity computation so that words disambiguation was studied [4]. Che adopted the synonym lexicons to conduct the sentence similarity computing research [5]. Liu [6] and Li [7] utilized the knowledge structure of HowNet to conduct the word similarity and sentence similarity studies. These methods were needed to improve the accuracy of the text similarity computation approach. Based on the semantic characteristics of the semantic web, Ganesan proposed a kind of text similarity algorithm [8]. However, he only considered the semantic relationship between words and words and ignored the characteristics of the words themselves, such as the frequency and the location of words. It reduced the accuracy of similarity computation.

Tibetan information processing falls behind Chinese and English information processing. Tibetan semantic computation research is fewer. The construction of Tibetan semantic ontology can enrich the Tibetan digital semantic resource. Making full use of ontology semantic relationship will greatly enhance the Tibetan information processing's accuracy. Furthermore, the effective similarity computation method should be proposed. In this paper, we will study on the text similarity computation based on tibetan ontology. The result of research can accelerate the development on the information retrieval and text classification and clustering. In the following, we first introduce the related background. The proposed approach is described next. We then present the experimental results, and conclude our work.

2. Background

The technology of the semantic text processing is the key to dispose and organize a mass of text data. It is convenient for the users to find the interesting information they need accurately. Semantics is

the interpretation of the data symbols. The core of semantics is knowledge. Ontology is an explicit formal specification of shared conceptual model, which is used to describe the domain-specific knowledge [9].

Many scholars studied on the text similarity calculation. In the traditional text information processing, it was general to use cosine similarity to measure the similarity of the two texts [10]. This method ignored the semantic relationship of terms.

[11] proposed text semantic similarity calculation based on weighted semantic Web and Hownet. It mainly considers three factors: semantic distance, semantic coincidence degree and hierarchical relationships of concept in the Ontology tree. S_i and S_j are the according concepts of features T_i and T_j. $D(S_i,S_j)$ represents the semantic distance of two concepts S_i and S_j. $C(S_i,S_j)$ represents the semantic coincidence degree of two concepts S_i and S_j. H_i and H_j represent the hierarchical relationships of two concepts S_i and S_j.

The feature's semantic similarity is as follows:

$$Sim(T_i,T_j)=Sim(S_i,S_j)=\frac{C(S_i,S_j)}{D(S_i,S_j)\times(|H_i-H_j|+1}. \tag{1}$$

When calculating texts' similarity, [24] considered synthetically the semantic similarity of text features and the importance of features in the text. The equation calculating texts' similarity is as follows:

$$Sim(D_i,D_j)=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{Sim(S_i,S_j)(P_i+P_j)}{m\times n\times(|P_i-P_j|)}. \tag{2}$$

$Sim(D_i,D_j)$ represents the similarity of texts D_i and D_j. m, n are features number of texts D_i and D_j. $Sim(T_i,T_j)$ represents the semantic similarity of features T_i and T_j. P_i and P_j are the percentages of the weight of features T_i and T_j in its text. The problem of this approach is that $D(S_i,S_j)=0$ in formula (1) and $P_i-P_j=0$ in formula (2) could be happened.

[12] calculated the text similarity combined with lexical features' semantic information and TF-IDF method. Let v_i,v_j are the vectors of two different texts.

$$v_i=(\omega_{i1},\omega_{i2},\omega_{i3},...,\omega_{im}), \tag{3}$$

$$v_j=(\omega_{j1},\omega_{j2},\omega_{j3},...,\omega_{jn}). \tag{4}$$

Define text similarity is as follows:

$$TextSim(v_i,v_j)=w_f\times VectSim(v_i,v_j). \tag{5}$$

In the above equation, $VectSim(v_i,v_j)$ represents similarity between vectors v_i and v_j, w_f represents the semantic factor of similarity computation based on the proportion of the feature's TF-IDF weight more than the threshold between the two vectors v_i and v_j. This approach only considered the semantic information with the higher features' weighting.

In the following, we will propose the text similarity computation approach considering five elements.

3. The Proposed Approach

Based on the Ontology tree, we propose the approach to calculate the text semantic similarity. It is used to the proposed approach and based on the concept semantic similarity that was proposed by ourselves in [13]. Now we introduce the concept semantic similarity in [13]. It considered four aspects. Let S_i and S_j express two concepts of Ontology. $Sim(S_i,S_j)$ expresses concept semantic similarity of S_i and S_j. Computation formula of $Sim(S_i,S_j)$ is as follows:

$$Sim(S_i, S_j) = \frac{C(S_i, S_j) * E(S_i, S_j)}{(D(S_i, S_j) + \beta) * \left(\left|H_i - H_j\right| + \gamma\right)}. \quad (6)$$

Here, $C(S_i, S_j)$ means the semantic coincidence degree and expresses the common ancestor node number of S_i and S_j in the Ontology tree. $D(S_i, S_j)$ means the semantic distance and expresses the shortest path length of connecting S_i and S_j of Ontology.

$$E(S_i, S_j) = \sqrt{1 - \frac{|M_i - M_j|}{M_i + M_j}}, \quad (7)$$

where M_i and M_j are the node densities of S_i and S_j. They are the children node numbers of S_i and S_j in the Ontology tree. H_i and H_j are the concept hierarchies of S_i and S_j in the tree. β and γ are used to prevent the denominator to be 0.

Now we introduce the proposed calculation of texts' semantic similarity. It considers two aspects: the concept similarity and the weight relationship of concepts. Assume D_i represents the i-th text's semantic vector space, and D_j represents the j-th text's semantic vector space. $D(S_i, S_j)$ is used to represent the similarity between D_i and D_j. It is calculated by the formula as follows:

$$Sim(D_i, D_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{Sim(S_i, S_j)(W_i + W_j)}{m \times n \times \left(\alpha + \left|W_i - W_j\right|\right)}. \quad (8)$$

In the above formula, m and n represent the number of concepts in the D_i and D_j respectively. S_i is one concept in the D_i, and S_j means one concept in D_j. W_i is the normalized weight of S_i in its text. W_i is the normalized weight of S_j in its text. From the formula we can see that the bigger the concept similarity is, the more similar the text similarity is. The smaller the difference between W_i and W_j is, the more similar the importance of concepts S_i and S_j respectively in the texts D_i and D_j is. That's to say, $\left|W_i - W_j\right|$ is in the inverse proportion to the text similarity. α is the smoothing factor which prevents $\left|W_i - W_j\right|$ from equalling zero.

X is the non-normalized weight of the concept in the text. X's value is TF (Term frequency) in the text. X can be normalized to W by the formula as follows:

$$W = \begin{cases} \dfrac{x - x_{\min}}{x_{\max} - x_{\min}}, x \neq x_{\min} \\ \dfrac{x - x_{\min} + 0.5}{x_{\max} - x_{\min}}, x = x_{\min} \end{cases}. \quad (9)$$

4. Experiment and Results

We build up two simple Tibetan ontology, one is the computer's ontology, the other is the animality ontology. Based on these two ontologies, we conduct the conceptual similarity experiments and text similarity computing experiments. These two ontologies are showed as Figure 1 and Figure 2. We use seven texts as the experiment texts. Every text includes 10 feature words which are the concepts in the ontologies. The texts from d_1 to d_7 are showed in Figure 3. TF means the term frequency of the feature. N-TF means the normalized weight of TF. For the concept similarity computation, the value range of β, γ is from 0.5 to 1. Two groups of concept similarity values based on computer ontology are showed in Table 1 $(\beta = 0.5, \gamma = 0.5)$ and Table 2 $(\beta = 1, \gamma = 1)$. It can measure the concept similar degree effectively. These results are required when we calculate text similarity.

For the text similarity computation, the value range of α is from 0.1 to 0.5. We calculate the semantic similarity value of d_1 and d_1, d_2,…, d_7 at four groups of different α, β, γ value. The results are shown in Table 3.

Observing the data in Table 3, we can see that the similarity trend is same. This means the reasonableness of the parameter range of α, β, γ in the proposed approach. The similarity between d_3 and d1 is higher than that between d_4 and d_1. Although the d_3 and d_1, d_4 and d_1 respectively have eight same features, the normalized weight distribution of the same features between d_3 and d_1 is closer than between d_4 and d_1. Thus, the above results are generated reasonably. d_5 and d_1, d_2 and d_1 respectively have five same features. However, the same features have bigger weights between d_5 and d_1 than that between d_2 and d_1. Therefore, similarity of d_5 and d_1 is bigger than that of d_2 and d_1. The similarities between d_3 and d_1, between d_4 and d_1 are higher than that between d_2 and d_1. Because d_3 and d_1, d_4 and d_1 have eight same features, while d_2 and d_1 have only five same features. That is to say, the co-occurence rate of features in the former is higher than in the latter. Therefore, the above result is reasonable. Similarly, similarity between d_6 and d_1 is less than that between d_5 and d_1. That's because the same features number of the former are two but the latter have five, what's more, the latter's same features contain the same features of the former. The similarity of d_7 and d_1 is about two times of d_5 and d_1. The number of same feature between d_5 and d_1 is more than d_7 and d_1, but the weight of features, which occupies dominant position effects the text similarity largely than the concept similarity among the features.

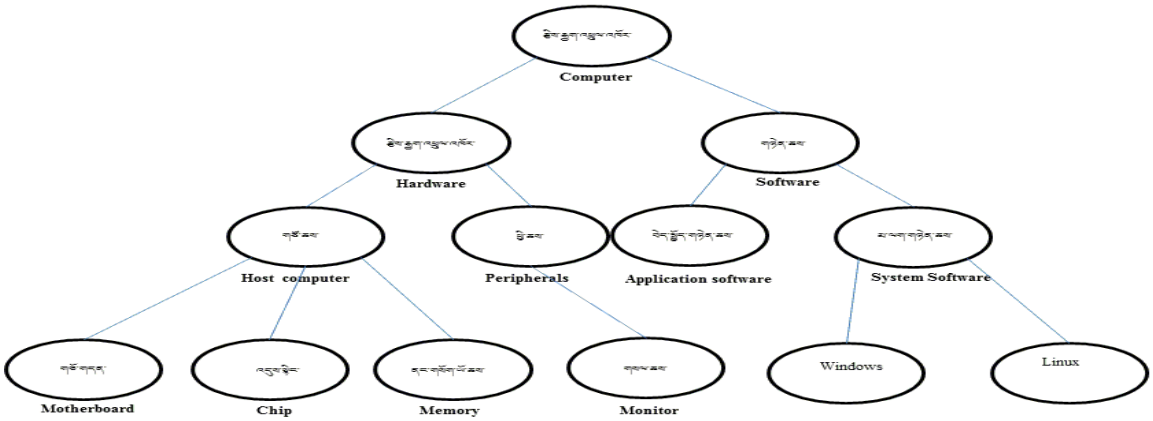


Figure 1. Computer Ontology

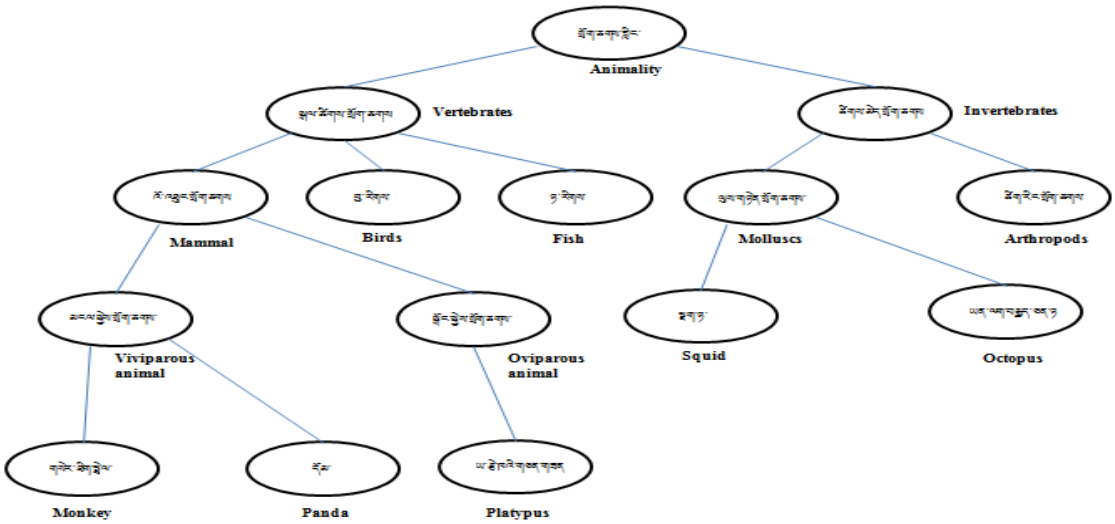


Figure 2. Animality Ontology

d_1			d_2			d_3			d_4			d_5			d_6			d_7		
Feature words	TF	N-TF	Feature words	TF	N-TF	Feature words	TF	N-TF	Feature words	TF	N-TF	Feature words	TF	N-TF	Feature words	Weigh	N-TF	Feature words	TF	N-TF
ཇོལ་ཆུག་འཕྲུལ་འཁོར་ (Computer)	3	0.0556	ཨིན་གུང་ཨེད་སྐྱོན་མེད་ (Invertebrates)	7	0.29	ཇོལ་ཆུག་འཕྲུལ་འཁོར་ (Computer)	3	0.06	ཇོལ་ཆུག་འཕྲུལ་འཁོར་ (Computer)	4	0.11	ཇོལ་ཆུག་འཕྲུལ་འཁོར་ (Computer)	3	0.06	ཇོལ་ཆུག་འཕྲུལ་འཁོར་ (Computer)	3	0.06	གཙོ་མས་ (Host computer)	3	0.06
བཤམས་མས་ (Monitor)	4	0.1111	ལྕང་ཨིན་གུང་སྐྱོན་ (Vertebrates)	6	0.14	བཤམས་མས་ (Monitor)	4	0.11	བཤམས་མས་ (Monitor)	7	0.44	བཤམས་མས་ (Monitor)	4	0.11	བཤམས་མས་ (Monitor)	4	0.11	རང་གཤོག་ཡོ་མས་ (Memory)	4	0.11
མ་ཡག་གཏེར་མས་ (System Software)	5	0.2222	ཕྱི་མས་ (Peripherals)	5	0.07	མ་ཡག་གཏེར་མས་ (System Software)	5	0.22	མ་ཡག་གཏེར་མས་ (System Software)	6	0.33	མ་ཡག་གཏེར་མས་ (System Software)	5	0.22	འཕྲིན་མས་ (Hardware)	5	0.22	ཉེར་སྤྱོད་གཏེར་མས་ (Application software)	5	0.22
ལྕང་ཨིན་གུང་སྐྱོན་ (Vertebrates)	6	0.3333	གཏེར་མས་ (Software)	6	0.14	ལྕང་ཨིན་གུང་སྐྱོན་ (Vertebrates)	6	0.33	ལྕང་ཨིན་གུང་སྐྱོན་ (Vertebrates)	5	0.22	ལྕང་ཨིན་གུང་སྐྱོན་ (Vertebrates)	6	0.33	ལྕང་གཏེར་གུང་སྐྱོན་ (Molluscs)	6	0.33	windows	6	0.33
གུང་ལྕགས་གུང་སྐྱོན་ (Oviparous animals)	7	0.4444	འཕྲུལ་ཁྱིང་ (Chip)	11	0.86	ཨཀ་ཡག་བཙན་ནག་ (Octopus)	9	0.67	ཨཀ་ཡག་བཙན་ནག་ (Octopus)	3	0.06	གུང་ལྕགས་གུང་སྐྱོན་ (Oviparous animals)	7	0.44	ཉེར་སྤྱོད་གཏེར་མས་ (Application software)	7	0.44	ཨཀ་ཡག་བཙན་ནག་ (Octopus	7	0.44
ཨ་རྡེ་ཁའི་གཅན་གཅན་ (Platypus)	8	0.5556	ཕྱི་རིན་པ་ (Birds)	8	0.43	ཨ་རྡེ་ཁའི་གཅན་གཅན་ (Platypus)	8	0.56	ཨ་རྡེ་ཁའི་གཅན་གཅན་ (Platypus)	8	0.56	linux	8	0.56	windows	8	0.56	མམས་ལྕགས་གུང་སྐྱོན་ (Viviparous animal)	8	0.56
ཐྲག་ཀ (Squid)	9	0.6667	ཐྲག་ཀ (Squid)	9	0.57	ཐྲག་ཀ (Squid)	9	0.67	ཐྲག་ཀ (Squid)	9	0.67	ཕྱི་རིན་པ་ (Birds)	9	0.67	ཉ་རིན་པ་ (Fish)	9	0.67	ཨིན་གུང་ཨེད་སྐྱོན་མེད་ (Invertebrates)	9	0.67
ཨིན་རིས་གུང་སྐྱོན་ (Arthropods)	10	0.7778	རྡ་ (Panda)	12	1	རང་གཤོག་ཡོ་མས་ (Memory)	6	0.33	རང་གཤོག་ཡོ་མས་ (Memory)	9	0.67	རང་གཤོག་ཡོ་མས་ (Memory)	10	0.78	མམས་ལྕགས་གུང་སྐྱོན་ (Viviparous animal)	10	0.78	གཙོ་ཉེར་སྤྱོད་ (Monkey)	10	0.78
འཕྲུལ་ཁྱིང་ (Chip)	11	0.8889	ཨ་རྡེ་ཁའི་གཅན་གཅན་ (Platypus)	8	0.43	འཕྲུལ་ཁྱིང་ (Chip)	11	0.89	འཕྲུལ་ཁྱིང་ (Chip)	11	0.89	ཨཀ་ཡག་བཙན་ནག་ (Octopus	11	0.89	གཙོ་གཤོག་ (Motherboard)	11	0.89	འཕྲུལ་ཁྱིང་ (Chip)	11	0.89
རྡ་ (Panda)	12	1	linux	12	1	རྡ་ (Panda)	12	1	རྡ་ (Panda)	12	1	ཕྱི་མས་ (Peripherals)	12	1	ཕྱི་རིན་པ་ (Birds)	12	1	རྡ་ (Panda)	12	1

Figure 3. The features and weights of texts

TABLE 1. The Computer Ontology concept similarity of β =0.5 and γ =0.5

Concept similarity	ཇོལ་ཆུག་འཕྲུལ་འཁོར་	འཕྲིན་མས་	གཏེར་མས་	གཙོ་མས་	ཕྱི་མས་	ཉེར་སྤྱོད་གཏེར་མས་	མ་ཡག་གཏེར་མས་	གཙོ་གཤོག་	འཕྲུལ་ཁྱིང་	རང་གཤོག་ཡོ་མས་	བཤམས་མས་	windows	linux
	Computer	Hardware	Software	Host computer	Peripherals	Application software	System Software	Mother board	Chip	Memory	Monitor		
ཇོལ་ཆུག་འཕྲུལ་འཁོར་ (Computer)	4												
འཕྲིན་མས་ (Hardware)	0.44444	8											
གཏེར་མས་ (Software)	0.44444	0.8	8										
གཙོ་མས་(Host computer)	0.1431	0.7950	0.1703	12									
ཕྱི་མས་(Peripherals)	0.1306	0.7257	0.1555	1.1313	12								
ཉེར་སྤྱོད་གཏེར་མས་ (Application software)	0.1306	0.1555	0.7257	0.3142	0.4444	12							
མ་ཡག་གཏེར་མས་ (System Software)	0.16	0.1904	0.8888	0.3975	0.3628	1.3063	12						
གཙོ་གཤོག་ (Motherboard)	0.0666	0.2612	0.0725	0.9428	0.3809	0.1212	0.0989	16					
འཕྲུལ་ཁྱིང་(Chip)	0.0666	0.2612	0.0725	0.9428	0.3809	0.1212	0.0989	2.4	16				
རང་གཤོག་ཡོ་མས་ (Memory)	0.0666	0.2612	0.0725	0.9428	0.3809	0.1212	0.0989	2.4	2.4	16			
བཤམས་མས་(Monitor)	0.0666	0.2612	0.0725	0.2693	1.3333	0.1212	0.0989	0.8888	0.8888	0.8888	16		
windows	0.0666	0.0725	0.2612	0.0857	0.1212	0.3809	1.0886	0.3076	0.3076	0.3076	0.3076	16	
linux	0.0666	0.0725	0.2612	0.0857	0.1212	0.3809	1.0886	0.3076	0.3076	0.3076	0.3076	2.4	16

Table 4 gives the comparison result of text consine similarity and text semantic classification. Here, we take consine similarity compared with the fourth group data of the Table 3.

In the Table 4, it is noticed that the similarity degree of d_1 and d_1, d_1 and d_2, d_1 and d_3, d_1 and d_4 of the first line data are a little smaller than that of the second line data. The reason is that text semantic similarity formula slightly shrinks consine similarity value of two texts for more same features through the multiplicative factor. The consine similarity of d_6 and d_1 is extremely small than that of the semantic similarity. However, their semantic similarity is obviously improved after considering the semantic relationship of two ontologies. The semantic similarities of d_5 and d_1, d_7 and d_1 are also improved.

The experimental results show that the proposed method considering these two factors of the semantics and the feature's weight is effective.

TABLE 2. The Computer Ontology concept similarity of β =1 and γ =1

Concept similarity	ཇོལ་ཆུག་འཕྲུལ་འཁོར་	འཕྲིན་མས་	གཏེར་མས་	གཙོ་མས་	ཕྱི་མས་	ཉེར་སྤྱོད་གཏེར་མས་	མ་ཡག་གཏེར་མས་	གཙོ་གཤོག་	འཕྲུལ་ཁྱིང་	རང་གཤོག་ཡོ་མས་	བཤམས་མས་	windows	linux
	Computer	Hardware	Software	Host computer	Peripherals	Application software	System Software	Motherboard	Chip	Memory	Monitor		
ཇོལ་ཆུག་འཕྲུལ་འཁོར་ (Computer)	1												
འཕྲིན་མས་ (Hardware)	0.25	2											
གཏེར་མས་ (Software)	0.25	0.3333	2										
གཙོ་མས་ (Host computer)	0.0993	0.4472	0.1118	3									
ཕྱི་མས་ (Peripherals)	0.09072	0.4082	0.1020	0.4714	3								
ཉེར་སྤྱོད་གཏེར་མས་ (Application	0.0907	0.1020	0.4082	0.1414	0.2	3							

software)													
སྒོ་ཡག་གཞུན་ཚན་ (System Software)	0.1111	0.125	0.5	0.1788	0.1633	0.5443	3						
གཙོ་གཡུན་ (Motherboard)	0.0510	0.1814	0.0544	0.5303	0.25	0.0833	0.0680	4					
འཇུག་ཤིང་ (Chip)	0.0510	0.1814	0.0544	0.5303	0.25	0.0833	0.0680	1	4				
རྒྱང་གཞིག་པོ་ཚན་ (Memory)	0.0510	0.1814	0.0544	0.5303	0.25	0.0833	0.0680	1	1	4			
བཤལ་ཚན་ (Monitor)	0.0510	0.1814	0.0544	0.1767	0.75	0.0833	0.0680	0.4	0.4	0.4	4		
windows	0.0510	0.0544	0.1814	0.0589	0.0833	0.25	0.6123	0.1428	0.1428	0.1428	0.1428	4	
linux	0.0510	0.0544	0.1814	0.0589	0.0833	0.25	0.6123	0.1428	0.1428	0.1428	0.1428	1	4

TABLE 3. The results of text similarity

α=0.1 β=0.5 γ=0.5							
	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	16.402	8.6603	13.383	12.434	3.5945	1.3851	7.8016
α=0.1 β=1 γ=1							
	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	4.2474	2.328	3.5196	3.2306	1.0868	0.5903	2.1882
α=0.5 β=0.5 γ=0.5							
	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	3.3878	2.2314	2.7852	2.7619	0.8675	0.4123	1.7551
α=0.5 β=1 γ=1							
	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	0.91	0.6203	0.7611	0.7478	0.2943	0.1899	0.5324

TABLE 4. The result comparison of semantic similarity and Cosine similarity

α=0.5 β=1 γ=1							
	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	0.91	0.6203	0.7611	0.7478	0.2943	0.1899	0.5324
Cosine Similarity							
	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	1	0.6353	0.7888	0.8026	0.2093	0.0388	0.4109

5. Conclusions

In this paper, the approach of the text semantic similarity computation is proposed. It considers feature's weight and concept similarity based on Tebitan ontology. After the experimental verification, the text similarity calculation method proposed can effectively measure the similarity between texts and further improve the rational calculation and accuracy. It has the better perspectives for the Tebitan information processing such as text classification, clustering, and information retrieval and so on.

Acknowledgment

This work is supported by "the National Natural Science Foundation of China (No.61309012)" and "Undergraduate Research and Training Program of Minzu University of China (No.URTP2013110053)".

References

1. Huang, J., Chen, C. (2009) A Physical Approach to Moving Cast Shadow Detection. In Proceedings of 2009 IEEE international conference on acoustics, speech, and signal processing, April 19-24 (2310-2317). Taipei, Taiwan.
2. Wang, J. (2012) Based on semantic similarity Web Text classification research, *Research On Library Science*, No. 9, 65-65.
3. Leone, A. (2007) Distante C. Shadow Detecting for Moving Objects Based on Texture Analysis, *Pattern Recognition*, 40(2), 1222-1233.
4. Agirree, R.G. (1995) A proposal for word sense disambiguation using conceptual distance. In Proceedings of International Conference on Recent Advances in Natural Language Processing (258-264).

5. Che, W., Liu, T., Qin, B. (2003) Facing the dual statement for retrieval of Chinese Sentence Similarity Computing. In Proceedings of the Seventh National Conference on Computational Linguistics (81-88).
6. Liu, Q., Li, S. (2002) The lexical semantic similarity calculation based on HowNet. In Proceedings of the third session of Chinese lexical semantic symposium (59-76).
7. Li, S. (2002) Study on the relevancy between sentences based on semantic computation, Computer engineering and Applications (75-78)
8. Ganesan, P., Garcia, H., Widom, J. (2003) Exploiting Hierarachical Domain Structure to Compute Similarity, ACM Trans on information Systems (64-93).
9. Studer, R., Beniamins, V.R., Fensel, D. (1998) Knowledge engineering: Principles and methods, *Data and Knowledge Engineering*, 25(1), 161-197.
10. Vapnik, V. (1998) Statistical Learning theory, Wiley-Interscience.
11. Liao, K., Yang, B. (2012) Similarity Computing of Documents Based on Weighted Semantic Network, *Journal of Intelligence*, 31(7), 182-186.
12. Huang, C., Yin, J., Hou, F. (2011) A combination of word semantic information and TF-IDF method of text similarity measure method, *Chinese Journal of Computer*, 34(5), 856-864.
13. Xu, G., He, Q., Zhao, X., Tang, R., Zhou, Q., Ding, B., Zhao, X. (2013) Tibetan Concept Similarity Computation Based On Ontology. In proceedings of The 6th International Conference on Intelligent Networks and Intelligent Systems, November 1-3, 2013 (292-295), Shenyang, China

Received on the 3rd of November 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 44-50
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

SURFACE OPTIMIZATION DESIGN ON LAST-BOTTOM USING SPORTS BIOMECHANICS COMFORT

***Zairan Li*[1]**

[1]*Wenzhou Vocational & Technical College*
Wenzhou, Zhejiang, P. R. China, 325035
Phone: (+86)-18958750163. E-mail: lzr_101@126.com

The rationality design on the surface morphology of shoe bottom has a major impact on foot health and users' comfort, and there has a high correlation between gait characteristics with the surface of shoe bottom and the foot. This paper firstly summarizes testing and analysis on the state of the foot biomechanics, and further to control the last-bottom surface morphology through gait parameters, secondly, critical curve optimization method on last body, foot biomechanics test by incorporate personalized last bottom surface design were proposed consequently. Finally, the footwear personalized comfort design on last bottom was realized.

Keywords: last-bottom surface, biomechanics, personalized shoe design

1. Introduction

Shoe last plays a key role in guiding the design of footwear model and design quality for wearing comfort. The body of the last is a three-dimensional surface, while the surface is determined by the regional characteristics of the foot, which is a comprehensive number of gait optimized design; it includes standing, walking, running and jumping, and other condition of foot [1]. In personalized design of shoes and in addition to basic foot length, circumference and other data, as well as characteristics of plantar pressure data analysis, individual gait characteristics needs to be focused. People in their daily life, the axis of the foot and walking (standing) direction to form a fixed angle range and enhanced coordination of various parts of the body functions, such as, physical activity to increases the scope of to improve the body's balance of performance, walking for increasing the force area decreases and the joint angles, and more effort of foot movement due to physiological gait characteristics and other reasons [2].

At present, domestic and foreign personalized shoes for improving comfort research methods are mainly focusing stress-testing analysis, which is to add personalization in the bottom of the shoe, change the shape of the sole or change various parts of the material to achieve, it also aims at the sole components of the design. As we all know, the shoe body shape is copied from the last body surfaces including the bottom of the chamber shoe uppers form and shape [3, 4]. So, for the last design and traditional design approach on last bottom surface relies on the experience of craftsmen, the site of the main considerations are centre point heel and forefoot the place, last body rocker degrees and other factors, and the lack of gait analysis application. Solution to this problem must rely on modern experimental techniques and data analysis techniques. In this paper, the state of motion of the foot corresponding gait data collection and analysis using data analysis and surface modelling techniques to personalize the last bottom was applied for surface parameter optimization, and furthermore, to improve comfort index of personalized shoes [5,6].

2. Foot State Biomechanical Testing and Analysis

Foot is the body upright posture important support, the main structure includes: bones, joints, arch, muscles, skin, etc., different gait characteristics play support the weight, transmit motion, absorbing shock and linkage levers effect. Biomechanics of the foot structure is not only a static standing pillar base, or to meet the body moving mechanical device, but also it is a basic function to maintain the body balance and stability.

2.1. Basic Characteristics Foot Gait

During exercise, joints is a particularly important role, it mainly consistent of the ankle, tarsal joints, tarsometatarsal joints, ametatarsophalangeal joint, and body constituted foot linkage (Shown as Figure 1). The full range of activities in the ankle is approximately 45^o; while the load is about 40^o, the

normal range of dorsiflexion is 20º-30º, and the plantar flexion of the normal range is 30º- 50º. Under normal circumstances, normally, the ankle range while walking is 24.4º, plantar flexion is 14.2º, normal walking barefoot average is 4º of plantar flexion. Ankle rotation axis tilted from the medial malleolus to the lateral malleolus, and the medial side and leg axis angle is 68º - 88º, in the plantar midline horizontal plane angle is 68.5º- 99º.

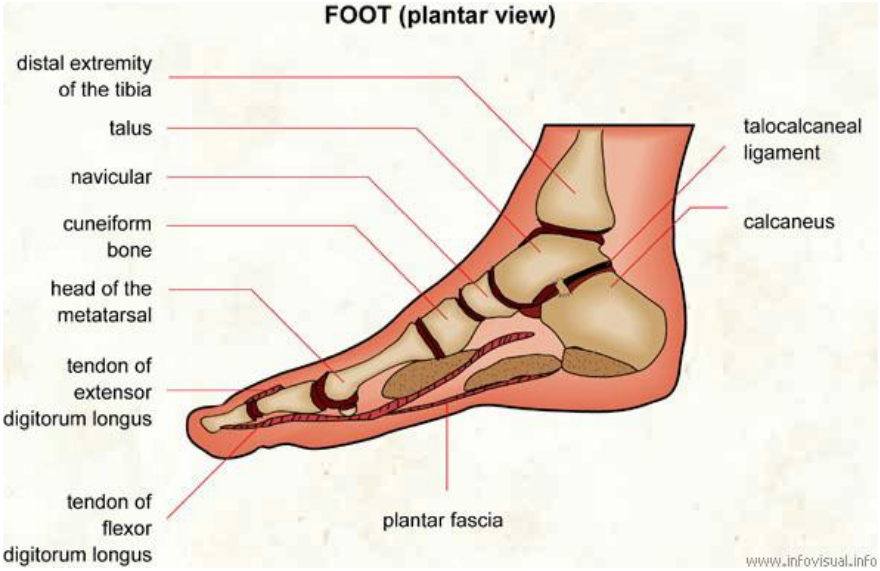


Figure 1. The biometric structure of foot

2.2. Gait Data Test

In order to solve the change of gait characteristics by personalized foot, different heel heights were used for experimental objects to collect gait data. Subjects selection criteria are: women, shoe size 235 #, body is 155 - 165mm, weight is 45 - 55kg; condition is added as follows: more than 40 mm high heels through the experience. Experimental procedure: each experimental subject was wearing heel height: 20mm, 30 mm ... 60 mm high heels and walking through a maximum plantar flexion motion capture three degrees, the maximum longitudinal degrees. Meanwhile, the acquisition of part of the load when plantar flexion, the experimental data obtained is shown in Table 1 below.

There have two First metatarsophalangeal joint motion axes, one is the vertical axis, the horizontal plane adduction and abduction; another is horizontal, do sagittal plane flexion and extension. During the experiment, the degree of longitudinal data collection subjects wearing different shoes with high, middle and front of the body when the plantar angle between the direction of the angle is a positive angle (abduction) or negative angle (adduction), shown in Figure 2.

TABLE 1. Characteristics of different states with pupil data

Index	Data1	Data2	Data3	Data4	Data5
Heel	20mm	30 mm	40 mm	50 mm	60 mm
Plantar flexion	15-20º	25-29º	33-36º	40-43º	46-48º
Longitudinal degrees	3-5ºor 0-3º	7-12ºor -3-8º	8-15º or -5-15º	8-18ºor-3-8º	10-30ºor-2-5º

Figure 2. Schematic vertical axis of foot

2.3. Data Analysis Gait

Walking motion is the process of transferring the heart heel, heel and foot centre point moving action has a direct relationship to the experimental data analysis, the changes brought about by the high heel foot plantar flexion and vertical variations, as shown in Figure 3.

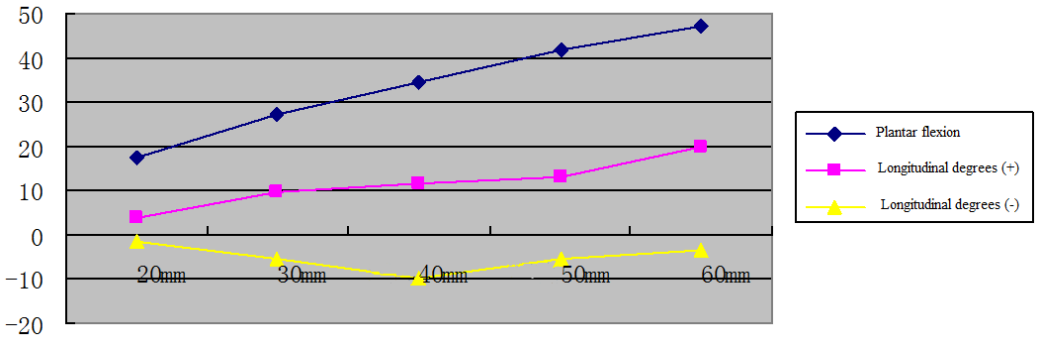


Figure 3. Different states with the pupil characteristic data

2.4. Design of the Axis of the Last Bottom

There has a direct impact on last bottom axis movement into shoes for comfort. The longitudinal direction characteristic features of Last bottom point have six parts: followed by the endpoint (O), last tip point (A), the place of the forefoot (H), the thumb protrudes from point (C), waist dens (G) and heel heart point (M) as the last bottom axis curve characteristic points. In the shoe modelling process of last bottom axis, this six feature points are followed by the endpoint location coordinates the main variable parameter set, and the remaining five were driven by the main variable parameters from varying parameters. Six characteristic points mathematical relationship between established are: the design of the shoe, high heel close the front rocker, last of the forefoot and heel of the base surface should be able to be placed on a flat smooth, followed by the base surface and the plane coincides exactly. Table 2 listed the heel before Rocker relationship of men and women shoes.

TABLE 2. Men and women with high heel shoes before Rocker relations unit: mm

Men	Height of Hell	25	30	35	40	45	50
	Height of front Rocker	17	16	15	14	13.5	13
Women	Former high Rocker	20	30	40	50	60	70
	Height of front Rocker	15	14	13	12	11	10

According to Table 2, the mathematical relationship between front high Rocker and heel can be obtained as:

Front Rocker height = 17 -heel height / 10. MATLAB was used to establish mathematical relationship with the high heel, heel height is ultimately achieved and other features from the point height varying parameters are changing with their last bottom in axis of parametric design. Table 3 shows the different standards compete heel women's shoes in the central axis of the last point height value other features.

TABLE 3. The central axis of the feature points of different high heel height unit: mm

O°	Aa	H	Cc	Gg	Mm
20	15	0	6.4	13.5	17.2
30	14	0	6.0	20.8	25.1
40	13	0	5.6	28.7	33.6
50	12	0	5.1	35.6	40.9

Input Table 3 to MATLAB program as,

```
x=[20 30 40 50 ];
y1=[6.4 6.0 5.6 5.1];
y2=[13.5 20.8 18.7 35.6];
y3=[17.2 25.1 33.6 40.9];
```

Where x is the value of the height of hell Oo, y1is the value of Cc, y2 is the value of Gg, y3is the value of Mm.

(1) 1-order linear fitting
By 1-order linear fitting, the linear relationship between x and y1, y2, y3 is,
y1=7.28-0.43*x //Cc and Oo
y2=-1.32+0.7420*x //Gg and Oo
y3=1.34+0.796*x //Mm and Oo

(2) Second -order polynomial fitting
By 2-order polynomial fitting, we get,
y1=7.005-0.0002*x^2 -0.0255*x
y2=-2.42-0.001*x^2 +0.812*x
y3=-0.31-0.0015*x ^2=0.901*x

In this paper, linear regression equation was applied to establish the last bottom axis, see Table 4.

TABLE 4. Data fitting results, Unit: mm

O°	Cc			Gg			Mm		
	Value	1-order	2-order	Value	1-order	2-order	Value	1-order	2-order
20	6.4	6.42	6.41	13.5	13.52	13.42	17.2	17.26	17.11
S0	6.0	5.99	6.06	20.8	20.94	21.04	25.1	25.22	25.37
40	5.6	5.56	5.66	28.7	28.36	28.46	33.6	33.18	33.33
50	5.1	5.13	5.23	35.6	35.78	35.68	40.9	41.14	40.99

Use the values in Table 4 with respect to each feature point, and then use the spline to connect them, and the main changes in the model parameters (high heel) and other self-varying parameters relationship was finally established, i.e., a high drive followed by a parameter the last bottom axis curve was acquired.

2.5. Gait and Walking Force Point Analysis

Standing in a static normal plantar surface, the weight can be simplified into three points, that is, in the calcaneus, the first metatarsal head of the fifth metatarsal head and the weight composition of the surface as shown in Figure 4. Composed of the calcaneus and metatarsal longitudinal arch rear arm, but also through the foot with metatarsal joint adduction, abduction varus or valgus, to accommodate the uneven roads, calcaneal tubercle of the Achilles tendon attachment, gastrocnemius, and soleus muscle contraction can be used as powerful plantar flexion movements [7]. Withstand sudden load plantar calcaneal talar neck, also with the boat ligament attachment and the boat ligaments supports the metatarsal head and bear weight. Leg with the flat state in the main force three points (A, B, C), the location of a schematic, and the specific value is: A point is of the fifth metatarsal joint head, located at the foot length 63.5%; B point is of the first metatarsal joint, located at the foot length 66.5%; C point is of the heel of the point which is located at the foot of 8%. Heart O point is for the heel which is located at the foot length 18%.

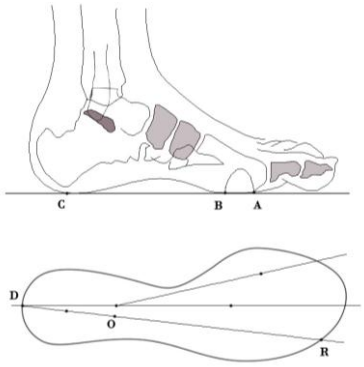


Figure 4. Distribution points on main affected plantar force

3. Personalized Last Bottom Surface Design

This personalized last bottom surface design needs to consider regular foot data and focus on the design of the plantar flexion and longitudinal changes that caused by changes in last bottom surface data.

3.1. Last Bottom Longitudinal Curve Design

Plantar flexion is combined with high variation, last bottom of the end point of the arc, the forefoot area points crown, concave bottom centre point, the last bottom end of the longitudinal axis of the arc endpoint and the last critical point which are shown in Figure 5 [8]. Among them, the key points of the forefoot portion crown point (forefoot landing point) is a point of first metatarsophalangeal joints (72.5% foot length) and fifth metatarsophalangeal joints points (63.5% foot length) of the two connections intersect at the intersection of the axis of the last bottom as a reference, while the point of the forefoot protrusion extents to the convexity of the key data points. In addition, another key point is the end point of cardiac concave flank area (41% foot length) at the corresponding point last bottom axis line as a reference, the point is the last bottom flank position relative to the forefoot and heel heart crown point recessed level (bottom centre concavity).

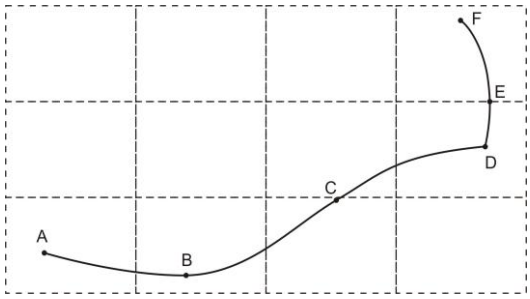


Figure 5. Last bottom longitudinal axes of the key points of the arc, where A is last bottom end point, B is forefoot area points crown, C is base-cantered pits, after the last bottom endpoint D, E is followed by a bump, F is after the arc endpoint (last decency mouth after the point)

3.2. Last Bottom Forefoot Surface Design

Combined with high and vertical variation of the first toe plantar grounding line design, it is referenced by key points: the location forefoot, the first toe plantar forward locations, and the fifth place as a last toe plantar forward before the end of the key point of the palm arc. While redefining the site locations, as shown in Figure 6, it is caused by changes in accordance with the longitudinal axis of the forefoot force point movement, the forefoot of place to first place or fifth toe with plantar locations move forward to design a personalized forefoot last bottom surface.

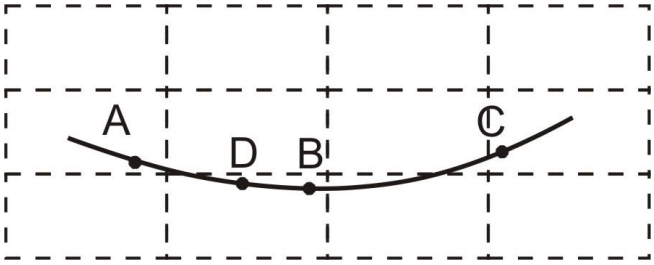


Figure 6. Key points last bottom arc, where A is first toe plantar forward locations, B is forefoot landing point, C is the fifth toe plantar forward locations, D is designed for locations

3.3. Personalization Last Bottom Surface Optimization

Currently, on the shoe surface design method for digital is still relatively small, the majority concentrated in the surface measurements and dimensional scanner rather experimental data analysis based on dynamic optimal design research is not any more. In this paper, by the state against the shoe foot gait data collection, and analysis and surface modelling techniques on last bottom parameter optimization is shown in Figure 7, further to the consideration of the shoes sport state for corresponding data collection and analysis, and finally to determine the optimized high heels last bottom surface design.

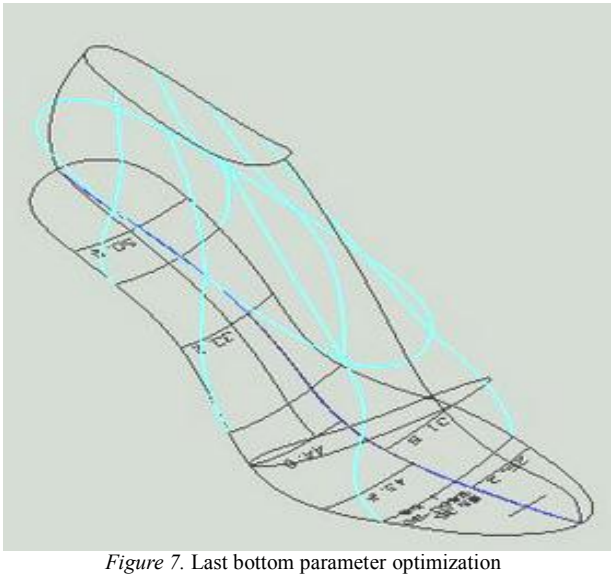


Figure 7. Last bottom parameter optimization

This situation lasts for a 50mm body with high experimental studies and last bottom surface optimization. By using interactive genetic algorithm (IGA) with a general of the basic processes, and doing random minor adjustments to the parameter to generate individual shoe populations tested and calculated; and optimization indexes T is its fitness value, and then select the T value that is greater in parameters to do crossover to generate the next batch program, finally, 8 program is acquired (see Figure 8.)

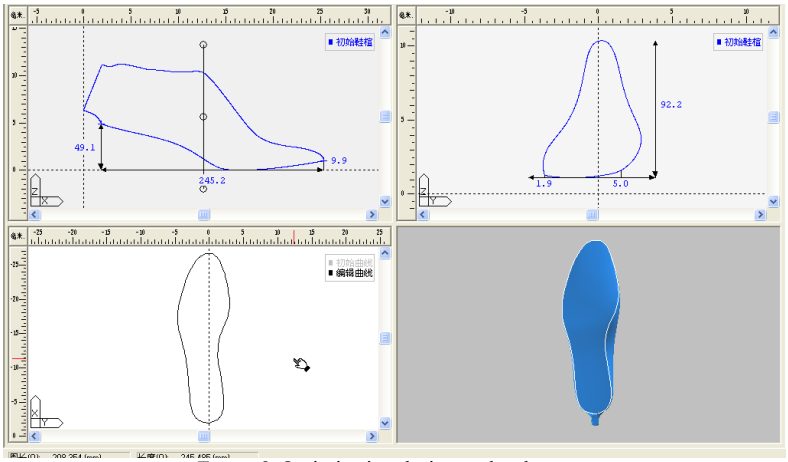


Figure 8. Optimization design on last bottoms

4. Conclusions

In this paper, experimental data evaluation based optimization methods for the shoe last surface design is a practical and reliable design method. Due to limitations of relatively small number of test samples, the data have a certain error to obtain more accurate results, the number of samples must be increased, but this will also bring experimental costs. How to effectively improve the optimization process makes it possible to cycle in as little as possible to get the optimal solution and to overcome the difficulties in future works.

Acknowledgement

This work was supported by Zhejiang Provincial Education Department Research Project-Biomechanics Personalized based for Last Bottom Surface Design.

References

1. Liu, Y., Chen, J., Tong, Sh., et al. (2008) Applied 3-order NURBS Curve to Last Surface and Simulation, *Journal of Zhejiang University*, 42(5), 875-879.
2. Kong, P.W., de Heer, H. (2009) Wearing the F-Scan mobile in-shoe pressure measurement system alters gait characteristics during running, *Gait & Posture*, 29, 143-145.
3. Li, G., Liu, J., Liu, Zh., Zheng, Yu, (2012) Genetic Algorithm with Pull Moves for Folding 2D Model Proteins, *International Journal of Digital Content Technology and its Applications*, 6(21), 412 – 418.
4. Cai, G., Jia, L., Yang, J., Liu, H. (2010) Improved Wavelet Neural Network Based on Hybrid Genetic Algorithm Application on Fault Diagnosis of Railway Rolling Bearing, *International Journal of Digital Content Technology and its Applications*, 4(2), 135 -141.
5. Shi, K. (2009) Research on Mechanical Model of Virtual Human-shoes System based on Fuzzy Association Rule Mining, in *proceedings of 2009 International Conference on Networking and Digital Society (ICNDS 2009)*, IEEE, 296-299.
6. Tessutti, V., Trombini-Souza, F., Ribeiro, A.-P., et al. (2010) In-shoe plantar pressure distribution during running on natural grass and asphalt in recreational runners, *Journal of Science and Medicine in Sport*, 13(1), 151-155.
7. Li, Z., Shi, K. (2013) Ergonomics Design on Bottom Curve Shape of Shoe-Last Based on Experimental Contacting Pressure Data, *International Journal of Digital Content Technology and its Applications*, 7, 86-93.
8. Hillmann, A., Rosenbaum, D., Winkelmannm, W. (2000) Plantar and dorsal foot loading measurements in patients after rotation plasty, *Clinical Biomechanics*, 15, 359-364.

Received on the 3rd of November 2013

Mathematical and Computer Modelling

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 51-57
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

RESEARCH ON FEATHER QUILL IMAGE DENOISING

H. W. Yue[a,b]

[a]Zhongkai University of Agriculture and Engineering
[b]Faculty of Automation, Guangdong University of Technology
No. 24 DongSha Road, Hai Zhu, Guangzhou City, China
Phone: +86-20-8900-2076; +86-13560401502. E-mail: yuehongwei420@163.com

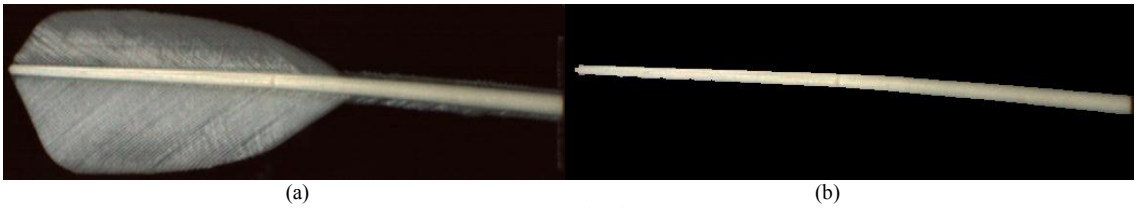
Tiny creases of feather quill is easily disturbed by noise that causing adverse effects on recognition performance. For the shortcomings of the traditional noise reduction algorithm in the process of image, i.e., it is sensitive to the noise resulting in the problem of weakened creases characteristics, this paper discusses probability measure of noisy image on manifolds based on heat equation theory, analyses regularity and uniqueness of heat kernels on manifolds, and gets the corresponding relationship between density function of noisy data and probability density of clean data. A denoising algorithm of feather quill based on the heat equation of graph is proposed. In the method, de-noised image can be obtained through iterative solution of equation based on feather quill image representing as an undirected weighted graph. The experimental results show that this suggested method can get better effect comparing with other transform domain algorithms.

Keywords: feather quill, image denoising, manifold, heat kernels, heat equation

1. Introduction

Badminton is a labour-intensive product with about ten detection steps from feather selecting to finished badminton. In the whole steps, parameter extraction of feather quill (Referred to as "FQ") is a key link for feather grading. There is a disadvantage in traditional detection methods, namely: manual operation, high labour intensity, instability in sorting quality. At present, we have done some study of the problem [1-3], the feather online detection system has been put into production line. The feather quill crease directly affects production quality of badminton, so the ability to accurately identify crease is an important indicator of the detection system. From sidelight image (Fig.1 (a)), we can get FQ image (Fig.1 (b)) by segmentation technique, we can observe the width of FQ is between 1mm to 3mm, crease width is approximately 0.2mm. This tiny defect is sensitive to noise interference. In recent years, manifold theory in image processing is a hot research topic [4], but the literature is still very rare for image denoising with manifold theory. However, the literature does not distinguish between manifold and European space and discuss uniqueness of heat equation solution. This paper will discuss noise sample set on manifold with lifting properties, through heat equation and geometric analysis, analyse regularity and uniqueness of heat kernels on manifolds, and construct a denoising algorithm of feather quill with heat equation based on graph to suppress image noise.

We continue the paper as follows. Section 2 begins with an overview of heat kernels theory on manifold. Section 3 analyses existence and uniqueness of heat kernels, followed by Section 4, which covers the problem of the proposed solution. In Section 5, we compare the performance of the proposed method with previous approaches. The main findings and possible future directions are summarized in Section 6.



(a) (b)

Figure 1. Feather image

2. Theoretical Background

Let noise data set be an open subset of R^d, which can be regarded as a d-dimension manifold. The true data set of X constitute a m-dimension manifold M. Meanwhile these true data is a subset

of X. So there is a smooth, regular embedding $i: M \to R^d(X)$, and M is the sub manifold of X. The Euclidean distance in $R^d(X)$ then induces a metric g on M. This metric depends on the embedding i.Then (M, g) is a Riemannian manifold. As the sub manifold of European space X, there is a natural induced Riemannian metric g on M. Furthermore let manifold M be equipped with a probability measure P_M which is absolutely continuous with respect to the natural volume element dV of M [5]. With these definitions the model for the noisy data-generating process in R^d has the following form: $X = i(\Theta) + \varepsilon$; where $\Theta \sim P_M$ and $\varepsilon \sim N(0, \sigma)$. A manifold with density is a Riemannian manifold M endowed with a measure μ that has a smooth positive density with respect to the Riemannian measure. The space (M, g) features the weighted Laplace operator Δ_μ, generalizing the Laplace-Beltrami operator, which is symmetric with respect to measure μ. It is possible to extend Δ_μ to a self-adjoint operator in $L^2(M, \mu)$.

Definition 1. Any self-adjoint semi-bounded above operator H in $L^2 = L^2(M, \mu)$ defines the heat semi group $\{e^{tH}\}_{t\geq 0}$, which is a family of positive definite bounded self-adjoint operators in L^2.

Theorem 1. For any $f \in L^2$, the function $\mathrm{u(t,x)}: (\mathrm{t,x}) \mapsto \mathrm{P_t} f(x)$ has a version $\mathrm{u(t,x)}$ that is C^∞ smooth in $(\mathrm{t,x}) \in \mathrm{R_+} \times M$. The function $\mathrm{u(t,x)}$ satisfies the heat equation $\dfrac{\partial \mathrm{u}}{\partial \mathrm{t}} = \Delta_\mu u$; the initial condition: $\mathrm{u(t,\cdot)} \xrightarrow{\mathrm{L}^2} f(\mathrm{x})$ as $t \to 0+$; and the estimate is $\operatorname{essinf} f \leq u(\mathrm{t,x}) \leq \operatorname{esssup} f$. Then discussing the following problems such as heat kernels theory on manifold (M, g) [6, 7].

Definition 2. A smooth function $\mathrm{u(t,x)}$ on $\mathrm{R_+} \times M$ is called a fundamental solution of the heat equation at a point $y \in M$ if the function $\mathrm{u(t,x)}$ satisfies in $\mathrm{R_+} \times M$ the heat equation: $\dfrac{\partial \mathrm{u}}{\partial \mathrm{t}} = \Delta_\mu u$.

Note: (1) the Dirac condition: if $t \to 0+$, then $\mathrm{u(t,\cdot)} \to \delta_{\mathrm{y}}$; and for any $\varphi \in \mathrm{D} := \mathrm{C}_0^\infty(\mathrm{M})$, $\int_M \mathrm{u(t,x)}\varphi(x)\mathrm{d}\mu(x) \to \varphi(y)$ as $t \to 0+$;

(2) If in addition $\mathrm{u(t,x)}$ is positive and $\forall \mathrm{t} \in R_+$, $\int_{\mathrm{M}} u(\mathrm{t,x})\mathrm{d}\mu(\mathrm{x}) \leq 1$; then function $\mathrm{u(t,x)}$ is a (regular) fundamental solution at y.

(3) if a function $\mathrm{u(t,x)}$ is positive and satisfies condition (1), then the Dirac condition (2) is equivalent to the following: for any open set U containing y, $\int_{\mathrm{U}} \mathrm{u(t,x)}\mathrm{d}\mu(x) \to 1$ as $t \to 0+$.

Definition 3. A function $\mathrm{q}_t(\mathrm{x,y})$ on $\mathrm{R_+} \times M \times M$ is called a (regular) fundamental solution of the heat equation if, for $\forall y \in M$, the function $(\mathrm{t,x}) \mapsto \mathrm{q}_t(\mathrm{x,y})$ is a (regular) fundamental solution at y.

Theorem 2. Manifold (M, μ) is stochastically complete, if and only if $\int^\infty \dfrac{V(r)}{S(r)} dr = \infty$.

Theorem 3. On a stochastically complete manifold (M, μ), any regular fundamental solution $\mathrm{u(t,x)}$ at a point $\forall y \in M$ coincides with the heat kernel $p_t(\mathrm{x,y})$ and hence is unique.

3. Existence and Uniqueness of the Solution

On Riemannian manifold (M, μ) with the Lebesgue measure in this case the Gauss-Weierstrass function $p_t(\mathrm{x,y})$, $(\mathrm{t,x,y}) \in R \times M \times M$ is:

$$p_t(\mathrm{x},\mathrm{y})=\frac{1}{(4\pi rt)^{\frac{\mathrm{d}}{2}}}\exp(-\frac{|\mathrm{x}-\mathrm{y}|^2}{4rt}) \tag{1}$$

With Lebesgue measure μ, $\Delta_\mu=\sum_{i=1}^{d}\frac{\partial^2}{(\partial x^i)^2}$, for $\forall y\in M$, and observe that:

$$\begin{cases}\dfrac{\partial u}{\partial t}=\dfrac{\partial p_t(\mathrm{x},\mathrm{y})}{\partial t}=-\dfrac{d}{2}(4\pi rt)^{-\frac{\mathrm{d}}{2}-1}4\pi\exp(-\dfrac{|\mathrm{x}-\mathrm{y}|^2}{4rt})+\dfrac{1}{(4\pi rt)^{\frac{\mathrm{d}}{2}}}\exp(-\dfrac{|\mathrm{x}-\mathrm{y}|^2}{4rt})\dfrac{|\mathrm{x}-\mathrm{y}|^2}{4rt^2},\\ \Delta_\mu u=\sum_{i=1}^{d}\dfrac{\partial^2 p_t(\mathrm{x},\mathrm{y})}{(\partial x^i)^2}.\end{cases}$$

Then we obtain $\Delta_\mu u=\sum_{i=1}^{d}\frac{\partial^2}{(\partial x^i)^2}=\frac{\partial u}{\partial t}$, namely $p_{\mathrm{t}}(\mathrm{x},\mathrm{y})$ is a regular fundamental solution to the heat equation $\frac{\partial \mathrm{u}}{\partial \mathrm{t}}=\Delta_\mu u$ at a point $y\in M$. For any open set U containing y, there exists a function $\int_{\mathrm{U}}p_t(\mathrm{x},\mathrm{y})\mathrm{d}\mu(x)\to 1,\quad \mathrm{t}\to 0^+$, using the formula (***Definition 2***), $p_{\mathrm{t}}(\mathrm{x},\mathrm{y})$ is a regular fundamental solution to the heat equation. For geodesic ball of (M,μ) is n-dimension ball with the Lebesgue measure, that is:

$$\begin{cases}V(x,r)=\dfrac{\pi^{\frac{d}{2}}r^d}{\Gamma\left[\frac{d}{2}+1\right]},\\ S(x,r)=\dfrac{d\pi^{\frac{d}{2}}r^{d-1}}{\Gamma\left[\frac{d}{2}+1\right]},\\ \int^{\infty}\dfrac{\mathrm{V}(\mathrm{x},\ r)}{\mathrm{S}(\mathrm{x},\ r)}\mathrm{dr}=\int^{\infty}\dfrac{r}{n}\mathrm{dr}=\infty.\end{cases}$$

by ***Theorem 2***, the manifold (M,μ) is stochastically complete. Therefore, formula (1) is the heat kernel.

In the following discussion on heat kernel is unique. In the definition of heat kernel, for $\forall y\in M$, if $\forall f\in L^2$ t>0, $x\in M$. It is easy to see that the function: $\mathrm{P_t}f(\mathrm{x})=\int_M p_{\mathrm{t}}(\mathrm{x},\mathrm{y})f(\mathrm{y})\mathrm{d}\mu(\mathrm{y})$, and let $f(x)=p(\theta)$ and $t=\frac{\sigma^2}{2}$, $p(\theta)$ is density function of Θ, we obtain the following formula [8]:

$$\mathrm{P}_X p(x)=\int_M\frac{1}{(4\pi r\frac{\sigma^2}{2})^{\frac{\mathrm{d}}{2}}}\exp(-\frac{|x-i(\theta)|^2}{4r\frac{\sigma^2}{2}})p(\theta)dV(\theta).$$

It follows from ***Theorem 1***, that for any $p(\theta)$, the function $\mathrm{P}_t p(x)=\mathrm{u}(\mathrm{t},\mathrm{x})$ is a bounded solution. Then from Theorem 3, we conclude that heat kernel on manifold (M,μ) is unique, the function (1) is the heat kernel in R^{n}. There is a function $p(\theta)$, $\mathrm{P}_t p(x)=\mathrm{P}_t(x,i(\theta)),t=\frac{\sigma^2}{2}$; namely

$$\int_M \frac{1}{(4\pi r\frac{\sigma^2}{2})^{\frac{d}{2}}}\exp(-\frac{\|x-i(\theta)\|^2}{4r\frac{\sigma^2}{2}})p(\theta)dV(\theta)=\frac{1}{(4\pi rt)^{\frac{d}{2}}}\exp(-\frac{\|x-i(\theta)\|^2}{4r\frac{\sigma^2}{2}}) \quad (2)$$

The Gaussian measure p_X is equivalent to the heat kernel $\mathrm{P_t}(x,i(\theta))$ of the diffusion process at time $t=\frac{\sigma^2}{2}$. The basic principle behind the denoising algorithm of this paper is to reverse this diffusion process.

4. Design Algorithm

4.1. Construction of Heat Equation

From ***Theorem 1***, the construction of heat kernel on manifold (M,μ) must first determine the Laplace operator. For European space is special manifold, according to the spectrum graph theory [9], we can construct similarity matrix W with symmetric k-nearest neighbour graph (k-NN), where $w_{ij}=\exp(-\frac{d^2(x_i,x_j)}{(\max(\sigma_i,\sigma_j))^2})$, $\sigma_i=d(x_i,x_k)$. If $i=j$, $w_{ij}=0$; let D be the diagonal matrix and $d_{ii}=\sum_{j=1}^{n}w_{ij}$, so Laplace operator $\Delta=1-D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$. Then we can formulate the algorithm by the following differential equation:

$$\partial_t X=-r\Delta X \ . \quad (3)$$

From the front conclusion, solution of equation (1) is unique. In order to solve the differential equation (1), we choose an implicit Euler-scheme, that is

$$X(n+1)=X(n)(1-hr\Delta+\frac{1}{2}h^2r^2\Delta^2), \quad (4)$$

where h is the time-step, r is the diffusion constant.

4.2. Design of Denoise Algorithm

Solving the equation (4) is an iterative process, if iterative parameter n is big, crease feature is too fuzzy. From the above analysis, we can conclude that: just iteration, the major component of noise is eliminated; Along with the iterative, part of the effective signal is eliminated. So need to set the threshold to control the number of iterations. The entire algorithm is defined as follows:

1) Initialization parameter h, r, k;

2) Structure an undirected graph, compute similarity matrix;

3) Compute Laplace operator Δ, and solve equation (4);

4) Compute $r(n)=\left|PSNR(X_n)-PSNR(X_{n+1})\right|$, if $r(n)\geq r(n+1)$, then terminating iteration, otherwise go to step 2.

5. Experiments

In this paper, experimental samples are collected from feather detection system, and raw material is duck feather. The detection system through CCD camera obtains feather image with side-lighting. Homemade feather acquisition system and its corresponding schematic diagram are shown in Fig.2.

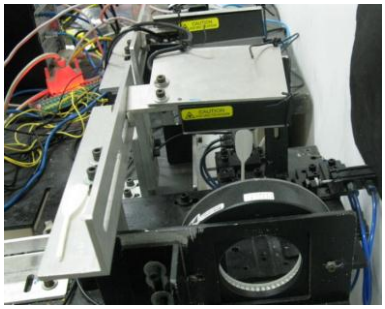


Figure 2. Acquisition system of feather

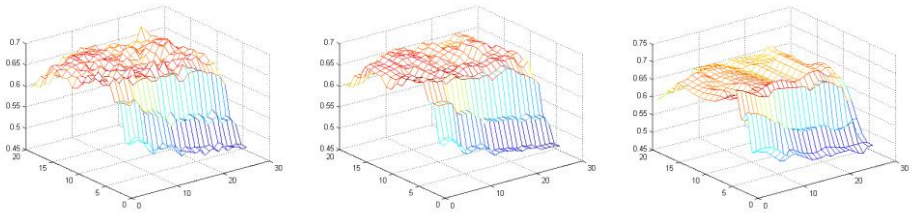
(a)No crease (b)Creased (c)Creased

Figure 3. Experimental samples

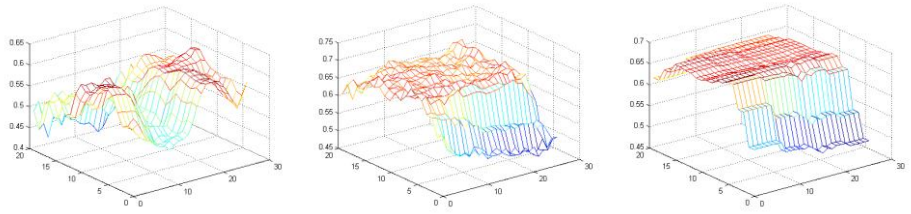
Because of slender structures and camber of FQ, these subimages have abandoned edge in order to reduce effects of edge plus noise, as shown in Fig.3. The above denoising algorithm is compared with DWT [10], Contourlet transform [11] and Curvelet transform [12] hard-thresholding method. Comparison of view denoising for Fig.3 with Gaussian noise using a fixed variance $\sigma = 0.01$, where the proposed algorithm parameters set as follows: $h = 1,\ r = 1,\ k = 7$. Table 1 shows the denoising results on the real image Fig.3. The proposed algorithm is clearly shown to be more effective than the DWT, Contourlet and Curvelet, as well as in term of PSNR's. Other algorithm's PSNR is less than the proposed algorithm. In order to observe the effect of the various noise reduction algorithms, the following is given respectively the spatial distribution of Fig. 3 (a)and(b) after the noise reduction, as shown in Fig. 4and 5.

TABLE 1. PSNR of different denoising methods

Image	**PSNR/db**			
	DWT	**Contourlet**	**Curvelet**	**Proposed**
Figs.2.a	35.12	37.23	39.08	39.31
Figs.2.b	34.19	36.59	38.15	38.18
Figs.2.c	36.4	37.98	38.77	39.86

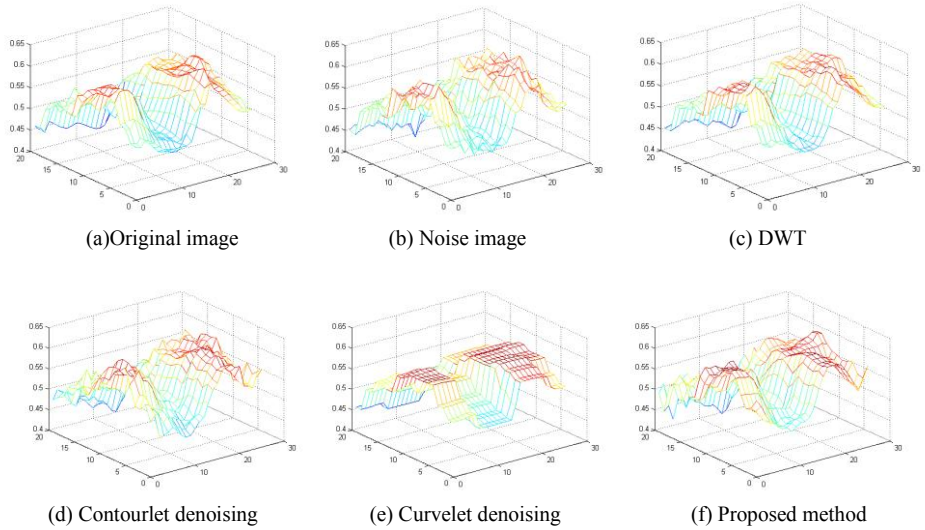


(a)Original image (b) Noise image (c) DWT denoising



(d)Contourlet denoising (e) Curvelet denoising (f) proposed method

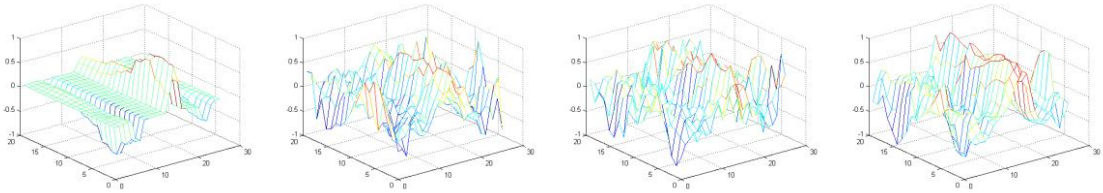
Figure 4. Denoising results of image 2(a)



(a)Original image (b) Noise image (c) DWT

(d) Contourlet denoising (e) Curvelet denoising (f) Proposed method

Figure 5. Denoising results contrast of image 2(b)



(a) Wavelet denoising (b) Contourlet denoising (c) Curvelet denoising (d) Proposed method

Figure 6. Gradient distribution of image 2(c)

According to the subjective visual qualities respectively, the proposed scheme has better smoothness than other algorithms. On the other hand, to further demonstrate visual quality of images, for example, Figure 6 presents graded distribution of Figure 2(c) after the noise reduction processing. For image processed with Dwt, Contourlet, Curvelet, from the subjective quality of the image one can observe that some of the impulse noises are not fully suppressed while crease of feather quill is blurring and distortional seriously. Both the performance of PSNR and the subjective visual qualities of our method outperform other algorithms in all cases.

6. Conclusion

In this paper, a new adaptive noise reduction scheme for removing salt and pepper noise is proposed. The first phase of the scheme efficiently identifies impulse noise while the other is to remove the noise from the corrupted image that is followed by image enhancement scheme to preserve the details and image quality. As per the experimental results, the proposed algorithm yields good filtering result using efficient noise detection mechanism. This is observed by numerical measurements like PSNR and visual observations through the experiments conducted. In addition, for characteristics of crease can be better retained, this will be conducive to automatic identification of feather quill with crease defect, and provide the necessary basis for further work in feature recognition.

References

1. Liu, H., Wang, R., Li, X. (2011) Improved Algorithm of Feather Image Segmentation based on Active Contour Model, *Journal of Computer Applications,* 32(8), 2246-2245.
2. Liu, H., Wang, R., He, Z. (2011) Slender Object Extraction for Image Segmentation Based on Centerline Snake Model, *Opto-Electronic Engineering*, 38(9), 124-129.
3. Liu, H., Wang, R., Li, X. (2011) The Finite Ridgelet Transform for Defeat Detection of Quill, *Journal of Computer Applications,* 32(1), 931-936.
4. Hein, M., Audibert. (2005) From Graphs to Manifolds - Weak and Strong Pointwise Consistency of

Graph Laplacians. In: *Proceeding of the 18th Conference on Learning Theory (COLT), Bertinoro, Italy, June 27-30, 2005.* (470-485). Springer Berlin Heidelberg.

5. Belkin, M., Niyogi, P. (2003) Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, *Neural computation*, 15(6), 1373-1396.
6. Cheeger, J., Yau, S. (1981) A Lower Bound for the Heat Kernel, *Comm. Pure Appl. Math.,* 34(4), 465-480.
7. Grigor'yan, A. (2009) *Heat Kernel and Analysis on Manifolds*. Providence: American Mathematical Society.
8. Hein, M., Maier, M. (2007) Manifold Denoising. In: *Advances in Neural Information Processing Systems 20: 21st Annual Conference on Neural Information Processing Systems 2007,Vancouver,Canada, December 3 -6, 2007,* (561–568). Cambridge: MIT.
9. Graham, F.C. (1997) *Spectral Graph Theory*. Fresno: American Mathematical Society.
10. Chen, F., Wang, W. (2009) Target Recognition in Clutter Scene based on Wavelet Transform, *Optics Communication,* 282(4), 523-526.
11. Mani, V., Arivazhagan, S., Braino, J. (2013) Multimodal Image Fusion using Multi-resolution Techniques, *International Journal of Biomedical Engineering and Technology,* 11(1), 3160-3163.
12. Ma, J., Plonka, G. (2010) The Curvelet Transform, *IEEE Signal Processing Magazine*, 27(2), 118-133.

Received on the 21st of June 2013

Mathematical and Computer Modelling

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 58-65
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

TUNING BG MULTI-PATTERN STRING MATCHING ALGORITHM WITH UNROLLING Q-GRAMS AND HASH

P. Yang[1], H. Fan[1], Qi. Huang[1], Li. Liu[2]

[1]Department of Computer Science Kunming University of Science and Technology,
China Yunnan Key Laboratory of Computer Technology Applications
[2]Department of Biomedical Engineering Kunming University of Science and Technology, China
E-mail: kmustailab@hotmail.com

In many important application fields such as network security and information filtering, fast increasing of network bandwidth demands the high performance multi-pattern string matching algorithms. In this paper, we improved the BG algorithm, which is one of the fastest multi-pattern string matching algorithms and present a serial of algorithms named BGqus. BGqus uses an unrolling q-grams method to simulate the original q-grams of big q value with unrolling q-grams of small q value. This method can reduce the branch cost and increase the filter capability with small space occupancy. In addition, BGqus uses a hash method to reduce the space occupancy of the bit mask table furthermore. The experimental results indicated that BGqus is very fast for big alphabet. It is very suits for the network security and information filtering on current computer for Gigabyte network.

Keywords: BG, hash, multi-pattern, unrolling q-grams

1. Introduction

Multi-patterns string matching is the performance bottleneck of many important fields, such as information filter, network and information security etc. For example, in snort, a well-known open source IDS, more than 70% computing time is for the multi-pattern string matching [1].In these fields, filter bandwidth must be bigger than the communication bandwidth. The Gilder's Law tells us that the total bandwidth of communication systems triples every 12 months, and the Moore's Law tells us that the hardware performance doubles every 18 months. Fast increasing of network bandwidth demands the research for fast multi-pattern string matching methods or algorithms, which the requirement for the performance of the multi-pattern string matching is endless. Especially in the main research fields of our team, the large-scale information retrieval and the mass medical information processing, poor performance of string matching caused a serious impact on our research.

For the given alphabet $\sum$ of size σ, $\sum^*$ is the closures of $\sum$, and for the given text string $T = t_0 t_1 .. t_{n-1} \in \sum^*$ and the set of patterns $P = \{P_1, P_2 ... P_{r-1}\}$, which $\forall P_i = p_0 p_1 ... p_{mi-1} \in \sum^*$ and mi is the length of the string P_i, the multi-patterns string matching is the problem of calculating the set: $O = \{pos \mid \exists k, P_k[j] = T[pos + j], 0 \le k < r, 0 \le j < mi, 0 \le pos < n\}$.

To date, BG [2] is one of fastest multi-pattern string matching algorithm. The author of this algorithm presented a q-grams bit-parallel filter method to quickly filter out most of positions can not match a pattern in the pattern set, and use the method of SBNDM [3] (a single pattern string matching algorithm) to realize string matching. This filter method can obviously increase the performance of algorithms. However, the SBNDM is an old single pattern string matching algorithm, and there are many improved algorithms based on SBNDM. If the mechanisms of the improved algorithms based on SBNDM are adopted on BG, faster multi-pattern string matching algorithms may be gained.

The q-grams technique is an important accelerating technique on pipeline processors. To date, many of the current fast algorithms use q-grams. In q-grams, q consecutive characters are processed as a single character. For example, the string "Hello" is processed as "Hel-ell-llo" for 3-grams, and each continuous q characters is called a q-gram. This can make the alphabet perceived by the algorithm larger. Because a fairly large alphabet leads to a low rate of branch prediction failure, q-grams can greatly reduce the branch cost and a high probability of jump, which enhance the algorithm performance obviously. A large alphabet can gain good filter capability for BG. In addition, another benefit of q-grams is that q-grams solutions are more flexible than original one.

Large-scale q-grams multi-pattern string matching needs big q value to gain an enough low branch prediction failure rate. However, BG is hard to support big q value because of exponential space ($4\sigma^q$ bytes) of the bit vector table. In addition, for large-scale multi-pattern method, the space

occupancy is commonly very big, and the cost of memory access has become the bottleneck of performance. Therefore, reducing the space occupancy is an important accelerating way for multi-pattern string matching field.

In this paper, we improved the BG algorithm. By introducing an Unrolling Q-grams method to BG that can simulate the q-grams method for big q value with the q-grams method for smaller q value q-grams, and by introducing the hash method into the q-grams bit vector table which can reduce the space of the table further, we presented a serial multi-pattern string matching algorithms named BGqus. Experimental results indicated that BGqus is faster than BG in most of case, and it can support gigabit network applications even for very large-scale matching.

2. The SBNDM Algorithm

SBNDM is the fundamental algorithm of BG. It uses the factor method to jump among the text. In the factor method, a sliding window of the same length with the pattern is sliding along the text. Text characters in the window are read backward. Let the string that has been read be U. if it is recognized that U is not a substring of the pattern, the string matching may not occur in this window and the window can safely jump over U. If the whole window is recognized as a substring of the pattern, because the length m substring of the pattern just is the pattern, a string matching is found in this window. Then, record the found location and slide the window1 character. The string matching can be realized by repeating above actions until the window has been over the end of the text which realize. The factor method is shown as Fig.1.

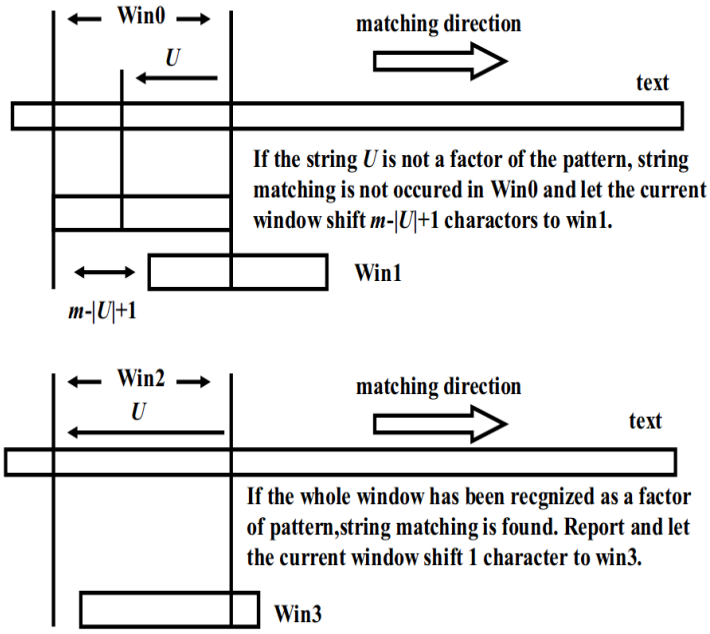


Figure 1. The factor method

To recognize all substrings of the pattern, SBNDM uses a backward NFA, which an example of the NFA is shown in Fig.2 for the pattern “baabbba”. In this NFA, all states are final.

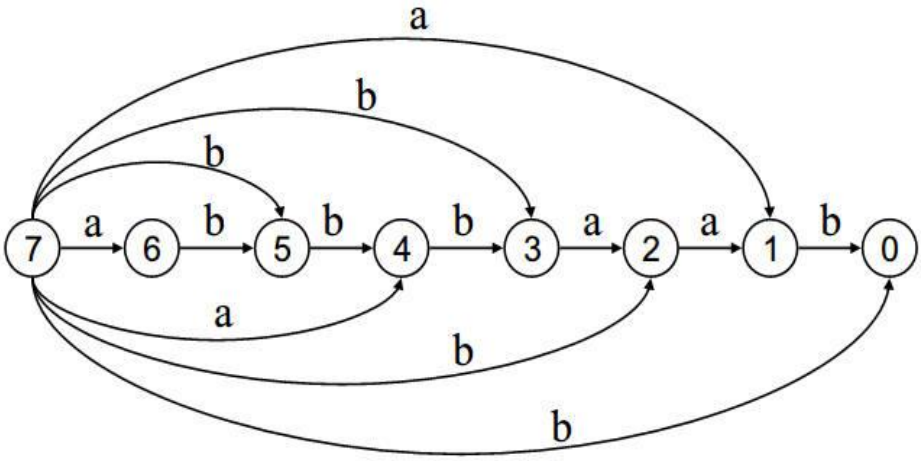


Figure 2. Example of the NFA to recognize all substrings of the pattern "baabbba"

Because of the poor performance to deal with NFA for computer, SBNDM uses a bit-parallel method used in Shift-And [4] to simulate the NFA. SBNDM pre-computes a bit vector table B, which stores a bit mask for each character in alphabet. The set bits in the bit mask for a character express the occurrences of this character in the reversal of pattern. When reading the characters backward, a state vector D is maintained by using the bit mask of the read character with the Shift-and technique, which the set bits in D express the occurrences of the window suffix has been read in reversal of the pattern. After a suffix of the window inputted, if and only if D is zero, the inputted string is not a substring of the pattern.

SBNDM can easily extended to deal with the extended string matching with character set by set the corresponding bits to 1 for bit masks for each character in the character set. The matching phase to deal with extended string with character set is same with the matching phase to deal with single pattern.

3. The BG Algorithm

To support multi-pattern matching, BG use extended string matching with character set based on SBNDM to be the filter method of multi-pattern matching. Let the pattern set P is {pattern, filters}, the BG algorithm can be described as follows:

If the q-grams method is not applied, BG processes the pattern set as an extended string with character set $P' = "\{p, f\},\{a,i\},\{t,l\},\{t\},\{e\},\{r\},\{n,s\}"$. If a substring of text can not match the extended string P', this substring can not match any pattern in the pattern set P. In opportune cases, the extended string matching is faster than the multi-pattern matching, and using extended string matching to filter out most of negative positions can obviously increase matching performance. For the locations that matches P', because the number of strings that matches P' is bigger than r, a multi-pattern string matching processing should be done to determine whether there is a sting matching of a pattern in the set of P. In the multi-pattern string matching phase, any correctly multi-pattern string matching method is allowed. BG uses the RKBT multi-pattern matching method. However, it is shown in our experimental results that using TRIE to realize multi-pattern matching may be faster.

For large-scale matching or small alphabet, the character set of a position in the extended pattern P' may cover most of characters in the alphabet, which deteriorates the performance seriously. The q-grams method is a good method to deal with this problem. Q-grams can exponential enlarge the alphabet perceived by the algorithm. For example, for 2-grams, the pattern set P can be processed as the extended pattern $P_2' = "\{pa, fa\},\{at,il\},\{tt,lt\},\{te\},\{er\},\ \{rn,rs\}"$ .The proportion of the q-grams characters of the character set in the q-grams alphabet is exponentially lower than the proportion which q-grams is not applied. Therefore, introducing q-grams can retard the performance deterioration for increasing of matching scale.

Smaller alphabet, longer patterns or larger scale leads to low branch prediction failure rate and low capability of filter. In these cases, it demands bigger q value to gain high performance. However, with the increasing of the q value, the space occupancy is increased exponential. Too big space

occupancy (64MB/3-grams and 16GB/4-grams on Extended ASCII alphabet) lead to that BG can not support big q value. To decrease the space and enhance the memory access performance, a variant of BG named BG6b that only build the bit mask for the lowest 6 bit of all characters has been presented to in [9], which reduce the space from $4\sigma^q$ bytes to 4×64^q bytes. Although this method reduces the space occupancy, but this has not solved the problem that that space is too big completely and BG6b can still not support bigger q value.

The code of 2-grams BG with TRIE multi-pattern matching is listed as Code.1.

//2-grams BG with Trie (P, T, m, n, r)
//preprocessing
$newstate \leftarrow 0;\ pos \leftarrow -1;$
for $i \leftarrow 0$ **to** $r-1$ **do**
$\{ m \leftarrow |P_i|;\ a_0a_1...a_{m-1} \leftarrow P_i^{rv};\ cur \leftarrow 0;\ j \leftarrow 0;\ k \leftarrow 1;$
while $g(cur,a_j) \neq \perp$ **and** $j < m$ **do** $\{ cur \leftarrow g(cur,a_j);\ j \leftarrow j+1;\}$
while $j < m$ **do**
$\{ newstate \leftarrow newstate+1;\ g(cur,a_j) \leftarrow newstate;\ cur \leftarrow newstate;\ j \leftarrow j+1;\}$
$output[cur] \leftarrow output[cur] \bigcup a_0a_1...a_{m-1};$
for $j \leftarrow m-1$ **to** 1 **do**
{
$B[(P_i[j] << 8) + P_i[j-1]] \leftarrow B[(P_i[j] << 8) + P_i[j-1]] | k;\ k \leftarrow k << 1;\}\}$
//*matching*
while $pos < n-m$ **do**
$\{ D \leftarrow \sim 0;\ j \leftarrow m-1;$
do $\{ D \leftarrow D \& B[(T[pos+j+1] << 8) + T[pos+j]];\ D \leftarrow D << 1;\ j \leftarrow j-1;$
$\}$**while** $D \neq 0$ **and** $j \neq 0;$
if $D \neq 0$ **then**
$\{ cur \leftarrow 0;$ **for** $i \leftarrow 1$ **to** m **and** $cur \neq \perp$ **do** $\{ cur \leftarrow \delta(cur, T[pos+i]);\}$
if $cur \neq \perp$ **then** *Report find matches of* $output[cur];\}$
$pos \leftarrow pos+1+j;\}$

Code 1. 2-grams BG with TRIE multi-pattern matching

4. The Unrolling Q-grams Method for BG

If a method to simulate the q-grams of big q value with the q-grams of small q value is presented, a faster algorithm can be gained, because the branch prediction failure rate is reduced and the filter capability is increased but the space occupancy is not increased.

The SBNDMQ [5] algorithm gave an unrolling q-grams method that can simulate the original q-grams method like BG used. In this method, the initial value of the bit vector D is set as follows: $D = B[t_i] \& (B[t_{i+1}] << 1) \& ... \& (B[t_{i+q-1}] << (q-1))$.

It can be proofed that this initial value of D is equal to the bit mask of the last q-gram character for original q-grams method. That is means the unrolling q-grams method and the q-grams method like BG used can simulate each other at the time to set the initial value of D. Although the unrolling q-grams needs more calculation, it just need 1KB bit mask table for Extended ASCII alphabet, and avoids the big space occupancy.

Use this idea, we presented an unrolling q-grams method for BG.

For a 3-gram $t_{pos}t_{pos-1}t_{pos-2}$, let the table B is the bit vector table of SBNDM, therefore, the 2-grams bit vector table B_2 can be defined as $B_2[t_{p_{os}}t_{pos-1}] = (B[t_{pos}] << 1) \& B[t_{pos-1}]$, the 3-grams bit vector table B_3 can be defined as $B_3[t_{pos}t_{pos-1}t_{pos-2}] = (B[t_{pos}] << 2) \& (B[t_{pos-1}] << 1) \& B[t_{pos-2}]$. The 3-grams bit mask table can be simulated by unrolling 2 2-grams bit mask table for

reducing the branch prediction failure rate because:

$$
\begin{aligned}
&(B[t_{pos}] << 2) \,\&\, (B[t_{pos-1}] << 1) \,\&\, B[t_{pos-2}] \\
&= (B[t_{pos}] << 2) \,\&\, (B[t_{pos-1}] << 1) \,\&\, (B[t_{pos-1}] << 1) \,\&\, B[t_{pos-2}] \\
&= (((B[t_{pos}] << 1) \,\&\, B[t_{pos-1}]) << 1) \,\&\, (B[t_{pos-1}] << 1) \,\&\, B[t_{pos-2}] \\
&= (B_2[t_{pos} t_{pos-1}] << 1) \,\&\, B_2[t_{pos-1} t_{pos-2}]
\end{aligned}
$$

It can be extended that the k unrolling of the q-grams bit mask table can simulate the $k+q-1$ grams bit mask table. In addition, if the result of bit vector D is zero after the k unrolling in q-grams BG, the sliding window can jump m-(q+u-1)+1=m-q-u+2 characters by the factor method.

5. The Hash Method

Because the bit vector table B of BG space occupied is still very big, to reduce the space occupancy of the bit mask table furthermore, we introduced the hash method. We use the hash method as follows:

Let the operations to calculate the position in the q-grams bit mask table of the string $P_i[j-q+1...j]$ is $(P_i[j] << ((q-1)\times s)) + (P_i[j-1] << (q-2)\times s)) + ... + P_i[j-q+1]$ instead of the operation that $(P_i[j] << ((q-1)\times \log_2 \sigma)) + (P_i[j-1] << (q-2)\times \log_2 \sigma)) + .. + P_i[j-q+1]$ in original BG. The value of s can modify the space of the bit mask table, smaller s can obviously reduce the space occupancy of the table but it leads to high probability of hash collisions and the s should be not bigger than $\log_2 \sigma$ because there is no collision when $s \geq \log_2 \sigma$.

After introducing of the unrolling q-grams method and above hash method, the result algorithm based on BG is named BGqus (BG with q-grams, u Unrolling and s-bit shift hash method). An example of BGqus for $q=3$, $u=2$ and $s=5$ is listed as Code.2

```
//BGqGuUsS (P, T, m, n) for q=3, u=2,s=5
//preprocessing
newstate ← 0; q ← 3; u ← 2; s ← 5
for i ← 0 to r − 1 do
  { m ←| P_i |; a_0 a_1 ... a_{m−1} ← P_i^rv; cur ← 0; j ← 0; k ← 1;
  while g(cur, a_j) ≠ ⊥ and j < m do { cur ← g(cur, a_j); j ← j + 1;}
while j < m do
  { newstate ← newstate + 1; g(cur, a_j) ← newstate; cur ← newstate; j ← j + 1;}
output[cur] ← output[cur] ⋃ a_0 a_1 ... a_{m−1};
for j ← m − 1 to q − 1 do
  {
  B[(P_i[j] << (2 * s)) + (P_i[j − 1] << s) + P_i[j − 2]] |= k; k ← k << 1;}}
//matching
pos ← m − 1; T[n...n + m − 1] ← P_0; edge ← n − m + q − 1;
while pos ≤ edge do
  {
  D ← B[(T[pos] << (2 * s)) + (T[pos − 1] << s) + T[pos − 2]]
        &B[(T[pos − 1] << (2 * s)) + (T[pos − 2] << s) + T[pos − 3]];
   while D = 0 do
    { pos ← pos + m − q − u + 2;
      D ← B[(T[pos] << (2 * s)) + (T[pos − 1] << s) + T[pos − 2]]
        &B[(T[pos − 1] << (2 * s)) + (T[pos − 2] << s) + T[pos − 3]];}
```

$j \leftarrow m-q-u+1$; $pos \leftarrow pos-u$;
$D \leftarrow (D << 1) \& B[(T[pos] << (2*s)) + (T[pos-1] << s) + T[pos-2]]$;
while $D \neq 0$ **and** $j \neq 0$ **do**
{ $pos \leftarrow pos-1$; $j \leftarrow j-1$;
$D \leftarrow (D << 1) \& B[(T[pos] << (2*s)) + (T[pos-1] << s) + T[pos-2]]$;}
if $j = 0$ **then**
{ $cur \leftarrow 0$; $pos \leftarrow pos+1$
for $i \leftarrow 0$ **to** $m-1$ **and** $cur \neq \perp$ **do** $cur \leftarrow g[cur][T[pos-q+1+i]]$;
if $cur \neq \perp$ **then** *Report find matches of* $output[cur]$;}
$pos \leftarrow pos+m-q+1$;}

Code 2. BGqus Algorithms for q=3, u=2, s=5

6. Experiments Results

To show the performance of BGqus, the following comparative experiment with AAC [6], Set BOM [7] and the serial of BG algorithms was done:

The platform is AMD Fx-5000 4-core@2.2Ghz/ 785G/ 4GB DDR2 RAM/ Ubuntu 10.04LTS 64-bit desktop edition/g++4.4.5 with -O3 optimize parameter. We tested under the three 20MB length random texts with alphabet size is 2,16,64 respectively, which the random function was the lib function of g++, rand(), and tested under the DNA sequence E.coli, English test Bible.txt, and natural language samples text world192.txt, which above 3 texts are from SMART [8]. For each matching condition, the pattern set were picked from the 400 prior random selected and non-overlapping substring of the text with the same length, and the average matching speed was recorded as the final result. The text had been read in memory to avoid the impact of disk and before each time of matching, a big table had been read sequentially to empty the Cache. Only the matching phase was timed by RDTSC* (the error range is ±30 CPU ticks). In this experiment, the CPU frequency was locked by cpufrequtils**; the network and the unrelated background service are closed to ensure the processor utilization was below to 3%.

The experimental results of the optimal parameter and the average matching speed for the fast four type algorithm which unit is MB/s are shown from Table 1 to Table 6.

TABLE1. The results on random text of alphabet size 2.

rand2			**m=8**			**m=24**		
r=	10	100	1000	10000	10	100	1000	10000
AAC	**285.2**	**143.0**	**120.5**	**120.5**	**246.7**	**101.3**	**14.9**	**9.6**
SetBOM	**78.4**	**18.2**	**12.6**	**10.3**	**253.1**	**82.6**	**14.4**	**5.8**
BG*q*G*x*B	31.4(3,6)	16.6(1,8)	18.8(1,8)	18.9(1,8)	127.9(3,6)	8.7(1,8)	3.2 (1,8)	1.2(1,8)
BG*qus*	75.7(5,3,2)	15.9(4,5,1)	16.0(4,5,1)	15.8(4,5,1)	**278.2**(5,3,1)	7.0(2,4,8)	2.9(2,4,8)	1.2(2,4,8)

TABLE2.The results on random text of alphabet size 16.

rand16			***m*=8**			***m*=24**		
r=	10	100	1000	10000	10	100	1000	10000
AAC	276.3	192.6	56.4	10.6	275.6	184.5	40.3	16.1
SetBOM	604.6	348.8	79.2	33.8	1124.1	656.0	223.9	93.5
BG*q*G*x*B	516.1(2,8)	334.2(3,6)	154.0(3,6)	8.5(3,6)	1264.3(2,8)	863.6(3,6)	511.5(3,6)	17.4(3,6)
BG*qus*	**1677.0(2,2,8)**	**937.3(3,2,5)**	**367.8(3,2,5)**	**141.1(4,2,5)**	**2939.5(3,2,3)**	**2475.5(3,2,5)**	**1218.0(4,3,3)**	**386.8(5,3,3)**

* http://en.wikipedia.org/wiki/Time_Stamp_Counter
** http://wiki.archlinux.org/index.php/Cpufrequtils

TABLE3.The results on random text of alphabet size 64.

rand64			**m=8**			**m=24**		
r=	10	100	1000	10000	10	100	1000	10000
AAC	289.5	196.1	65.2	10.4	274.1	198.9	55.9	18.9
SetBOM	896.2	648.2	147.5	94.4	1622.9	888.8	339.8	134.4
BG*q*GxB	888.3(2,8)	794.7(2,6)	410.3(2,8)	224.0(3,6)	1829.7(2,8)	1386.3(2,6)	892.2(2,8)	354.7(3,6)
BG*qus*	**2151.0(2,2,8)**	**2161.7(2,2,8)**	**1162.1(3,2,4)**	**442.0(3,2,5)**	**2754.7(2,2,8)**	**2572.0(2,2)**	**1995.0(3,2,4)**	**521.8(3,3,5)**

TABLE 4.The results on DNA sequence.

E.coli			***m*=8**			***m*=24**		
r=	10	100	1000	10000	10	100	1000	10000
AAC	270.1	138.4	**43.8**	**10.8**	267.3	133.4	16.8	11.6
SetBOM	238.0	94.6	23.8	3.2	624.7	260.8	56.3	**23.8**
BG*q*GxB	180.5(3,6)	30.7(3,6)	11.1(1,8)	4.9(3,6)	552.6(3,6)	92.3(3,6)	5.7(1,8)	3.5(1,8)
BG*qus*	**384.1(3,2,3)**	**148.8(5,1,3)**	21.6(5,3,3)	5.2(5,3,2)	**1414.0(5,3,1)**	**675.5(5,3,2)**	**125.7(5,3,3)**	3.3(2,5)

TABLE 5.The results on pure English text.

bible			***m*=8**			***m*=24**		
r=	10	100	1000	10000	10	100	1000	10000
AAC	275.6	183.8	61.8	**27.6**	274.4	179.2	27.7	11.6
SetBOM	377.5	151.6	27.8	7.9	826.2	361.4	61.0	16.4
BG*q*GxB	348.8(2,8)	140.3(3,6)	33.1(3,6)	8.5(3,6)	903.5(2,8)	509.8(2,8)	156.4(3,6)	6.6(3,6)
BG*qus*	**738.6(3,2,2)**	**227.7(3,2,3)**	**53.5(5,1,3)**	12.0(5,3,3)	**1830.9(3,3,2)**	**799.7(3,5,3)**	**263.0(5,3,2)**	**31.9(5,3,3)**

TABLE 6.The results on a sample of natural language.

world192			***m*=8**			***m*=24**		
r=	10	100	1000	10000	10	100	1000	10000
AAC	277.2	206.4	59.1	**27.1**	273.9	195.2	26.5	**13.7**
world192			***m*=8**			***m*=24**		
r=	10	100	*r*=	10	100	*r*=	10	100
SetBOM	402.9	174.8	24.2	5.8	855.3	324.1	23.4	3.5
BGqGxB	395.1(2,8)	170.3(2,8)	47.9(3,6)	9.2(3,6)	932.1(2,8)	517.7(2,8)	86.7(3,6)	7.0(3,6)
BG*qus*	**821.8(2,2)**	**291.6(3,2,3)**	**61.0(4,2,3)**	12.0(4,4,3)	**1690.5(3,2,2)**	**692.6(3,3,3)**	**105.5(3,3,3)**	9.3(4,4,3)

It is shown in the experimental results that BGqus is faster than BG in any case. For small alphabet and large-scale matching, because there are too string matching occurs, the extended string matching most of positions of text, the filter method is ineffective. Then AAC is the fastest algorithm in comparative algorithms because of its liner worst time complexity. In addition, in other cases, BGqus is always the fastest one and faster than other algorithms of times on our platform. Especially for alphabet size 64 and pattern length 8, even for the scale 10000 patterns, BGqus can reach 3.5Gbps performance. Since it is point out that the alphabet size of the equivalent character set of random network flow is about 64 [9], BGqus is very suits for the network security and information filtering on current computer for Gigabyte network.

7. Conclusions

In this paper, a serial of multi-pattern string matching algorithms named BGqus was presented based on the BG algorithm. BGqus uses an unrolling q-grams method to simulate the original q-grams of big q value with unrolling q-grams of small q value. This method can reduce the branch cost and increase the filter capability with small space occupancy. Moreover, BGqus uses a hash method to

reduce the space occupancy of the bit mask table furthermore. The experimental results indicated that BGqus is very fast for big alphabet. It is very suits for the network security and information filtering on current computer for Gigabyte network.

Acknowledgment

This paper is supported by the National Natural Science Foundation of China (NO.81360230) , the Natural Science Foundation of Yunnan Province Science and Technology Department (NO. 2012FB131),the Yunnan Province Social Development Science and Technology Project of China (NO.2010CA016) and the Innovation Fund for Technology Based Firms of China (NO. 10C26215305130).

References

1. Calder, B., Tuch, N., Sherwood, T., Varghese, G. (2009) *Deterministic memory-efficient string matching algorithms for intrusion detection*. IEEE INFOCOM.
2. Salmela, L., Tarhio, J., Kyt¨ojoki, J. (2006). *Multi-pattern string matching with q-grams*. ACM Journal of Experimental Algorithmics, (11).
3. Peltola, H., Tarhio, J. (2003) Alternative algorithms for bit-parallel string matching. In Proceedings of the 10th International Symposium on String Processing and Information Retrieval SPIRE'03, volume 2857 of LNCS, pp. 80–94, Manaus, Brazil, Springer-Verlag, Berlin.
4. Baeza-Yates, R., Gonnet, G.H. (1992) *A new approach to text searching*. Commun. ACM, 35(10), 74–82
5. Durian, B., Holub, J., Peltola, H., Tarhio, J. (2009) Tuning BNDM with q-grams. Proceedings of the Workshop on Algorithm Engineering and Experiments (ALENEX), pp. 29–37
6. Aho, A.V., Corasick, M.J. (1975) *Efficient string matching: an aid to bibliographic search*. Communications of the ACM, 18, 333–340
7. Allauzen, C., Crochemore, M., Raffinot, M. (2001) Efficient experimental string matching by weak factor recognition. In proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching, volume 2089 of LNCS Springer-Verlag, pp. 51–72
8. http://www.dmi.unict.it/~faro/smart/[OL,2013/3/10]
9. Salmela, L. (2009) *Improved Algorithms for String Searching Problems* [Ph.D. Thesis]. Helsinki: Helsinki University.

Received on the 21st of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 66-73
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

THREE-DIMENSIONAL DNA STRUCTURE SOLUTION TO NEAREST-NEIGHBOR CLUSTERING

X. Ren[1], X. Li[2], X. Liu[3]

[1]School of Management Science and Engineering
Shandong Normal University, Wenhua Esat str. 88, Jinan, Shandong, China
Phone: (+86)152531362756. Email: heaventallxiaozhu@gmail.com
[2]Phone: (+86)13608927382. Email: sunnysddy@126.com
[3]Phone: (+86)15965776043. Email: sdxyliu@163.com

In this study, we propose to construct a 3-dimensional structure to solve the nearest-neighbor clustering problem, making subtly use of fantastic characteristics of three kinds of DNA structures, 3-armed DNA, DNA hairpin structures and triple-stranded DNA. Because nearest-neighbor clustering problem can be transformed into MST problem, the 3-armed DNA molecules can be treated as the limb structures of a binary tree and the hairpin structures can do a good job to cap off the cleaved ends of DNA to make them double-stranded. Then the triple-stranded DNA structures are used to do the screening job with a lower rate of wrong matches to get the trees containing all the data points. Furthermore, the final result is detected by finding the well qualified minimum spanning trees using the fluorescence analysis. In theory, the algorithm takes *O(n)* experimental steps to get the final clustering result.

Keywords: Nearest-neighbor clustering, 3-armed DNA, MST, 3-dimensional DNA structure

1. Introduction

Clustering as a module in data mining can be used as a separate tool to uncover some deep knowledge in database, and to summarize characteristics of each category, then focus on a particular class for further analysis. A hierarchical way is wildly used, which can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion. However, a pure hierarchical clustering method easily suffers its incapable of performing adjusting once a merge has been done [1]. However, DNA computing with its strong parallelism and powerful storage capability can make a difference.

DNA computing has made great achievements since 1994 [2]. On the one hand, it has solved many NP-complete problems by transforming them into permutation and combination problems or sometimes into graph theory problems. In addition, without exception the nearest-neighbour clustering here is transformed into the minimum spanning tree, the graph theory problem. On the other, during the past nearly two decades, various DNA structures, basic double-helix DNA, closed circle DNA [3], etc., performing as specific models are used to solve problems. Yet, few studies focus on the mixed use of this DNA structures to make a better solution to complex problems, such as Carlos [4], combing DNA sticker model and double-stranded DNA model to solve the traditional Bin-Packing Problem. Aim to make it more flexible to navigate biological operations and the advantage of different models, he figured out a hybrid DNA-sticker model, which used not only the basic structure of sticker model but changed in the code manner by Adlem-Liption model's so that to use their operations as well. In Christian et als study [5], the plasmid DNA model was subtly combined with protein specialty which served for output. Inspired by all those combination ideas, in this paper, three DNA structures, the 3-armed, hairpin structures and the triple-stranded, are combined to form a new model realizing their virtus respectively to solve nearest-neighbour clustering.

We start with the description of those three DNA structures in section 2. Then in section 3, we introduce the nearest-neighbour clustering problem and how to transform into the MST. Later, the DNA algorithm and its procedures are proposed, with an instance to illustrate this algorithm in the following part. Discussions and conclusion are drawn in the last two sections.

2. Three Kinds of DNA Structures

2.1. K-Armed DNA Molecules

K-armed DNA molecules, of which the 3' ends there is a signal of helic extension, which is generally up to 30 to 45 base pairs of length, have good nature to show graph structures intuitively and can perform to solve computational problems at molecule level. In addition, they indeed exist in nature, like intermediates called Holliday in the process of homologous recombination of E. coli. [6] Actually, specially designed *k*-armed DNA structures, particularly 3-armed and 4-armed ones can be made stably and furthermore, under certain conditions 8-armed and 12-armed ones can be obtained [7]. In addition, both the double-helix part and the cleaved single strand can be encoded to present information. N. Jonoska and his partners [8] firstly used them to generate 3-dimensional DNA graph structures to solve 3-SAT and 3-VCP. In that design, the *k*-armed structure was used to present the corresponding vertex of *k*-degree, called vertex building blocks, and it is same in what Fang Gang and his partners [9] did to work out the edge connectivity. Then they proposed an improved algorithm by introducing evolutionary algorithm in 2007 [10], which could avoid theoretically the trouble of exponential increase of solution space in traditional DNA computations. In addition, the *k*-armed DNA structures with the good quality to intuitively represent the vertex in a graph have been used to form the actual graph structures to deal with the MCP [11]. In this paper, 3-armed molecules (in Figure1a) also serve as basic building blocks but with different encoding strategy.

2.2. DNA Hairpin Structure

DNA hairpin model was first proposed by Hagiya [12], and used to calculate the Boolean circuit. Winfree [13] made further study of this structure, calling it the terminator for polymerization. Later, Hagiya [14] took advantage of its self-government to solve 3-SAT problem. Martínez-Pérez [15], etc. realized Adleman's first experiment using hairpin model. In addition, here, the hairpin structures we need are the same ones as the cap structures in Jonoska's study [8]. In addition, its special structure shows in Figure. 1b.

2.3. Triple-Stranded DNA Structures

In 2004, Shigemori found that the double-helix DNA and oligodeoxyribonucleotides (ODN) could form a stable three-chain structure mediated by RecA-protein and ATPγS. Chinese scholar, Yang Jing [16] took use of triple-stranded DNA can be formed readily at any location in double-stranded DNA with a homologous deoxyoligonucleotide to solve a simple 0-1 integer programming problem. In addition, later many Chinese scholars developed and used this structure to solve SAT, process arrangements [17-18]. The technics has been changed somewhat so that it may decrease the error rate in DNA computation and be adopted to solve several problems in graph and combinatorial optimization. The key idea of using triple-stranded DNA model is that sequence-specific recognition of double-stranded DNA by homologous deoxyoligonucleotides is fulfilled when a single-stranded DNA complements to some part of the double-stranded fragment with which it can form into a triple-helix structure. And the triple-stranded DNA are then separated and the desirable double-stranded DNA strands can be retained and extracted after adsorbed by the magnetic particles coated with streptavidin. In addition, the three-stranded DNA structure is shown in Figure 1c.

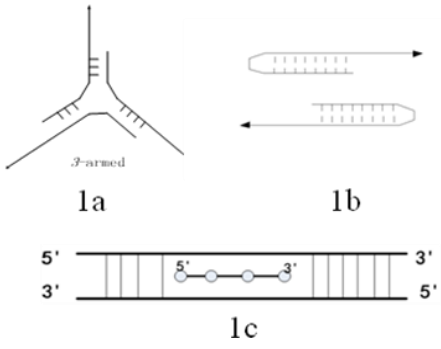


Figure 1. Three DNA structures mainly used in this study

3. Nearest-Neighbour Clustering

In this study, we focus on the cluster in a 2-dimensional plane, which is often used in practice, especially when encountering clustering problems in graph.

3.1. Problem Discribtion

A tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering. Moreover, figure 2 shows a dendrogram of five points. Firstly, each points are treated as one group. Then according to similarity scale, they are grouped step by step until all data points in one cluster. For example, when the similarity of two groups of points, {*a*, *b*} and {*c*, *d*, *e*}, is roughly 0.15, and at this time this similarity is the smallest, they are merged into one bigger cluster.

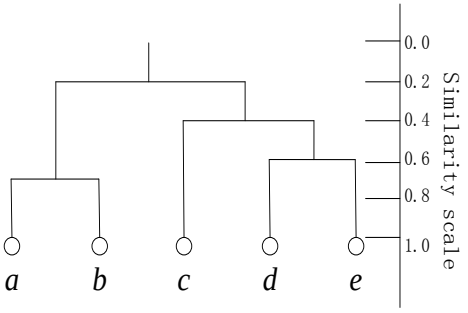


Figure 2. Dendrogram representation for hierarchical clustering of five points

In 2D data plane, Euclidean distance is used to measure the distance between two objects. In addition, the measure for distance between clusters is as follows, where $|p\text{-}q|$ is the Euclidean distance $d(C_i, C_j) = \min_{p \in C_i, q \in C_j} | p - q |$ between two p of C_i and q of C_j.

3.2. Transforming Of Nearest-Neighbour Clustering Into MST

MST is defined as following: Given a connected and undirected graph $G=(V, E)$, a spanning tree of G is a subgraph that is a sub-tree connecting all vertices together. E_{ij} is the edge joining vertex v_i and v_j together ($e_{ij} \in E$, v_i, $v_j \in V$). We assign a weight w_{ij} to e_{ij}, and define the weight of a spanning tree by computing the sum of all weights in that tree. A minimum spanning tree has the weight less than or equal to the others. The mapping from nearest neighbor clustering into MST is as follows:

If we treat the data points as vertices of a graph, with edges forming a path between the vertices in a cluster, then the merging of two clusters corresponds to adding an edge between the nearest nodes in two clusters. Because edges always go between distinct clusters, the resulting graph will generate a tree. Therefore, the nearest-neighbour clustering algorithm is also called a minimal spanning tree algorithm [1]. Note that, only integer distances are concerned here.

4. Solving the Nearest-Neighbour Clustering Problem

In this section, we propose an algorithm to solve the nearest-neighbour clustering. Actually, what we did is more likely to construct a binary tree. In addition, the 3-armed DNA molecule is the best choice to be the basic limb structure of binary trees. Firstly, we list the encoding related things as follows.

4.1. Encoding

Three arms of a 3-armed molecule are logically divided into two parts, two vertex arms and one distance arm. Let $V=\{v_1, v_2, \ldots, v_n\}$ be the vertices set, and w_{ij} is the corresponding edge weight between v_i and v_j. ($i=1,\ldots, n$; $j=i, i+1, \ldots, n$; $i \neq j$)

Vertex arms: each of them is used to respectively represent two vertices, v_i and v_j ($i \neq j$). In addition, those two vertex coding sequences lie in the cleaved ends.

Distance arm: As two arms are labelled with v_i and v_j, correspondingly, the distance w_{ij} is coded with ($w_{ij} \times k$) base pairs into the double-stranded part of the third one and k is a constant. The extened cleaved end is encoded complement to v_i or v_j. A restriction endonuclease sequence must be coded into the double-stranded section of distance arm if the distance is larger than the threshold.

For example, if the distance between A and B is two less than the threshold, 3-armed DNA molecules needed are in figure 3. The one with complement sequence of vertex A coded in the distance arm is located on the left. The one with B is located on the right. Two pairs of GC means the distance between A and B is two.

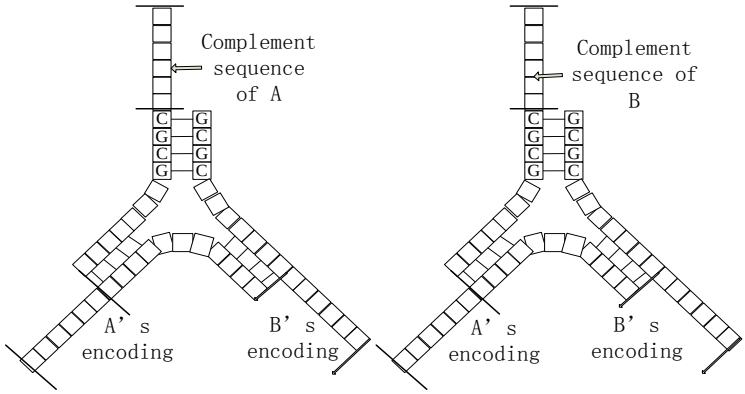


Figure 3. Two kinds of 3-armed DNA molecules coded with vertices A and B

Aim to cap off the cleaved ends of 3-armed molecules, the single-stranded parts of hairpin structures are encoded complement to vertices sequences.

In order to get structures with all vertices, encode single-stranded DNA fragments called verex strands complement to vertex sequences and then make them to be corresponding nucleoprotein filaments.

Encoding vertex probes complement to vertices. Because nanoparticles [19] of different sizes can be stimulated to show different colours by a single wavelength of light, fluorescence analysis can be used to detect the marked probes, which show the composition of clusters.

4.2. Biological Algorithm

Basic algorithm

Step1 Make the 3D DNA structures containing vertices and edges.

Step2 Retain and extract those 3D DNA structures which have all the vertices or data points.

Step3 Isolate the lightest ones from result solution of Step 2, and that is the MST structures.

Step4 Detect products from Step 4 to get the final classification result.

Biological procedures

Step1 Encode all the DNA structures and fragments following the instructions in the *encoding* part. Moreover, prepare nucleoprotein filaments (Mix and incubate the strands that make up the vertex strands and RecA protein in a reaction mixture which contain ATPγS under certain conditions. Thus, the nucleoprotein filament will come into being.). Label the vertices probes with different sizes of metal nanoparticles.

Step2 Combine multiple copies of all kinds of 3-armed DNA molecules with buffer solution and DNA connected enzyme to hybridize and ligate. Then put enough hairpin structures to close the cleaved ends. Thus, we could get data pool to be screened.

Step3 Remove those still with open ends, not fully matched and partly formed. Then remove redundant hairpin structures by using an *exonuliease* enzyme. Then purify the solution.

Step4 Remove by gel electrophoresis the covering graphs that are larger than the original graphs formed in the above steps. The details about covering graph are explained in the literature of Fang Gang [9].

Step5 Mix the nucleoprotein filaments of i-th vertex with the data pool of Step4 to get the triple-stranded DNA structures. Then mix the magnetic particles wrapped by biotin-streptavidin with the product. Due to the high affinity of biotin-streptavidin interaction, the formed triple-stranded DNA structures will be arrested by magnetic particles and isolated from the rest. After the elution and deproteinization, those desired double-stranded DNA contain the sequence of vertex v_i can be obtained.

Step6 Repeat Step5 from i equals 1 to n to get the spanning tree structures with all vertices.

Step7 Use gel electrophoresis on products in Step 6 to select the lightest 3D DNA structures and those are the MST structures.

Step8 Denature the results of Step7, and then amplified them by PCR.

Step9 Firstly, hybrid the amplified products with vertex probes marked by metal nanoparticles on the slide under certain conditions. Secondly, illuminate them by fibre-optic illuminator to identify the formation of colours scattered by different size metal nanoparticles. If k-1 parts of sites could be detected, k clusters could be got. Therefore, the final classification result can be read.

5. Example

For simplicity, an undirected complete graph $G=(V, E)$ with four vertices was given in figure 4. $V=\{v_1, v_2, v_3, v_4, v_5\}$ is the data set, $E=\{e_1, e_2, e_3, e_4, e_5, e_6\}$ is the edge set with each edge being distance edge between two vertices, and $W=\{w_1, w_2, w_3, w_4, w_5, w_6\}=\{1,2,4,9,6,8\}$ is the weights set, and the threshold $\tau=4$.

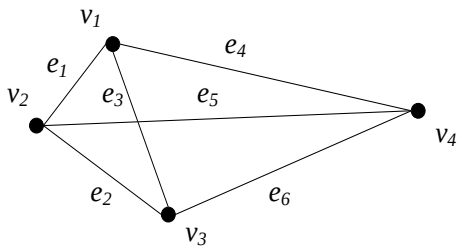


Figure 4. Graph G with four vertices

First of all, we encode 12 kinds of 3-armed DNA molecules and combine enough copies in a single tube to form all the possible 3D DNA structures. Then mix hairpin structures labelled with sequence v_i or those with complement sequence to vi to make the former 3D DNA structures to be closed. That is the data pool containing all the combinations of points and their linking edges. Then we screen those containing all the five vertices by constructing and then deconstructing triple-stranded DNA structures repeatedly. After that, use gel electrophoresis to get the lightest 3D DNA structures which is the MST. One possible structure is in figure 5. After detecting the color of enzyme site, two subtrees are obtained. Then read the combination of vertices colours, it has known that v_1, v_2, v_3 are clustered into one, and v_4 in the other. Moreover, the clustering result is shown in figure 7. Just note that, what figure 5 shows is just one possible 3D DNA structure that qualifies the minimum spanning tree, and we can detect other minimum spanning tree structures of DNA, like showing in figure 6. (In figure 5 and figure 6, two bold red lines represent our special designed DNA sequence of enzyme cutting sites, because the length of e_5 exceeds the fixed threshold.)

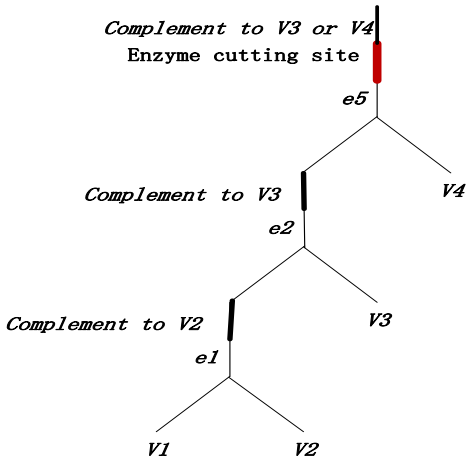


Figure 5. One possible MST structure been chosen by this algorithm

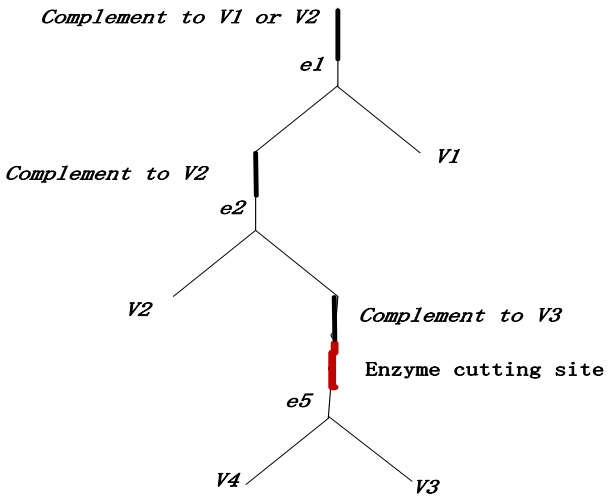


Figure 6. Another possible MST structure been chosen by this algorithm

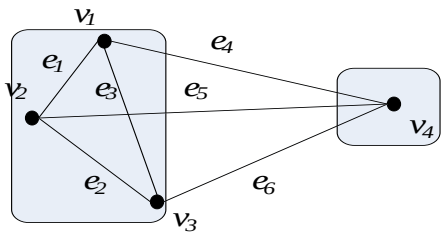


Figure 7. The clusting result of five points

6. Discussion

6.1. Some Notes on this Algorithm

The key point of the design here is that the vertices having not clustered in some group all have a chance to connect with any vertex in the group. According to the construction before, each vertex in the group has a cleaved end to connect to the others. Like the instance in figure.4, firstly, v_1 and v_2 were grouped together and then v_3 was added in the group. It could be seen that, when v_4 was going to be added, it could be connected with v_1, v_2 or v_3.

The cleaved open ends must be capped off by some special designed hairpin structures in order to get the MST by comparing their weights. Because if still being single strands, DNA of the same length may having different weight owing to different DNA bases composition.

Although the distance here between vertices are limited to be integer, in fact, we could consider using binary or quaternary coding to encode weight of real number under certain precision [20].

At the same time, we do not have specific experimental results, therefore, are unable to discuss the feasibility of procedure above. However, we have used limited number of operations, which can be easily implemented, in the laboratory.

6.2. Algorithm Complexity

According to the algorithm, what we have done all depends on a complete graph of n vertices. The quantity of different types of 3-armed DNA structures is $n(n-1)$, since each pair of two vertices corresponds to two 3-armed structures. Moreover, the hairpin structure types are $2n$ because of the complement of v_i sequence and v_i all have a chance to be uncovered. Moreover, n types of vertex fragments are needed to make up the triple-stranded structures. Therefore, the quantity of sequences needed to encode is $O(n^2)$.

Space complexity: Same as other DNA computing algorithms, the algorithm here also makes all the combinations of vertices. As a result, the space complexity is exponential. That means that for problem sizes larger than some value, the number of required initial strands might exceed a practical generated number. Thus, it is important to know that the proposed algorithms can be applied on the moderate and small graph cases.

Operational complexity: With the help of nanoparticle tagging the vertex DNA probes, detection of the result can be done in a few constant steps. Therefore, the main operation is to detect if all the vertices fragments were in the 3D DNA structures. According to algorithm design, the algorithm complexity is *O(n)*.

7. Conclusions

Although it is believed that the theoretical procedures here are empirically possible, there are also some technical problems to be studied further. Especially the cost of DNA experiment is so high that free of us get easy access to do in practice. Moreover, the amount of DNA molecules dealing with a real problem is always beyond our imagination. Despite of the troubles we face, DNA computing still has broad prospects. We are encouraged by the progress of DNA computing and the rapid development of biological technologies. Also together with the development of biology and chemistry, we can figure out more flexible and efficient access to problems.

Acknowledgements

This paper was finally supported by National Natural Science Foundation of China (61170038), Shandong Province Natural Science Foundation (ZR2011FM001), Humanities and Social Sciences project of Ministry of Education (12YJA630152).

References

1. Han, J., Kamber, M., Pei, J. (2012) *Data Ming, Concepts and Techniques*. 3nd Edn. Beijing: China Machine Press.
2. Adleman, L.M. (1994) Molecular computation of solutions to combinatorial problems. *Science.* 266(5187), 1021–1024,
3. Henkel, Ch.V., Bäck, Th., Kok, J.N., et.al. (2007) DNA computing of solutions to knapsack problems, *BioSystems*. 88(No 1-2), 156-162.
4. Sanches, C.A.A., Soma, N.Yo. (2009) A polynomial-time DNA computing solution for the Bin-Packing Problem. *Applied Mathematics and Computation*. 215(6), 2055-2062.
5. Henkel, Ch.V., Bladergroen, R.S., et.al. (2005) Protein output for DNA computing. *Natural Computing*. 4(1), 1-10.
6. Ma, R.I., Kallenbach, N.R., Sheardy, R.D., et.al. (1996) Three-arm nucletic acid junctions are flexible. *Nucleic Acids Res.* 14(24), 9725-9753.
7. Wang, X., Seeman, N.C. (2007) Assembly and characterization of 8-arm and 12-arm DNA branched junctions. *J. Am. Chem. Soc*. 129(26), 8169-8176.
8. Jonoska, N., Stephen, A.K., Saito, M. (1999) Three dimensional DNA structure in computing, *Biosystems.* 52(1-3), 143-153.
9. Fang, G., Zhang, S.M., Xu, J. (2006) Three dimensional DNA graph structure solution to edge connectivity. *Chinese Jonrnal of Aerospace Electronics Information Engineering and Control.* 28(1), 119-121.
10. Zhang, S.M., Fang, G. (2007) Three dimensional DNA structure solution to connectivity based on evolutionary algorithm. *Journal of Computer Engineering and Applications.* 43(7), 41- 44.
11. Ren, X., Liu, X. (2013) Three Dimensional DNA Graph Structure Solution to Maximum Clique Problem. *Proceedings of the 2012 International Conference on Information Technology and Software Engineering*, *Beijing, China*, *December 8-10, 2012.* LNEE, 211, 89-97.
12. Hagiya, M., Arita, M. (1998) Towards parallel evaluation and learning of Boolean formulas with molecules. *3rd DIMACS Workshop DNA Based Computers, Massachusetts Institute of Technology,* June 14-15, 105-114.
13. Winfree, E. (1998) Algorithmic self-assembly of DNA. PhD thesis, Caltec, Pasadena.
14. Sakamato, K., Gouzu, H., Kiga, D, et.al. (2000) Molecular computation by DNA hairpin formation.

Science. 288, 1223-1226.
15. Martínez-Pérez, I., Zhong, G., Ignatova, Z., Zimmermann, K.H. (2010) *DNA Computing Models.* New York: Springer-Verlag New York Inc.
16. Yang, J., Yin, Z. X. (2008) 0-1 integer programming problem based on RecA-m ediated triple-stranded DNA structure. *Chinese Journal of Computer Engineering and Applications.* 44(2), 76-79.
17. Fang, G., Xu, J., et.al. (2009) Performing DNA computation with RecA-mediated triple-stranded DNA structure. *Natural Computing.* 8(1), 149-156.
18. Yang, J., Yin, Z.X., Huang, K.F. (2012) The Traveling Salesman Problem Base on Triple-Stranded DNA Structure Model. *Informatics in Control, Automation and Robotics*. 133, 769-775.
19. Bao, H., Jia, C.P., Zhou, Z.L., et.al. (2009) A Novel DNA Detection Method Based on Gold Nanoparticle Probes and Gene Chips. *Acta Chimica Sinica*. 67(18), 2144-2148.
20. Rohani, B., Abu, B., Junzo, W. (2008) A Biologically Inspired Computing Approach to Solve Cluster-Based Determination of Logistic Problem. *Biomedical Soft Computing and Human Sciences.* 13(1), 59-66.

Received on the 21st of June 2013

Mathematical and Computer Modelling

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 74-82
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

INFLUENCE OF NONESSENTIAL POINT DAMAGE ON THE CROPPING TIME AND THE CROSS-SECTION QUALITY IN PRECISION CROPPING

L. Zhang, H. Xiang

College of Electromechanical Engineering
China University of Petroleum, Qingdao, 266580, China
Phone: (+532)-86983500. E-mail: zlj-2@163.com

The finite element software, ANSYS, is applied to build the three-dimensional simulation model of the slotted 0.45%C steel bar. The nonessential point damage and potential damage point in the bar caused by heat stress, which are very important parameters to evaluate the heat stress effect of V-shaped notch tip, are proposed in this paper and their mechanisms of action are analysed. The numerical simulation results show that the fine cracks in the inner of metal bar caused by heat stress are actually an annular damage strip, very close to the V-shaped notch tip. The damage life of bar material is also obtained under the action of heat stress. Compared with the existing experimental device, a new experimental device, which can better prefabricate the fine crack at V-shaped notch bottom of metal bar by heat stress, is built and the corresponding experiments of heat treatment are carried out. Based on it, the cropping experimental results show that the high-quality cropping cross-section is not only kept well, but also the cropping time is reduced remarkably compared with the traditional cropping method without the action of the heat treatment. In addition, the cropping time for the bars without applying the asbestos clapboard to conduct directly heat treatment is less than that for the bars without heat treatment.

Keywords: Precision cropping, nonessential point damage, heat stress, fine crack, damage strip

1. Introduction

The precision cropping of metal bar is widely used in industrial production, such as chain pins, micro-motor shafts and combustion engineer appurtenances and so on. High-speed shearing cropping, radial holding differential cropping, and blanking under the high axis load are effective and economical methods of bar cutting [1]. The shear behaviours of bar fine blanking and cropping processes through their experimental performance have been studied. In order to investigate the phenomena of shearing behaviour and effects during the cutting operation, some major parameters, such as the surface finish, hardness distribution are also examined by employing certain equipment [2]. Unlike these common shear methods, where a cropping die is used so as to cause the huge shearing force and the defective fracture surfaces, a low-stress cropping method is proposed where the gap effect and bend effect of annular V-shaped notch of bar surface are considered as factors to improve breakability and surface finish and to decrease cutting force and tool wear. In the cropping process, V-shaped notches are cut at specific length intervals in metal bar to engender the stress concentration at the bottom of the V-shaped notch. Then the slotted bar is fed to crop under the action of continuous bending fatigue force and would rupture after certain periods. It is obvious that the stress concentration degree of V-shaped notch bottom has a direct relation to the surface quality of rupture section and the cropping time. However, by large numbers of the low-stress cropping experiments, it is confirmed that the cropping cross-section quality can be kept high, but the cropping time is still longer because the radius at the V-shaped notch bottom cannot be sufficiently small, where the V-shaped notch is mainly obtained by general turning method. On the basis of the low-stress cropping, the method of prefabricating the fine crack at V-shaped notch bottom by heat stress was proposed and the influences of water velocity, the width of the adiabatic boundary, cooling water temperature and initial heat temperature of bar on heat stress at the V-shaped notch bottom have been also analysed [3, 4]. However, the studies mentioned above are based on the two-dimensional analytical model of bar in the numerical simulation and do not give the mechanism of action of heat stress prefabricating the fine crack at V-shaped notch bottom. In addition, the very quantificational experimental results are also not verified because of the simplification of simulation model and the defects of existing experimental device. Therefore, the three-dimensional finite element model of the slotted bar is built and the corresponding experimental device is improved in this paper. The nonessential point damage and the potential damage point in the bar caused by heat stress are proposed. The damage life of bar material under the action of heat stress is also obtained. The three-dimensional simulation results show that the fine cracks in the bar caused by heat stress are actually an annular damage strip, very

close to the V-shaped notch tip. The cropping experimental results show that the cropping time, by prefabricating the fine crack at V-shaped notch bottom, is reduced remarkably and the high-quality cropping cross-section of the bar is still kept well compared with the cropping method without the action of heat stress. The experimental results also testify the usefulness of the asbestos clapboard.

2. Mechanism of High-Speed Circumferential Low-Stress Bending Fracture Cropping

The mechanical parts of the cropping system with high-speed circumferential low-stress bending fracture mainly consist of the transmission system, the blanking die, the fine-tuning system, the hydraulic clamping and the feeding mechanism, and the body. The working mechanism of blanking die and fine tuning system is shown in Fig.1. When the cropping machine works, the variable frequency motor drives the column hammer mounted on the eccentric spindle by the two-stage accelerative transmission system. The column hammer made by antifriction composite provides the bar with circumferentially continuous eccentric loading. The ball screw controlled by the AC servo motor drives the feed sleeve to adjust the compressive deformation of the column hammer to make the crack of V-shaped notch tip extend flexibly and regularly. In the course of cropping, the clamping die can confine the radial movement, the radial torsion and the axial torsion of metal bar. The feed-in device, by which the arm of the exciting force can be adjusted, can restrict the axial movement of metal bar.

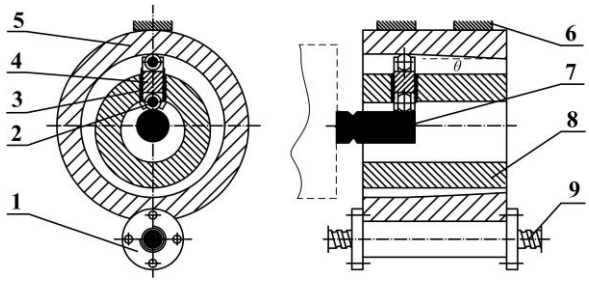


Figure 1. Working mechanisms of blanking die and fine tuning system
1. Nut and accessory, 2. Roller, 3. Ball sleeve, 4. Column hammer, 5. Feed sleeve,
6. Sliding guide, 7. Bar, 8. Eccentric spindle, 9. Ball screw

3. Numerical Simulation Analysis of Heat Stress Prefabricating the Fine Crack

3.1. Fundamental of Prefabricating the Fine Crack at V-Shaped Notch Bottom

The basic principle of heat stress prefabricating the fine crack of V-shaped notch bottom is mainly based on heat expansion and cold retraction effect of metal material. As a result of temperature gradient along the axial different parts of the bar, the heat tension stress occurs in the inner of the slotted bar. The heat tension stress in the vicinity of V-shaped notch tip is maximal because of stress concentration effect. The fine crack in the vicinity of V-shaped notch tip will appear when the heat tension stress reaches the ultimate strength of the metal material [5].

According to above principle, the cooling mode of the metal bar in heat treatment is shown in Fig.2. The bar with annular slots is heated to the settled temperature firstly, about 500℃, and the V-shaped parts of the bar are protected with asbestos clapboards and then the bar is cooled rapidly by cooling water. In the course of cooling, the temperature gradient between the inner and surface of the bar and that between V-shaped notch part and the bar surface acted on the cooling water will appear. By setting reasonably parameters, the fine crack will occur in the vicinity of V-shaped notch tip firstly due to the action of three-dimensional stress of the V-shaped notch.

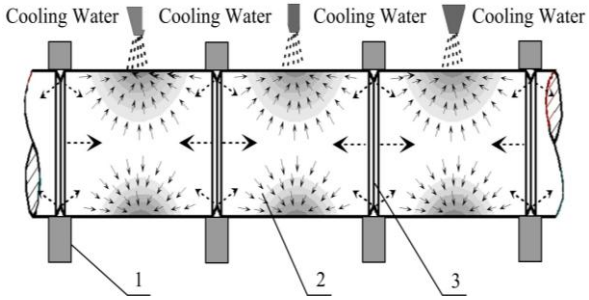


Figure 2. Cooling mode of the metal bar

3.2. Simulation Model and its Boundary Conditions

Because the axial heat tension stress of V-shaped notch tip is mainly analysed, V-shaped notch part of the metal bar is not regarded as the boundary of simulation model of metal bar. The approximate differential equation of heat conduction is given by

$$\rho c\frac{\partial T}{\partial \tau}=\frac{1}{r}\frac{\partial}{\partial r}(\lambda r\frac{\partial T}{\partial r})+\frac{\partial}{\partial z}(\lambda\frac{\partial T}{\partial z}), \tag{1}$$

where ρ is density of metal bar, c is specific heat, T is bar node temperature in either time, τ is cooling time, λ is heat conductivity coefficient, r and z are polar coordinates.

Initial condition of the bar:

$$T(r,z,0)=T_0, \tag{2}$$

where T_0 is initial heating temperature of the bar.

$$T(0,0,\tau)=\text{limited value} \tag{3}$$

Boundary condition of the heat convection:

$$\lambda(\frac{\partial T}{\partial n})_w=-h(T-T_1)\ (r=r_0, l\le z\le 20-l;\ r=r_0, -(20-l)\le z\le -l), \tag{4}$$

where h is coefficient of mean heat convection load, n is direction vector and T_1 is temperature of cooling water.

Boundary condition of the heat insulation:

$$q_w=-\lambda(\frac{\partial T}{\partial n})_w=0\quad (r=r_0, 0.4\sqrt{3}\le z\le l;\ r=r_0, -l\le z\le -0.4\sqrt{3};$$
$$(r_0-0.8)^2+z^2=0.04,\ r_0-1\le r\le r_0-0.9;\ z=\frac{\sqrt{3}}{3}(r-r_0)+0.4\sqrt{3},\ r_0-0.9\le r\le r_0;$$
$$z=-\frac{\sqrt{3}}{3}(r-r_0)-0.4\sqrt{3},\ r_0-0.9\le r\le r_0), \tag{5}$$

where q_w is heat condition, r_0 is radius of bar, l is width of adiabatic boundary.

The expression of mean heat convection load coefficient h is calculated as [6]

$$h=0.5c_1(2ur_0/v_m)^{n_1}\,\mathrm{Pr}_m^{1/3}\lambda_m/r_0, \tag{6}$$

where c_1 is coefficient, u is water velocity, v_m is kinematic viscosity of cooling water, Pr_m is prandtl number of cooling water, and λ_m is heat conductivity coefficient of cooling water.

The finite element software, ANSYS, is applied to build the non-steady temperature field model of the slotted bar. The three-dimensional simulation model of slotted metal bar is shown in Fig. 3. When the metal bar is heated to 500℃, the material of V-shaped notch tip will go over the elastic region and bring the plastic deformation. Therefore, the bilinear kinematic hardening model of material needs to be defined in the process of finite element analysis. The tangential slope ET in the bilinear kinematic hardening model can be obtained by

$$E_T = \frac{\sigma_b - \sigma_{0.2}}{\delta - 0.2\%}, \tag{7}$$

where σ_b is ultimate strength of material, $\sigma_{0.2}$ is yield stress in the condition of plastic strain 0.2% and δ is elongation percentage.

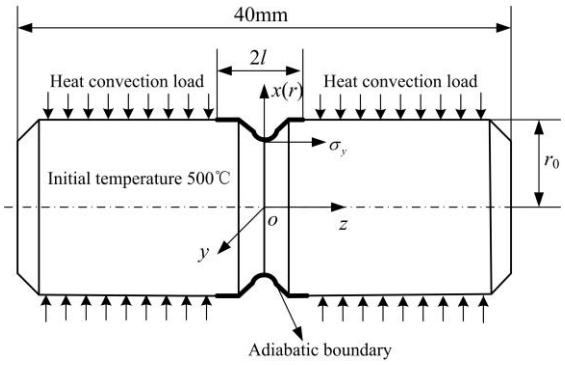


Figure 3. The three-dimensional analytical model of the bar
1. Asbestos clapboard, 2. Slotted bar, 3. V-shaped notch

Because the heat stress in the inner of the bar is jointly affected by the temperature field and the structure field, the three-dimension solid element SOLID 87 in ANSYS software is applied to divide the mesh of the model. The whole simulation model has 28804 nodes and 17360 elements. Both two ends of the model are the heat insulation boundary of the displacement coupling and the heat insulation boundary of the displacement restriction respectively. In addition, the initial temperature of every node in simulation model is 500℃ and the heat convection loads only act on the outer surface of bar except V-shaped notch part. Based on it, the Von Mises equivalent stress distribution nephogram near V-shaped notch for 45 steel bar is shown in Fig.4. As is shown in Fig.4, the maximum stress occurs in the vicinity of V-shaped notch bottom, but not in the tip of V-shaped notch.

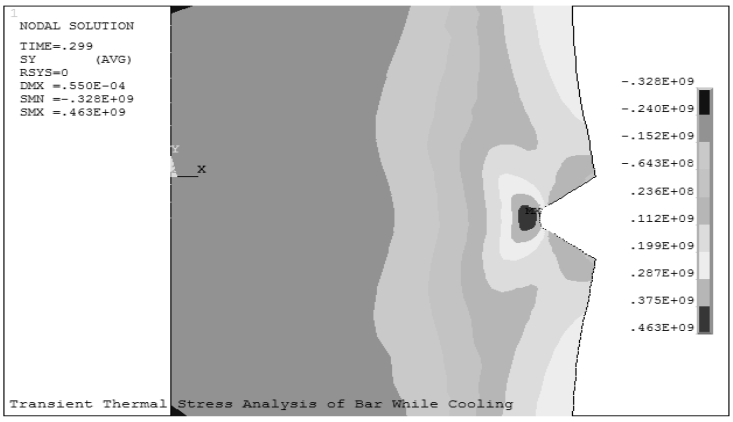


Figure 4. Von Mises equivalent stress distribution nephogram

4. Damage Mechanism of Heat Stress Prefabricating the Fine Crack

4.1. Nonessential Point Damage Caused by Heat Stress in the Inner of the Bar

The inherent damage in the bar, such as micropore, is called essential point damage. The damage caused by external load in the bar is called nonessential point damage. Therefore, the damage in the bar caused by heat load is the nonessential point damage. The influence of nonessential point damage on the material of the bar in this paper is that when the axial tension stress at some point near V-shaped notch tip arrives at ultimate strength of the material, the material at the point is damaged actually and the fine heat crack is also considered to occur here. In the cooling process, the axial tension stresses of other points around the first actual damage point can also reach ultimate strength of the material and hence the material at these points can also be damaged. Some points around the first actual damage point, whose axial tension stresses can not reach ultimate strength of the material, are called potential damage points. Because every point in the bar is not separate, the potential damage points are affected to a great extent by actual damage points. Therefore, the potential damage points caused by heat stress also exert an important influence on the damage life of the bar under the action of external fatigue loads. The influence can not appear in this circulation, but it can occur in next circulation. The actual damage and potential damage belong to the nonessential point damage mentioned above.

4.2. Damage Analysis of Heat Stress Prefabricating the Fine Crack at V-Shaped Notch Bottom

As an example of a section of 45 steel bar, its initial heat temperature is 500℃, its total length is 40mm, depth of V-shaped notch is 1mm, flare angle is 600, bottom radius is 0.2mm, and radius of the bar is 7.5 mm. The temperature of the cooling water is 20℃.

When the water velocity is 4m/s and the width of adiabatic boundary is 2mm as shown in Fig.3, the numerical simulation results show that the projective positions of node 6989, node27762, node7269, node 28416 and node 8240 in the bar in XZ plane located at V-shaped notch in ANSYS software are shown in Fig.5 and these nodes are regarded as the analytical objects. Node 36 is the tip of V-shaped notch, node 6989 is the representative node. The nodes of 27762, 7269, 28416 and node 8240 are closest to the representative node 6989. The numerical simulation results show that the axial tension stress at node 6989 in the bar as shown in Fig.5 is the first to arrive at ultimate strength of 45 steel in corresponding node temperature. The axial heat tension stresses at node 6989, 7269 and node 28416 also arrive at ultimate strength of 45 steel in current condition. In addition, the axial tension stresses at node 27762 and node 8240 do not reach ultimate strength of 45 steel under the present condition; hence, the heat cracks at the two nodes do not appear. However, the two nodes are the potential damage points, and node 8240 is closest to the tip of V-shaped notch and node 27762 is farthest to the tip of V-shaped notch. The cropping experimental results show that these potential damage points can also play an active part in reducing the cropping time. From the above analysis, the fine cracks in the bar caused by heat stress are actually an annular damage strip, very close to the V-shaped notch tip.

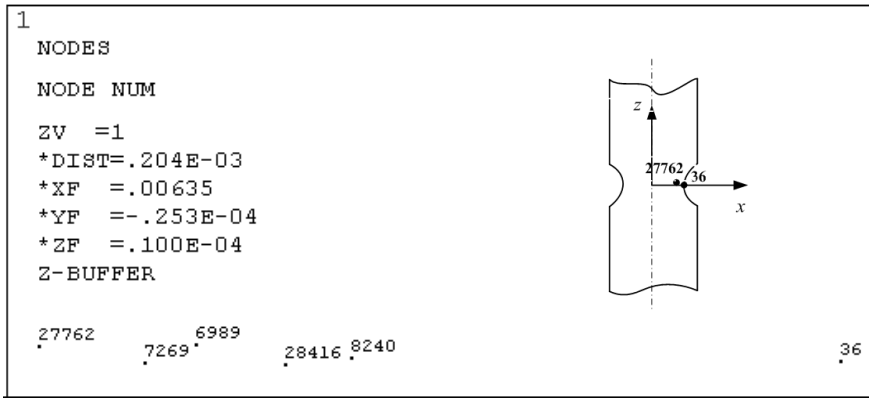


Figure 5. The projective positions of representative nodes in XZ plane in ANSYS

Under the different water velocity and width of adiabatic boundary, the widths of annular damage strip mentioned above are shown in Table 1. From the Table, under the same water velocity, when the

fine crack appears, the width of annular damage strip increases first, and then decreases, hence there is an optimal width of adiabatic boundary theoretically to make the damage degree reach a maximum. Under the same width of adiabatic boundary, the width of annular damage strip increases with the increase of water velocity. When the water velocity is smaller, the heat crack will not occur.

TABLE 1. The widths of annular damage strip (mm)

Water velocity($m \cdot s^{-1}$)	Width of adiabatic boundary *l* (mm)				
	1	2	3	4	5
0.5					
1					
2	0.0342	0.0461	0.0402		
3	0.0686	0.0982	0.0725	0.0523	
4	0.0786	0.1140	0.0932	0.0758	
5	0.0837	0.1218	0.1109	0.0896	0.0634

5. The Damage Life of the Material Caused by Heat Stress

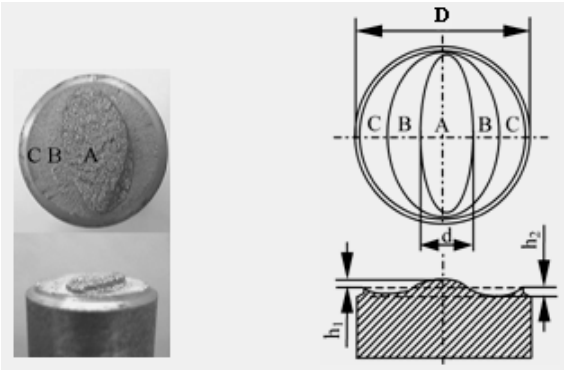
The damage life of the material under the action of heat stress is defined as the time from no actual damage in the bar to the formation of damage strip at the V-shaped notch bottom. When water velocities are 0.5m/s, 1m/s, 2m/s, 3m/s, 4m/s and 5m/s respectively and the widths of adiabatic boundary are 1mm, 2mm, 3mm, 4mm and 5mm respectively, the damage life is shown in Table 2. From the table, under the same width of adiabatic boundary, when the fine crack appears, the damage life decreases with the increase of water velocity; under the same water velocity, the damage life increases with the increase of the width of adiabatic boundary. It is an optimal choice that the water velocity is as large as possible and the width of adiabatic boundary is least, almost equal to half width of V-shaped notch.

TABLE 2. The damage life in the inner of the bar (s)

Water velocity($m \cdot s^{-1}$)	Width of adiabatic boundary *l* (mm)				
	1	2	3	4	5
0.5					
1					
2	0.204	0.323	0.501		
3	0.180	0.305	0.512	0.660	
4	0.168	0.287	0.473	0.658	
5	0.157	0.264	0.456	0.651	0.763

6. The Method of Cross-Section Quality Assessment

A method to assess the bar cross-section quality was proposed according to the characteristic of the obtained cropping cross-section and the typical cross-section photo is shown in Fig. 6(a) [7]. There are three parts in fracture surface. The part ‘A’ section is the region of final fracture, which should be as narrow as possible in practice. The part ‘B’ and part ‘C’ are produced by the slow crack initiation and crack propagation, and the two regions are smooth. The measurement method is shown in Fig. 6(b), where d is the width of the part ‘A’ region and h1 and h2 are the height of bulge and dent in cross-section respectively. In addition, to further assess the cross-section quality of the bar with different geometrical diameters, the ratio i of the final fracture to extended region is also proposed. For the bars with the same diameters, the less d, h1, h2 and i are, the higher the cross-section quality of the bar is.



(a) Photo (b) Parameters definition.
Figure 6. Parameters in quality assessment of bar cross-section.

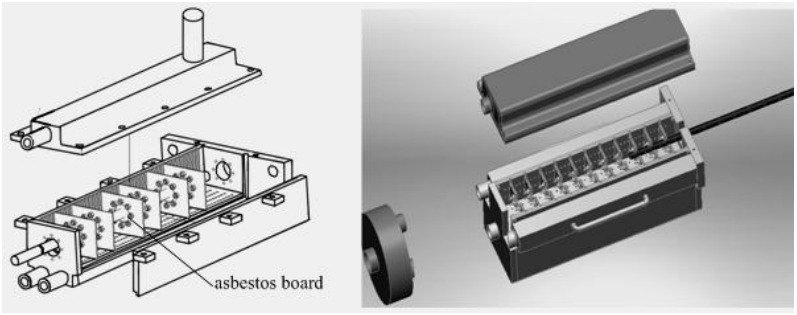
7. Experimental Results and Discussion

7.1. Cropping Experiments for 45 Steel Bars

As an example of sections of 45 steel bars, the water velocity is 5m/s, the width of adiabatic boundary is 0.618mm, which is equal to half width of V-shaped notch, and the geometrical parameters of the bar are the same as ones of the bar in the course of numerical simulation. The cropping experimental installation with the high-speed circumferential low-stress bending fracture presented in the paper is now in the process of development, but the cropping method has the same static mechanics model with the cropping method with variable frequency vibration [8, 9]. In the cropping method with variable frequency vibration, the relationship between the exciting force and the excited frequency has been acquired and the influences of five types of excited frequency control curves on the cropping cross-section and the cropping time have been investigated. Therefore, to further prove the validity of heat stress prefabricating the fine crack, the following experiments are carried out in the cropping machine with variable frequency vibration.

For the experimental device of heat stress prefabricating the fine crack, the structure of original experimental device is shown in Fig.7 (a). Although the device can carry out the task of heat stress prefabricating the fine crack, there are still some defects. For example, the bar heated to 500℃ is very difficult to pass through many of asbestos board hole, which brings the inconvenience of bar feeding. In addition, only the upper surface of the bar is first sprayed by cooling water because there is only a water inlet as shown in Fig.7 (a). All these will bring the experimental errors. Therefore, the experimental device of heat stress prefabricating the fine crack is improved and designed as shown in Fig.7 (b). The asbestos plate is divided into two parts, and one part is movable. The water inlet is increased to 4. Based on it, the feeding of metal bar is very convenient and the cooling effect of bar surface is improved.

The initial excited frequencies are 25Hz and 30Hz respectively and the cropping experimental data results are shown in Table 3. The cropping results of bar cross-section quality are shown in Fig. 8 and Fig.9 respectively. From table 3 and the two figures, by prefabricating the fine crack at the bottom of V-shaped notch with heat treatment, the high-quality cross-section is not only kept well, but also the cropping time is reduced remarkably compared with the traditional cropping without the action of the heat treatment. In addition, in the course of cropping, the exciting force is always applied to different action points on the surface of every segment of the bar along the circumferential direction of the bar. It is also shown that the cropping time for every segment of the bar is basically same at a time. The cropping experimental results further proves that the fine cracks in the inner of metal bar caused by heat stress are actually an annular damage strip, very close to the V-shaped notch tip.

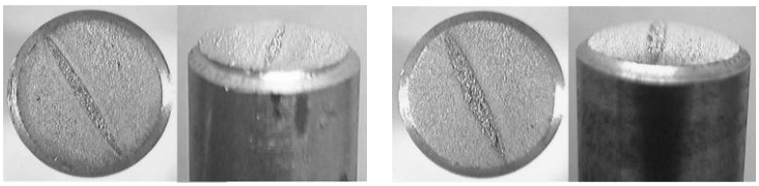


(a) Original experimental device (b) Improved experimental device

Figure 7. Experimental device schematic of heat stress prefabricating the fine crack

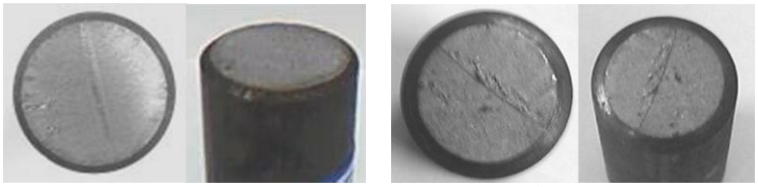
TABLE 3. Cropping experimental results for 45 steel bars without heat treatment and with heat treatment

Whether the heat treatment exists	Initial excited frequency (Hz)	Time (s)	q (mm)	h_1 (mm)	h_2 (mm)	i
yes	25	20.02	1.1	0.24	0.06	0.0924
no	25	26.06	1.2	0.23	0.06	0.1016
yes	30	27.14	1.2	0.25	0.08	0.1016
no	30	32.65	1.3	0.25	0.07	0.1110



(a) 25 Hz (b) 30 Hz

Figure 8. The cropping cross-sections for bars without heat treatment



(a) 25 Hz (b) 30 Hz

Figure 9. The cropping cross-sections for bars with heat treatment

7.2. The Necessity of the Asbestos Clapboard

To further analyse the necessity of the asbestos clapboard and the influence of the nonessential point on the precision cropping, the heat treatment experiments are conducted many times without applying the asbestos clapboard. The experimental results of the cropping time for 45 steel bars are shown in Table 4. From the table, it is testified again that the cropping time is reduced by applying the asbestos clapboard to protect the V-shaped notch parts. According to Table 3 and Table 4, it is also shown that the cropping time for the bars without asbestos clapboard to conduct directly heat treatment is less than that for the bars without heat treatment. The author thinks that it is the potential damage points in the bar that play an important role in heat stress prefabricating the fine crack.

TABLE 4. Experimental results of the cropping time without asbestos clapboard and with asbestos clapboard

yes	25	20.02
no	25	25.28
yes	30	27.14
no	30	30.35

8. Conclusions

The cropping time and the cross-section quality directly affect the productive efficiency, the production cost and the performance of the parts. Therefore, a method by which the cropping time can be reduced remarkably and the cross-section quality can also be kept well, is improved in this paper.

According to the above research, the following conclusions may be drawn:

1) The nonessential point damage and the potential damage point caused by heat stress, which are very important parameters to evaluate the heat stress effect of V-shaped notch, are proposed in this paper and their mechanisms of action are analysed in detail.

2) The three-dimensional finite element model of the slotted bar is built. The simulation results show that the fine cracks in the inner of metal bar caused by heat stress are actually an annular damage strip, very close to the V-shaped notch tip. Under the same water velocity, when the heat crack at the V-shaped notch bottom appears, the width of annular damage strip increases first, and then decreases, hence there is an optimal width of adiabatic boundary theoretically to make the damage degree reach a maximum. However, the optimal width of adiabatic boundary is very difficult to obtain because of constraints of grid size and grid quantity. Under the same width of adiabatic boundary, the width of annular damage strip increases with the increase of water velocity.

3) It is optimal choices in the experiment of heat stress prefabricating the fine crack of V-shaped notch bottom that the water velocity is as large as possible and the width of adiabatic boundary is least, almost equal to the half width of V-shaped notch. Based on it, the damage life is very little.

4) The cropping time for the metal bars without asbestos clapboard to conduct directly heat treatment is less than that for the bars without heat treatment to some degree. This further proves that the potential damage points in metal bar can play an important role in the course of heat stress prefabricating the fine crack of V-shaped notch tip region.

5) Large numbers of experimental results show that the method of heat stress prefabricating the fine crack of V-shaped notch bottom has the great advantage of reducing the cropping time, but its effect on the cropping cross-section quality is not very obvious.

Acknowledgements

The authors wish to express gratitude to the National Natural Science Foundation of PR China (Approval no. 51105382), Shandong Natural Science Foundation of PR China (Approval no. ZR2011EL039) and Qingdao Development Zone Science and Technology Development Project (Approval no. 2011-2-50), which have supported this work.

References

1. Chen, J.D., Wang, Y.W., Yu, D.H., et.al. (1992) Brittle precision cropping of metal materials. Int. J. Mach. *Tools Manuf.*, 32(3), 415-424.
2. Chan, L.C., Lee, T.C., Wu, B.J., et.al. (1998) Experimental study on the shearing behaviour of fine-blanking versus bar cropping. *J. Mater. Process. Technol.*, 80-81, 126-130.
3. Zhao, S.D, Zhang, L.J., Lei, J., et.al. (2007) Numerical study on heat stress prefabricating ideal crack at the bottom of V shaped notch in precision cropping. *J. Mater. Process. Technol.*, 187-188, 363-367.
4. Zhang, L.J., Zhao, S.D. (2008) Influence of cooling water temperature and initial heat temperature of bar on heat stress at the bottom of V-shaped notch. AIP Conference Proceedings, 1060(1), 208-211.
5. Min, H., Yue, Y., Ling, W. (1999) Mechanics of materials, Xi'an: Xi'an Jiaotong University Press.
6. Ma, Q.F. (1979) Heat Transfer, Beijing: Peoples Education Press.
7. Zhang, L.J., Zhao, S.D., Lei, J., et.al. (2007) Investigation on the bar clamping position of a new type of precision cropping system with variable frequency vibration. *Int. J. Mach. Tools Manuf.*, 47(7-8), 1125-1131.
8. Zhang, L.J., Zhao, S.D., Hua, C.J., et.al. (2008) Investigation on a new type of low-stress cropping system with variable frequency vibration. *Int. J. Adv. Manuf. Technol.*, 36(3-4), 288-295.
9. Hua, C.J., Zhao, S.D. (2009) Research on the computer measurement and control system of the cropping method with variable-frequency vibration and low energy-consumption. *Machinery Design & Manufacture*, (9), 72-74.

Received on the 21st of June 2013

Mathematical and Computer Modelling

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 83-93
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

EXPERIMENTAL STUDY ON DRAG REDUCTION OF A SHIP DUE TO A DRAINAGE SLIT

A. M. Liu, J. M. Wang

School of Transportation, Wuhan university of Technology
No.1040, Heping Street, Wuhan, China
Phone: (+86)-02786544877. E-mail: lamfluid@126.com

In order to reduce the ship's resistance, the drag reduction effect is figured out by a drainage slit on a medium or low speed ship in wind tunnel. The double-model was adopted as the testing ship model. Two cases including the original ship model and that with a drainage slit from bow to stern were investigated. The velocity distribution for the ship models were measured by five-holes Pitot tube, wall surface pressure were obtained through the silicon-based micro pressure sensors, and the viscous drag force was measured using the self-made strain gauge as well. The test results showed that the flow velocity of around the ship model with the slit changed not considerably comparing with the original ship model except near the inlet and outlet of the slit. However, the pressure coefficient decreases obviously on the bow and varies on the stern; the drag reduction ratio was about 5%.

Keywords: drag reduction, pressure measurement, force measurement, velocity measurement

1. Introduction

As is well known, when the fluid flows past a blunt body, because of the boundary layer separation, the average pressure on the rear of the body is considerably less than that on the front. Thus, a pressure drag, which composes the most of the total viscous drag in some cases is developed. In the light of these results, we can assume that the total drag maybe decrease if the pressure difference between on the rear of the body and that on the front reduces by means of some method. The most straightforward way perhaps is to open a flow passage between the front stagnation point and the rear stagnation point. Whereupon the pressure on the front portion of the body probably decreases and that on the rear portion may increases, therefore the total drag may be decreased. From the published literatures the flow passage is commonly slit or hole.

Izutsu [1] carried out fluid force measurements for a circular cylinder with a slit, and Yajima and Sano [2] investigated the drag reduction of a circular cylinder due to double rows of hole. These researches indicated that the reduction effect is good. Moreover, Tan [3] studied the flow around a hollow circular cylinder with several splits and the drag reduction mechanism by using the hydrogen bubble flow visualization technique. The experimental result demonstrated that the drag force decreased obviously and the attack of a split can hardly affect the drag of the cylinder.

As to medium-speed or low-speed ship, especially to plump ship, the viscous drag can reach an amount up to 80% of the total drag. Accordingly, it is important to reduce the viscous drag as far as drag reduction of ship is concerned. Chen et al. [4] carried out research on drag reduction of a ship due to leading edge hole using numerical simulation method. The results showed that if the position and form of flow passage reasonably is chosen, the pressure drag decreases drastically and then the total viscous drag reduction ratio reaches about 2%. Based on the above research results, in order to make sure of the effect of this kind of the drag reduction method, an experimental study of drag reduction was made in this paper.

2. Experiment Arrangement

Experimental research is carried out in wind tunnel. The total drag force, wall surface pressure and velocity distribution for the ship models were respectively measured at Reynolds number of $Re=1.8\times10^6$ (based on the length of waterline of the testing ship model). The Strain gauge was employed to measure total drag force. The ship surface pressure distribution was obtained through the pressure sensor. The velocity distribution around the ship model was obtained by the five-hole Pitot tube.

2.1. Ship Model

In the present experiment, according to the results of Chen et al. [4], one two-tailed ship with bulbous bow and no rudder were selected as the test ship. For the wind tunnel tests, a model ship was

made of wood with a scale ratio of 1/97.5, which was 0.8 m in the length of waterline and 0.16 m in the molded breadth and 0.041 m in the molded draft. Table 1 and figure 1 provide the principal particulars and the body plan of the ship. Two cases including the original three dimensional ship model and that with a drainage slit from bow to stern are shown in figure 2. The inlet of the drainage slit is at the high pressure region of bow, and outlet, below stern transom plate (shown in figure 3). The projection area of inlet in the direction, which is perpendicular to the incoming velocity, equals that of outlet, i.e., they are 11.87% of the ship's maximal cross-sectional area, or 0.45% of ship wet area. It is important to point out that a double-model is adopted, that's to say, two same ship models stick together at the water plane.

TABLE 1. Principal particulars of the test ship

Particular	Lpp(m)	B(m)	Draft(m)	Wetted surface area(m^2)
Prototype	78	31.2	4	1592
Model	0.8	0.32	0.041	0.1675

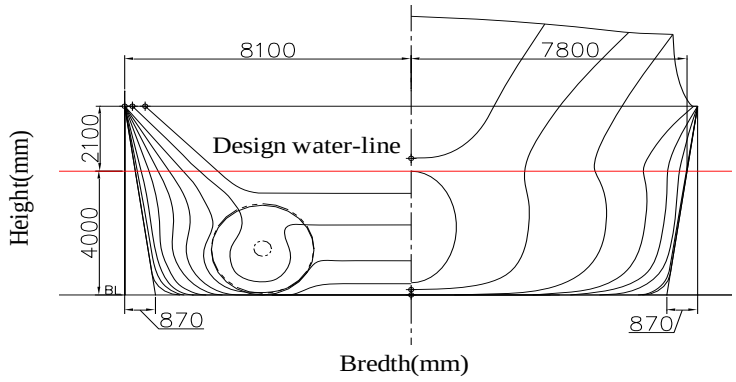
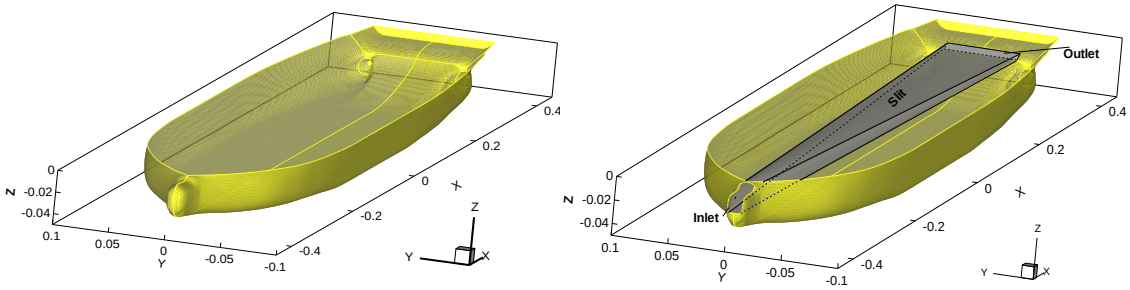


Figure 1. The body plan of the ship(mm)



Case 1. Original ship model

Case 2. Ship model with a drainage slit from bow to stern

Figure 2. Ship models

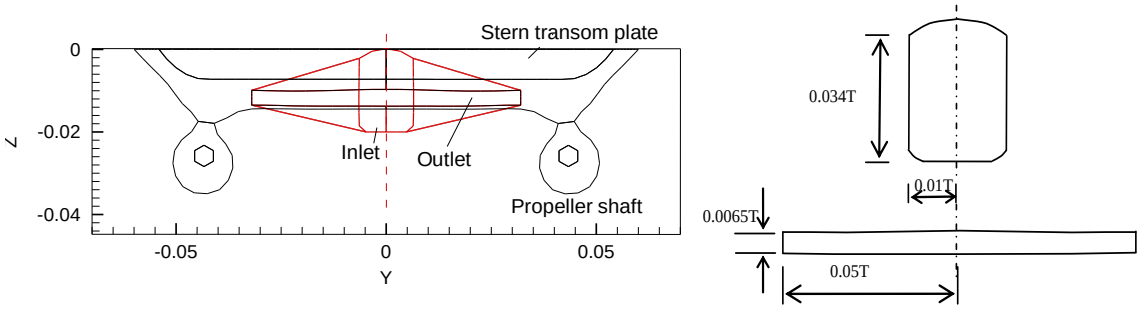


Figure 3. Inlet and outlet of the slit (T is draught)

2.2. Wind Tunnel

The measurement of mean velocity components and pressure were carried out in a single circuit wind tunnel with open test section. The open diameter is 1 meter and the length of the test is 2 meter. Blockage coefficient, defined as the ratio of sectional area of the model and the wind tunnel, was less than 1.6% for the model ship. Because test section of the wind tunnel is open, no correction is made for the blockage effect. The turbulence intensity of the wind tunnel is less than 1%. For the wind tunnel tests, a double-model ship was used. The model ship was fixed in the middle of the test section of wind tunnel. A traversing mechanism driven by three servomotors with a precision of 1 mm for each direction was installed for efficient and accurate positioning of the measurement probes. Experimental facilities and double-model are given in figure 4.

In the wind tunnel tests, similarity law of Re was complied. That is to say, Reynolds number in the wind tunnel tests must be equal to that in navigating conditions. This means the wind velocity must be very high. However, it is very difficult in the wind tunnel tests. While from Qin [5] experimental results we can see that the viscosity drag coefficient is not changed basically when Reynolds number is larger than 1.0×10^6. Since the wind speed was 32 m/s, the resulting Reynolds number ($\mathrm{Re}=\frac{UL}{\nu}$) was 1.84×10^6. On the other hand, the aim of the experiment is mainly that the drag reduction mechanism of a ship due to a slit is studied by comparison of the drag between case 1 and case 2. Therefore, as long as the Reynolds number of these two cases is the same, the experimental requirements are satisfied. Here, L is the length at design waterline of the prototype and the model ship, respectively. ν is the kinematic viscosity of water or air. All measurements were carried out in the fixed condition. It is noteworthy that the average temperature of the tunnel air during the experiment was 5°C, although it changed slightly every day.

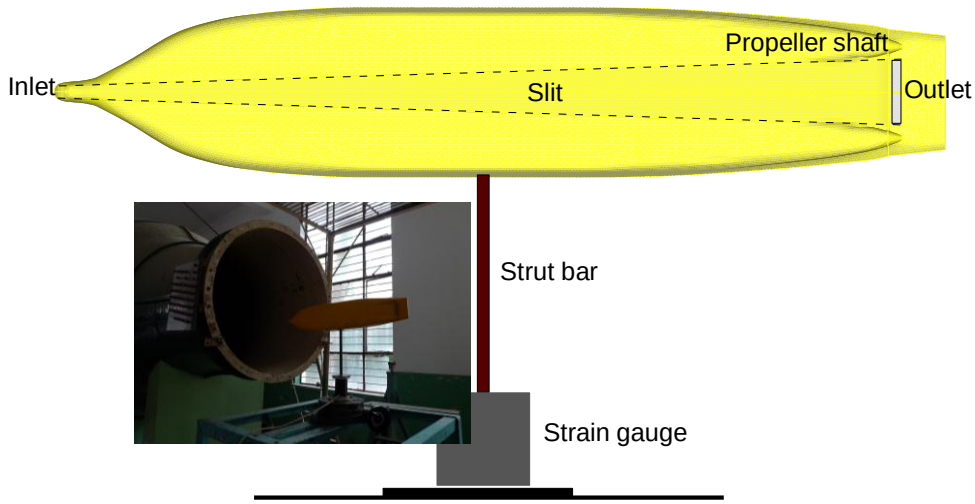


Figure 4. Sketch of experimental facilities and double-model

3. Experimental Results

3.1. Local Mean Velocity Measurements

For the measurements of the three-dimensional mean velocity field around stern and bow region, a five-hole Pitot tube was utilized. The diameter of pressure hole is 0.25mm and that of Pitot tube, 2.5mm as well. The measuring range of velocity is 0m/s~50m/s and the accuracy is ±0.025%FS.

The fixed coordinates of ship are designated (x, y, z), as shown in figure 2. The origin of the coordinates is located at the intersection of the midship and waterplane. Where x denotes the downstream direction, y is transverse and to starboard, and z is the upward direction. x coordinate is nondimensionalized by the length between perpendiculars (Lpp) of the model ship, but y and z coordinate are nondimensionalized by the draft of the model ship.

For the sake of understanding how the velocity near the inlet and outlet of the slit change, velocity

measurements for the original ship were carried out at 4 cross sections, while the same measurements with a drainage slit were performed at the same sections. Section 1 (x/Lpp =-0.58) and 2 (x/Lpp =-0.49) is located at upstream and downstream of the slit inlet respectively, while section 3(x/Lpp =0.46) and 4(x/Lpp =51), upstream and downstream of the slit outlet.

Two dimensional velocity vectors at different cross sections are given in figures 5-12. These figures indicate that the velocity of the cross section change very slightly at bow and stern after opening the slit except near the inlet and outlet of the slit. Because the inlet and outlet projection area perpendicular to the incoming velocity are the same, the velocity of the outlet varied slightly comparing with that of inlet, that has to say, the difference between the velocity of the outlet and that of ambient flow around the bottom of the ship is not considerable. Therefore, the change of the wake velocity is slight after opening the slit. Furthermore, due to the propeller shaft, circumfluence exists in the wake of case 1 as well as case 2. In addition, the area of the circumfluence is becoming larger and larger along downstream direction. In general, the flow velocity is not changed obviously after opening the slit.

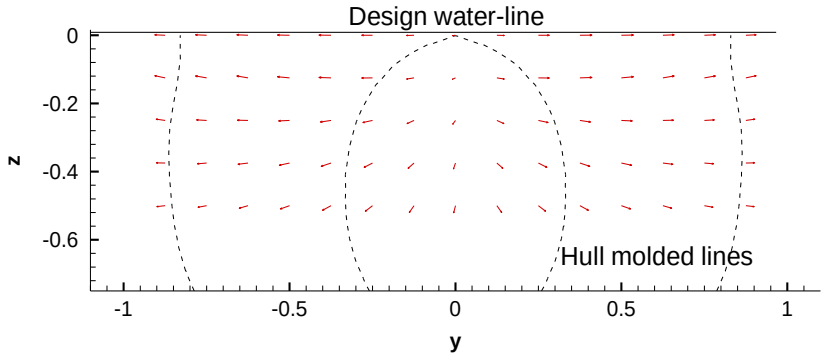


Figure 5. Velocity vector at section 1 (x/Lpp=-0.58, no slit)

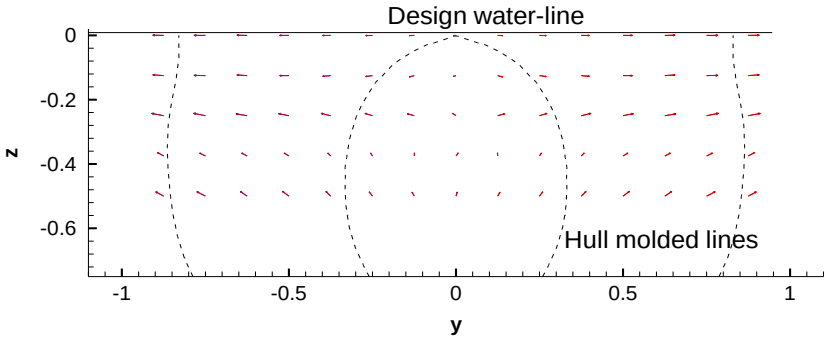


Figure 6. Velocity vector at section 1 (x/Lpp=-0.58, with slit)

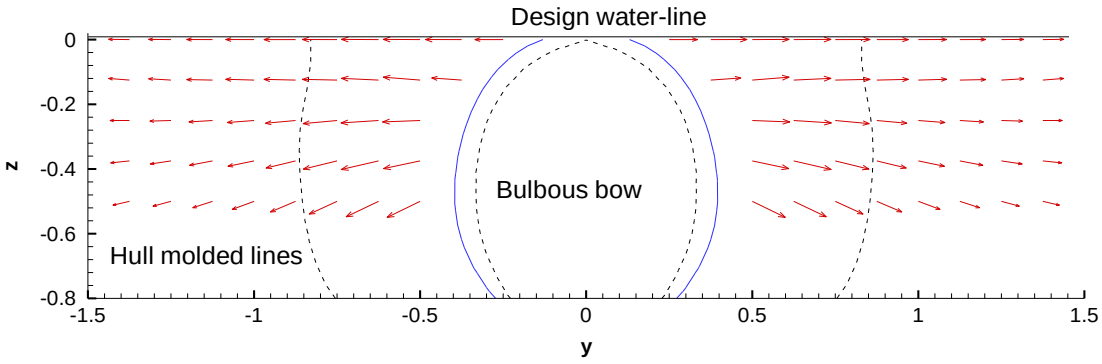


Figure 7. Velocity vector at section 2 (x/Lpp=-0.49, no slit)

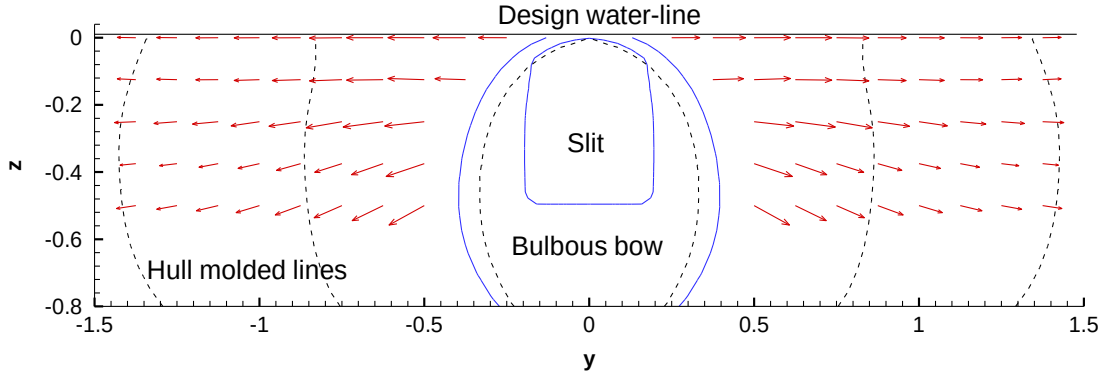


Figure 8. Velocity vector at section 2 (x/Lpp=-0.49, with slit)

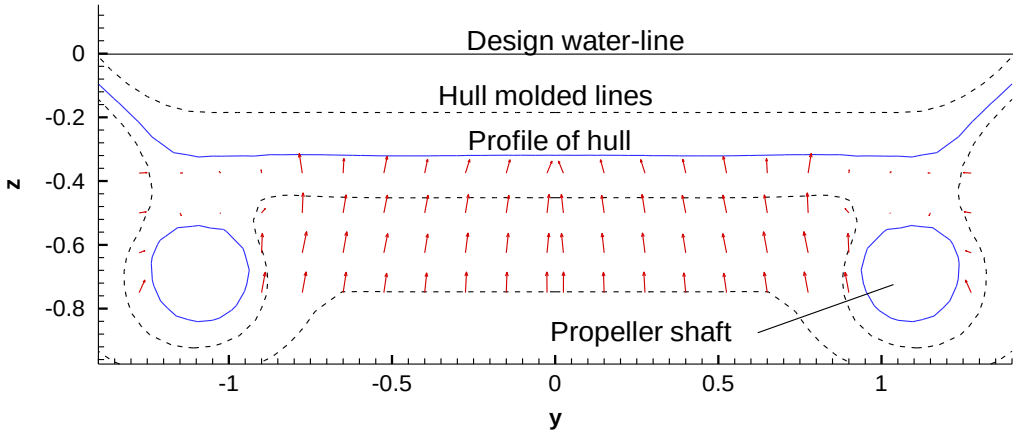


Figure 9. Velocity vector at section 3 (x/Lpp=0.46, no slit)

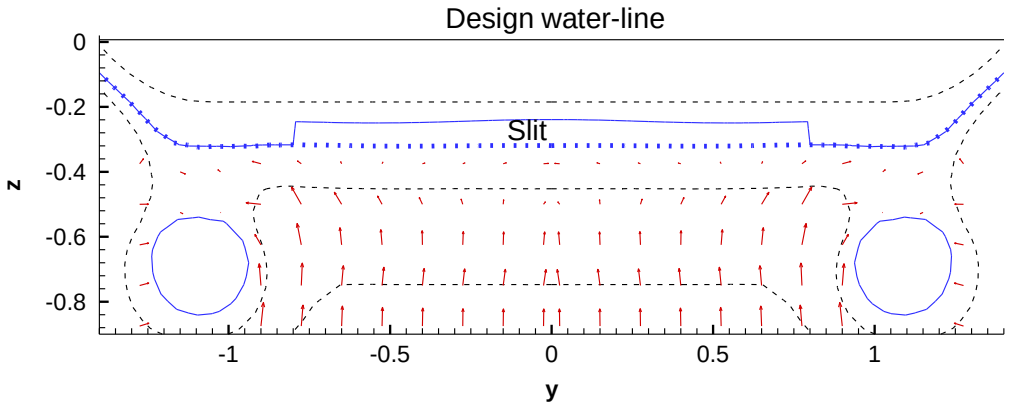


Figure 10. Velocity vector at section 3 (x/Lpp=0.46, with slit)

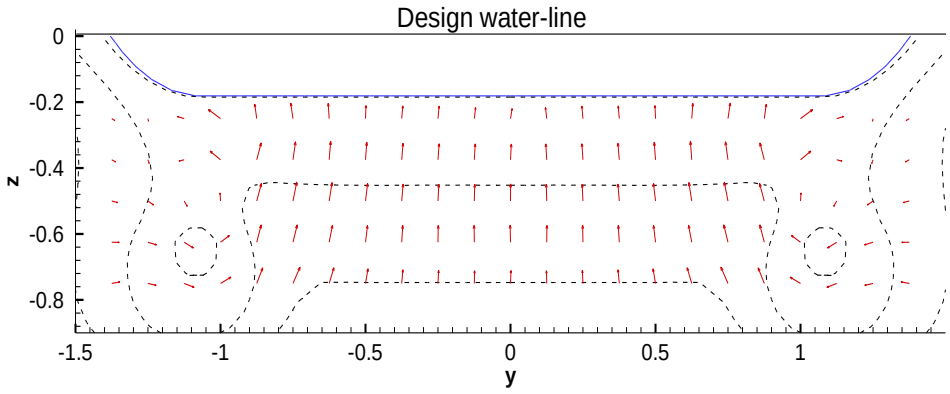


Figure 11. Velocity vector at section 4 (x/Lpp=0.51, no slit)

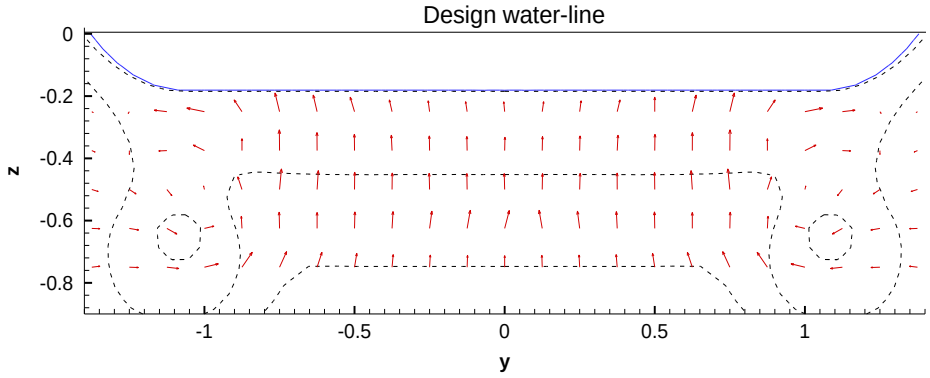


Figure 12. Velocity vector at section 4 (x/Lpp=0.51, with slit)

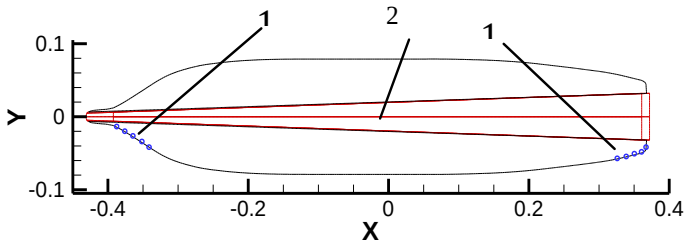
3.2. Pressure Measurement

In order to explore the mechanism for the drag reduction, the ship surface pressure distribution was obtained through the silicon-based micro pressure sensors. The pressure measurement ranges from 0Pa to1000Pa and the accuracy is ±0.5%FS. The diameter of pressure hole is 1mm.

Considering the greater change of pressure just happen near the inlet and outlet of the slit, 8 pressure holes are set up at bow and 6 pressure holes, at stern of the ship body surface, where is close to the inlet and outlet of the slit(shown in figure 13). In addition, here the waterline is z=-0.27. The location of the experimental measuring point is listed table 2. Experimental results of pressure coefficient of Cp are given in figure 14.

TABLE 2. The location of the measuring point (z=-0.27)

	bow								stern					
Points No.	1	2	3	4	5	6	7	8	1	2	3	4	5	6
x/L_{pp}	-0.54	-0.525	-0.51	-0.485	-0.47	-0.455	-0.44	-0.425	0.4	0.4125	0.425	0.4375	0.45	0.4625



1. measuring point; 2.slit
Figure 13 .Location of the measuring point (z=-0.27)

As a whole, pressure of eight points at bow decreases along x direction, but at stern increases gradually, primarily because that velocity is becoming larger and larger with the increasing of the ship cross section area at bow along x direction, while the contrary is at stern.

Except for three points (NO.1 and 2 at bow, NO.6 at stern) where are very close to the inlet or outlet of the slit, Cp of measuring point changes slightly after the slit is opened. The Cp (point NO.1) on the bow of original ship model is close to 0.7. However, Cp at bow of case 2 decreases to about 0.25, which that pressure coefficient at bow reduces clearly due to the slit. However, the change of Cp at stern before and after opening the slit is less than that on bow. Therefore, the pressure difference between bow and stern reduces due to the slit. Consequently, the conclusion that the pressure drag decrease can be made.

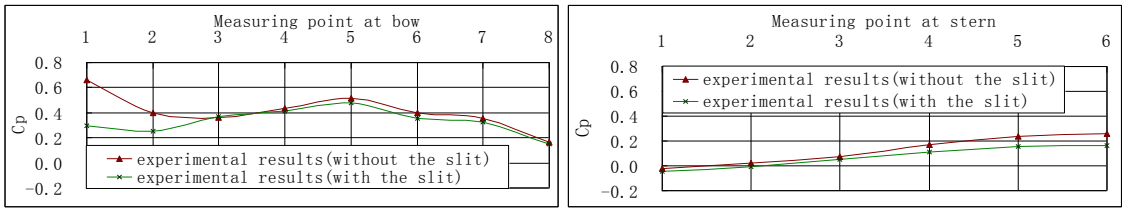


Figure 14. Cp of the 10 measuring point

3.3. Viscous Drag Measurement

In order to make sure of the effect of drag reduction due to the slit, viscous drag measurements are performed. The self-made strain gauge was employed to measure drag force. The force measurement ranges from 0g to 1Kg and the accuracy is ±0. 1%FS. Experimental drag of two cases is 166.1g and 157.5g, and the corresponding coefficient C_f is 0.0076 and 0.0072 respectively, so the reduction rate which is defined as follows is about 5%.

$$M_f = (C_{fi} - C_{f0}) / C_{f0} \tag{1}$$

Here, subscript 0 and 1 represent the data of case 1 and case 2 respectively.

After opening the slit the wet area increase, the friction drag may be increased. However, the pressure drag decrease obviously, i.e., the decreasing of the pressure drag is larger than increasing of friction drag. Thus, the end result is that the total viscous drag force decrease. This also is the reason of the drag reduction due to a drainage slit.

4. Conclusions

The following conclusions can be made:

(1) Because the difference between the velocity of the outlet and that of ambient flow around the bottom of the ship is not considerable, the change of the wake velocity is slight after opening the slit.

(2) The change of Cp at stern after opening the slit is less than that on bow. Therefore, the pressure difference between bow and stern reduces due to the slit. Consequently, the pressure drag decreases.

(3) The test results show that the drag force of the ship model with drainage slit decreases. The drag reduction ratio is greater than 5%.

(4) The free surface is not considered during carrying out the experiment, which is not consistent with the actual conditions. Therefore further research work should be done.

It can be well-recognized that in the financial models of human behaviour outlined above the company acts as an active player. Thus, the active system under consideration is a single-level one.

Acknowledgements

The present research was supported by the Natural Science Foundation of Hubei Province (Grant No.2011CDB248) and National Natural Science Foundation of China (Grant No.NSFC-10972167).

References

1. Izutsu, N. (1993) Fluid Force Measurements for a Circular Cylinder with a Slit, *Nagare*, 12, 293-301.
2. Yajima, Y., Sano, O. (1996) A Note on the Drag Reduction of a Circular Cylinder due to Double Rows of Hole, *Fluid Dynamics Research,* 18, 237–243.
3. Tan, G., Wang, J., Li, Q. (2001) Drag Reduction Technique of Cylinder and Mechanism Research, *Journal of Beijing University of Aeronautics and Astronautics*, 27(6), 658–661. (*In Chinese*)
4. Cheng, K., Wang, J., Wei, W. (2012) Numerical Simulation on Leading Edge Holes for Drag Reduction and Flow Field Characteristics of the Craft. *Journal of Ship Mechanics*, 16(5), 472-476. (*In Chinese*)
5. Qin, J. (2007) Numerical Simulation of Viscous Flow Around Low-speed Full Form Ship. Wuhan: Wuhan University of Technology Press.

Received on the 21st of June 2013

Mathematical and Computer Modelling

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 91-101
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

THE COMPREHENSIVE EVALUATION MODELS OF ECO-SLOPE BASED ON AHP AND FUZZY THEORY

X.-B.Xiong, J.-M. Li*, H.-F. Dai*

Nantong University
**Jinggangshan University; China*
Phone: (+86)-13328090772. E-mail: thongtao2006@163.com

This paper investigates the factors influencing the quality of slope vegetation on highway and then constructs three comprehensive evaluation models. In the evaluation, three functions of slope vegetation as main factors are analysed respectively and corresponding indices are selected. Five levels are set up for evaluation and standards of sub-factors are given by the experts. In the first model, the weights of main factors and sub-factors are determined by Analytic Hierarchy Process (AHP), and the fuzzy comprehensive evaluation is applied to analyse the overall performance. In the other two models, the overall performance is evaluated by fuzzy integral. The research data of six herbs is used to examine these models, and the analysis shows that the third model can overcome the confusion of ranking in the first model and heavy information loss in the second model and its evaluation result is consistent with previous research.

Keywords: slope vegetation, Analytic Hierarchy Process, fuzzy comprehensive evaluation, fuzzy measure, fuzzy integral.

1. Introduction

The highway is an important infrastructure of the country, but its construction and operation have also brought a series of neglect influence on the area along highway, mainly reflecting in the following aspects: changes in topography, disturbing of soil structure, destruction of plant communities, reduction of biological diversity, local climate impact, increasing the risk of water loss and soil erosion. Without certain protective measures, not only the safety of expressway's facilities and traffic safety cannot be guaranteed, but also the ecological environment of surrounding areas. The slope is an important part of expressway, and application of slope vegetation can prevent slope exposure, fracture and deformation effectively by reducing the water loss and soil erosion, thus protecting the structure of road, contributing a lot to the extension of service life and protection of roadbed. The green vegetation along the highway can not only reduce or eliminate the noise and toxic contaminants from vehicles, but also purify the air, thus forming a clean, environmentally friendly and comfortable driving environment.

There have been numerous investigations into the ecological effect of highway. Barry et al. investigated the variation of biotic communities in a southern Ontario stream along the highway [1]. Reijnen et al. studied the relation between addition of noise pollution from highway and the decrease in breeding bird populations [2]. Ullmann surveyed the exotic species in roadside vegetation in southern New Zealand, and the result of survey revealed that the range of such species is liable to expand, particularly in the habitat-complex provided by roadsides [3]. Forman et al. evaluated the effect of roads with different traffic volumes on surrounding avian distributions with its importance relative to other variables [4].

Palmer et al. explored the reliability of visual quality ratings of landscapes and the validity of photographic representations used when making these ratings [5]. Viles et al. analysed the degree to which the 'greenway' and 'connectivity' concepts can be applied in an integrated approach to managing road reserves and the adjacent landscape to facilitate the extension of ecological corridors and mitigate adverse environmental effects of the road network [6]. Clay et al. evaluated descriptor variables used by agencies to assess scenic quality along roads in their jurisdiction [7].

The mechanism of root reinforcement function of the slopes has been widely studied. Xie proposed complex μx method, and applied it to test the force of pulling roots [8]. Osman and Barakbah suggested that soil water content (SWC) and root length density (RLD), which could be used as indicators of slope stability and had the potential to predict slope failure in the future [9].

Influence of slope vegetation in the areas of ecology, landscape and structural stability has been well documented, however, there is few research undertaken to evaluate the overall effect of the highway in the round. The purpose of this paper is to set comprehensive evaluation systems to assess slope vegetation from its ecology, landscape and stability functions.

2. Methodology

2.1. Weight Obtained by AHP Analyses

Analytic Hierarchy Process (AHP) has become an important method to determine the weight of indices since it has been proposed by T. L. Satty in 1970s. AHP aims to settle the conflict between practical demand and scientific decision-making, and it aims to find a way to blend process of qualitative analysis and quantitative analysis. It has been widely used to determine influence of different factors in areas of maintenance policy selection, resources planning, environmental decision making and management of conflict [10]. It is necessary to estimate the weight of indices that influence the overall evaluation of the highway reasonably, and AHP analyses can provide a credible result.

The weight of factors and sub-factors of quality evaluation can be calculated based on AHP as follows:

1.The Pair-wise comparison matrix is constructed.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}, \tag{1}$$

where, x_{ij} is the scale of factor (or sub-factor) T_i comparing with factor (or sub-factor) T_j, and $x_{ij}=(x_{ji})^{-1}$.

2. The maximum eigenvalue λ_{max} and corresponding eigenvector w of the matrix are obtained.

3. The weights are calculated by the normalization of eigenvector w.

2.2. Evaluation Vector Calculated by Fuzzy Comprehensive Evaluation

Fuzzy math theory was first set by Zadeh in 1965. It was specifically designed to mathematically represent uncertainty and vagueness and provide formalized tools for dealing with the imprecision intrinsic to many problems [11]. Fuzzy comprehensive evaluation is an important branch of fuzzy theory, and has become a practical assessment method in many areas.

Based on the weight parameters determined, the process of a fuzzy comprehensive evaluation is shown as follows:

1. The factors ($B_i, i=1,2,\cdots,n$) and sub-factors ($C_{ij}, j=1,2,\cdots m_i$) that influence the quality evaluation are determined after analysing the problem.

2. An evaluation set V is established to judge levels of the factors and sub-factors, for example, the set $V = \{ v_1(excellent), v_2(good), v_3(general), v_4(bad), v_5(worse) \}$.

3. A series of criterions are set up to evaluate levels of sub-factors.

4. The evaluation vectors of sub-factors C_{ij} are calculated as follows:

$$C_{ij} = \frac{1}{N_{ij}}(N_{ij}^1, N_{ij}^2, \cdots, N_{ij}^k) = (f_{ij}^1, f_{ij}^2, \cdots, f_{ij}^k), \tag{2}$$

where, N_{ij} is the total number of samples, N_{ij}^k is number of samples which belong to level k, f_{ij}^k is the frequency that the sample belongs to level k.

5. The membership matrix R_i can be obtained by Eq. 3

$$R_i = (C_{ij})1_{1\times m_i} = \begin{pmatrix} C_{i1} \\ C_{i2} \\ \cdots \\ C_{im_i} \end{pmatrix} \tag{3}$$

6. The evaluation vector of the factor E_i can be calculated by Eq. 4

$$E_i = W_{Bi} R_i , \tag{4}$$

where W_{Bi} is the weight vector of sub-factors.

7. The overall evaluation can be calculated by Eq. 5

$$EV = W_A E = W_A \begin{pmatrix} E_1 \\ E_2 \\ \cdots \\ E_n \end{pmatrix} \tag{5}$$

where W_B is the weight vector of factors.

2.3. Evaluation Based on Fuzzy Measure and Fuzzy Integral

Traditional theories of multi-criteria decision making with weighted arithmetic mean is widely applied in various domains, but the interaction between factors (sub-factors) remains unsettled. In these traditional methods, the effects of factors are considered independent and the coordination of these effects is not considered in the determination of weights. In the analyses of practical problems, the evaluation result provided by weighted arithmetic mean methods may deviate from common sense significantly [12]. The coordination of factors can be achieved with the introduction of fuzzy measure and fuzzy integral. Since the concept of fuzzy integral was first defined by Sugeno in 1974, the theory has been successfully used in many decision making problems [13].

The fuzzy integral is constructed on the basis of fuzzy measure proposed to consider the superadditivity or subadditivity of indices [12]. The relevant concepts are presented as follows [13]:

Fuzzy measure:

Fuzzy measure g is defined on the power set $\beta(X)$ of set X, and satisfies the following properties:

$g : \beta(X) \to [0,1]$

(a) $g(\phi) = 0, g(X) = 1$;

(b) if $A, B \in \beta(X)$ and $A \subset B$, then $g(A) \le g(B)$;

(c) if $F_k \in \beta(X)$ ($1 < k < +\infty$), and the sequence $\{F_k\}$ is monotone, then $\lim_{k \to +\infty} g(F_k) = g(\lim_{k \to +\infty} F_k)$

The λ-fuzzy measure g_λ is a fuzzy measure with the following properties:

$\forall A, B \in \beta(X), A \cap B = \phi$;

$g_\lambda(A \cup B) = g_\lambda(A) + g_\lambda(B) + \lambda g_\lambda(A) g_\lambda(B)$, $(-1 < \lambda < +\infty)$;

If $X = \{x_1, x_2, \cdots, x_n\}$, fuzzy density $g_i = g_\lambda(\{x_i\})$, then

$$g_\lambda(\{x_1, x_2, \cdots, x_n\}) = g(X) = \frac{1}{\lambda}[\prod_{i=1}^{n}(1 + \lambda g_i) - 1] = 1 \tag{6}$$

If $\lambda > 0$, then $g_\lambda(A \cup B) > g_\lambda(A) + g_\lambda(B)$, which means the fuzzy measure is of superadditivity; If $\lambda < 0$, then $g_\lambda(A \cup B) < g_\lambda(A) + g_\lambda(B)$, which means the fuzzy measure is of subadditivity; If $\lambda = 0$, then $g_\lambda(A \cup B) = g_\lambda(A) + g_\lambda(B)$,which means the evaluation of the set $\{A, B\}$ equals to the sum of the evaluations for set A and B. When fuzzy density of elements in the set is determined, the coefficient λ can be calculated by Eq.6

Fuzzy Integral

Fuzzy measure g is defined as above $A \in \beta(X)$, the Sugeno integral of a satisfaction function $f : X \to [0,1]$ with respect to g is defined as [12]:

$$(s)\int f\mathrm{d}g = \sup_{\alpha\in[0,+\infty)} [\alpha \wedge g(A\cap F_\alpha)], \tag{7}$$

where $F_\alpha = \{x \mid f(x) \geq \alpha\}$.

The set X ($X = \{x_1, x_2, \cdots, x_n\}$) is finite, then the index i of element x_i is permuted so that $0 \leq f(x_{(1)}) \leq f(x_{(2)}) \leq \cdots \leq f(x_{(n)}) \leq 1$, the integral can be calculated by following equation:

$$(s)\int f\mathrm{d}g = \bigvee_{i=1}^{n} [f(x_{(i)}) \wedge g(A_{(i)})], \tag{8}$$

where $A_{(i)} = \{x_{(i)}, \cdots, x_{(n)}\}$.

The Choquet integral based on the concept introduced by Choquet in capacity theory is proposed by Murofushi and Sugeno [14].

$$(c)\int f\mathrm{d}g = \int_0^\infty g(F_\alpha)d\alpha \tag{9}$$

with the same notation as above, the integral can be calculated by following equation:

$$(c)\int f\mathrm{d}g = \sum_{i=1}^{n} (f(x_{(i)}) - f(x_{(i-1)}) g(A_{(i)}). \tag{10}$$

The evaluation of different plans can be provided by comparison of their value of fuzzy integral with reasonable definition of the satisfaction function.

3. Fuzzy Comprehensive Evaluation Model Based on AHP

A. Establishment of factors and sub-factors for highway quality evaluation

The evaluation of slope vegetation protection of highway is a basic method to ensure quality of the project, relating to road engineering, plant ecology, restoration ecology, grassland ecology, soil and water conservation science, landscape science theory etc. Quality evaluation index system of the vegetation slope protection should be established reasonably to compensate for flaws when isolated theories of different disciplines are combined, hence the result can be convincing.

1) Ecology index (B_1)

The original landform and stability of ecology system can be influenced with the excavation during the construction of expressway, and a series of environmental problems arise in the construction and operation, such as soil erosion, pollution, biodiversity reduction etc. The increased noise levels and toxic contaminants from vehicle are other two significant environmental effects of highways, and are considered a nuisance to human. The ecological function of the slope vegetation lies on alleviating the detrimental impact of highway on ecological environment, such as suction noise control, air purification, water and soil conservation, and maintaining biodiversity along the highway. According to experts from related fields and previous researches, the ecology indices can be selected as follows:

Growth rate (C_{11}): the average increment in the height of plant per day. The growth rate reflects the growth situation of the plant.

Coverage (C_{12}): the percentage of area covered by vegetation. The coverage can reflect the biomass of plant aboveground objectively, and high level of coverage can prevent erosion of slope and decrease the surface runoff.

Disease resistance (C_{13}): the disease resistance is an important index to value the quality of revetment vegetation. The disease resistance is divided into 5 levels: (1). excellent: without pathological changes; (2). good: area with pathological changes accounts for less than 10%; (3). general: the percentage of area with pathological changes range from 10% to 30%; (4). bad: the percentage of area with pathological changes range from 30% to 70%; (5) worse: area with pathological changes accounts for more than 70%. Their corresponding value ranges are V_1= {9, 7~8, 5~6, 3~4, 0~2}.

Drought resistance (C_{14}): the drought resistance is an index to reflect vegetation's ability to resist drought, and can also be divided into 5 levels.

Arid resistance (C_{15}): the arid resistance can be valued by colour and situation of growth in field of different nutritional levels.

Planting density (C_{16}): the number of plants per square centimetre.

Summer survival rate (C_{17}): the percentage of plants survives after high temperature in summer. In summer, high temperature and drought are the most important factors that influence the growth of plants. The activity of root system will be inhibited at high temperature, thereby affecting shoot growth, causing leaf wilting to die due to a lack of water in stem. The environmental conditions of slope are relatively adverse, and only the plant with heat resistance can survive.

2) Landscape index (B_2)

While driving, landscape appreciation occurs during a constant and rapid progression along a linear path. Roadside vegetation contributing a lot to the aesthetic quality of highway landscapes, and can harmonize the visual characteristics of roadway corridors with their context to improve the general quality of roadways [15]. As the roadside vegetation provides drivers aesthetic enjoyment, it also improves comfort and safety of the driving. According to experts from related areas and previous researches, the landscape indices can be selected as follows:

Colour of vegetation (C_{21}): the colour of vegetation can reflect the growth status of lawn plant and their landscape effect. It is an important indicator to assess the aesthetic quality of the vegetation, and reflect the drivers' feeling of the landscape in great degree. The colour of the plant can be divided into 5 levels according to observation of the investigator: (1) blackish green; (2) dark green; (3) green; (4) light green; (5) yellow green, and the corresponding value ranges are V_2= {9, 7~8, 5~6, 3~4, 0~2}.

Degree of uniformity (C_{22}): the level of uniformity in species distribution of vegetation and coordination degree of colour. The index can be divided into 3 levels: (1) good: the lawn is completely flat, highly consistent, no weeds or fewer weeds; (2) bad: there is large difference between the height of the lawn and full of weeds; (3) general: between the two situations above, and corresponding value ranges are V3= {7~9, 0~4, 5~6}.

3) Stability index (B_3)

Root of the plant plays a significant role in anchorage function of vegetation because of the interface friction between the soil and roots and thus forms a composite reinforced material. The root system of the roadside vegetation can improve the strength parameters of the soil and therefore improve the stability of the slope. With the absorption of moisture and transpiration, the pore pressure of slope decreases and the stability of slope improves. According to experts from related field and previous researches, the stability indices can be selected as follows:

Stretching resistance (C_{31}): tension that can break the roots or pull them out of the soil. The root will crack and even be pulled out under certain tension, and then the reinforcing function of the root will cripple, therefore the stability of the slope decreases.

Safety factor (C_{32}): The comprehensive index that reflects the probability of shear failure of the slope. The stability of the roadside slope is directly decided by the shear strength of soil-mass, and the presence of root system will improve the shear strength parameters significantly. The stability of highway slope is also decided by the slope angle and the bulk density of soil-mass, so the safety factor index can be set up to reflect the stability of slope.

B. Engineering Case

Before the quality evaluation, a decision group of 20 experts from related field was set up to provide the criterions which determine the standards of sub-factors and relative importance of different sub-factors and factors to obtain pair-wise comparison matrixes.

The parameters of sub-factors were acquired from the research of Jiang [16]. In Jiang's research, the indices of six different common herbs for slope vegetation (①. Trifolium pretense Linn, ②. Agrostis stolonifera Linn, ③. Trifolium repens L, ④. Bromus inermis Leyss, ⑤. Lolium perenne L., ⑥. Elymus tranchycaulum) were measured, and the data from previous research is listed in Table 1. In addition, the corresponding criterions of standards determined by experts are listed in Table 2.

C. Calculation of weight parameters by AHP

The Pair-wise comparison matrixes and weights of indices of different levels are determined by AHP as following (Table 3~5).

D. Calculation of evaluation vector

Based on the weight of indices in Table 3~5, the evaluation vectors of the factors and sub-

factors can be calculated by Eq. 4. Then the overall evaluation vectors can be calculated by Eq. 5. The result is presented in Table 6.

TABLE 1. Value of indices for highway quality evaluation

Vegetation \ Indices	Growth rate(cm/d)	Coverage (%)	Disease resistance	Drought resistance	Arid resistance	Planting density(/cm^2)
Trifolium pretense Linn	0.14	86	7	7	7	1.7
Agrostis stolonifera Linn	0.25	76	6	6	6	1.4
Trifolium repens L	0.14	87	7	7	8	1.8
Bromus inermis Leyss	0.32	78	8	8	8	1.5
Lolium perenne L.	0.30	88	9	7	8	2.5
Elymus tranchycaulum	0.19	73	9	9	9	1.5

Vegetation \ Indices	Summer survival rate (%)	Colour of vegetation	Degree of uniformity	Stretching resistance(N)	Safety factor of shear failure
Trifolium pretense Linn	78	6	8	22.45	1.31
Agrostis stolonifera Linn	66	5	6	16.64	1.40
Trifolium repens L	80	6	8	22.46	1.34
Bromus inermis Leyss	82	8	4	28.13	1.64
Lolium perenne L.	78	9	9	25.68	1.53
Elymus tranchycaulum	83	7	5	26.40	1.14

TABLE 2. Effect evaluation indices and standard classification of slope vegetation

Level	Growth rate(cm/d)	Coverage (%)	Disease resistance	Drought resistance	Arid resistance	Planting density(/cm^2)
Ⅰ	>0.3	>90	9	9	9	>2.2
Ⅱ	0.24~0.30	80~90	7~8	7~8	7~8	1.8~2.2
Ⅲ	0.18~0.24	70~80	5~6	5~6	5~6	1.4~1.8
Ⅳ	0.12~0.18	60~70	3~4	3~4	3~4	1.0~1.4
Ⅴ	<0.12	<70	0~2	0~2	0~2	<1.0

Level	Summer survival rate (%)	Colour of vegetation	Degree of uniformity	Stretching resistance (*N*)	Safety factor of shear failure
Ⅰ	>80	9	9	>30	>1.6
Ⅱ	70~80	7~8	7~8	25~30	1.4~1.6
Ⅲ	60~70	5~6	5~6	20~25	1.2~1.4
Ⅳ	50~60	3~4	3~4	15~20	1.0~1.2
Ⅴ	<50	0~2	0~2	<15	<1.0

3.1. The Satisfaction Function of Indices

The satisfaction function $\boldsymbol{f}$ of ecology, landscape and stability indices should be determined before the calculation of fuzzy integral. The range of function $\boldsymbol{f}$ is [0, 1], and the value of it increases with the increase of variable value $\boldsymbol{x}$ as the evaluation of highway slope vegetation increases with the increasing value of factors and sub-factors. According to these properties, the satisfaction can be defined as follows:

$$f(x_i) = \frac{x_i}{x_{\max}}, \tag{11}$$

where $x_{\max}$ is 1 for the indices valued with the percentage or 9 for that with grade in standard classification in Table 2. For example, the satisfaction function value of coverage (C_{12}) is itself, and the function value of Disease resistance (C_{13}) can be obtained when the grade of Disease resistance is divided by 9. The result is presented in Table 7.

TABLE 3. Pair-wise comparison matrix of the factors and corresponding weights

A	B_1	B_2	B_3	W_A
B_1	1	4	2	0.56
B_2	1/4	1	1/3	0.12
B_3	1/2	3	1	0.32
λ_{max}=3.018 CR=0.017<0.1				

TABLE 4. Pair-wise comparison matrix of the sub-factors within 'ecology functions' (B_1) and corresponding weights

B_1	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	W_{B1i}
C_{11}	1	1/4	1/3	1/3	1/3	1	1/3	0.05
C_{12}	4	1	3	3	3	4	2	0.33
C_{13}	3	1/3	1	1	1	2	1/2	0.12
C_{14}	3	1/3	1	1	1	2	1/2	0.12
C_{15}	3	1/3	1	1	1	2	1/2	0.12
C_{16}	1	1/4	1/2	1/2	1/2	1	1/3	0.06
C_{17}	3	1/2	2	2	2	3	1	0.20
λ_{max}=7.107 CR=0.013<0.1								

TABLE 5. Pair-wise comparison matrix of the sub-factors within "landscape function" (B_2) and "structure function" (B_3) and corresponding weights

B_2	C_{21}	C_{22}	W_{B2}	B_3	C_{31}	C_{32}	W_{B3}
C_{21}	1	1/2	0.33	C_{31}	1	1/3	0.25
C_{22}	2	1	0.67	C_{32}	3	1	0.75
λ_{max}=2 CR=0 <0.1							

TABLE 6. Overall evaluation vectors of six herbs

Result	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
EV_1	0.000	0.579	0.393	0.028	0.000
EV_2	0.000	0.268	0.652	0.080	0.000
EV_3	0.000	0.579	0.393	0.028	0.000
EV_4	0.380	0.322	0.218	0.080	0.000
EV_4	0.221	0.779	0.000	0.000	0.000
EV_6	0.314	0.120	0.326	0.240	0.000

TABLE 7. The satisfaction function value of the indices

Indices / Vegetation	Growth rate	Coverage	Disease resistance	Drought resistance	Arid resistance	Planting density
Trifolium pretense Linn	0.467	0.860	0.778	0.778	0.778	0.773
Agrostis stolonifera Linn	0.833	0.760	0.667	0.667	0.667	0.636
Trifolium repens L	0.467	0.870	0.778	0.778	0.889	0.818
Bromus inermis Leyss	1.000	0.780	0.889	0.889	0.889	0.682
Lolium perenne L.	1.000	0.880	1.000	0.778	0.889	1.000
Elymus tranchycaulum	0.633	0.730	1.000	1.000	1.000	0.682

Indices / Vegetation	Summer survival rate	Color of vegetation	Degree of uniformity	Stretching resistance	Safety factor of shear failure
Trifolium pretense Linn	0.780	0.667	0.889	0.748	0.819
Agrostis stolonifera Linn	0.660	0.556	0.667	0.555	0.875
Trifolium repens L	0.800	0.667	0.889	0.749	0.838
Bromus inermis Leyss	0.820	0.889	0.444	0.938	1.000
Lolium perenne L.	0.780	1.000	1.000	0.856	0.956
Elymus tranchycaulum	0.830	0.778	0.556	0.880	0.713

3.2. The Fuzzy Density of Indices and Coefficient λ

In the evaluation system of slope vegetation protection, the effects of factors and sub-factors are not independent actually. For instance, A represents the event that the ecology function is fine when growth rate (C_{11}) is at the first level, B represents the event that the ecology function is fine when coverage (C_{12}) is at the first level, then C represents the event that the ecology function is fine when growth rate (C_{11}) and coverage (C_{12}) are both at the first level. The probabilities of three events often satisfy the relationship as follows:

$$P(C) > P(A) + P(B) \tag{12}$$

Because the ecology function is more likely to be fine with two indices of first level, the comprehensive effects of two indices may surpass the simple sum of them. It can be depicted by the superadditivity of indices well when λ-fuzzy measure is applied to measure the probability of corresponding functions in good condition.

The fuzzy density g_i of indices should be obtained before the calculation of coefficient λ, the fuzzy

density of factors and sub-factors can be given by following method:

1. 20 experts are asked to value the importance of factors and sub-factors. If one expert considers the sub-factor essential for the evaluation of factors or factors essential for the overall evaluation of slope vegetation, he will tick the box below it. There will be 20 boxes for every factor or sub-factor.

2. The fuzzy density g_i can be calculated by Eq.13

$$g_i = \frac{m}{20} w_i , \tag{13}$$

where m is the number of boxes with tick below the sub-factors or factors and w_i is the weight of sub-factors or factors calculated in the first model by AHP. For example, the evaluation of sub-factors in landscape function by experts is presented in Table 8.

TABLE 8. Assessment on sub-factors of landscape function by experts

Expert index	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
C_{21}	√	√		√	√		√	√	√	√
C_{22}	√	√	√	√	√	√	√	√	√	√
Expert index	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
C_{21}	√	√	√		√	√	√	√	√	
C_{22}	√	√		√	√	√	√	√	√	√

the fuzzy density can be calculated by the method above:

$$g_\lambda(\{C_{21}\}) = \frac{16}{20} \times 0.33 = 0.264 ;\ g_\lambda(\{C_{21}\}) = \frac{19}{20} \times 0.67 = 0.637$$

After calculation of fuzzy density g_i, the coefficient λ can be solved by Eq.6. As for the sub-factors calculated above, λ is the positive solution of equation as follows: $(1+0.264\lambda)(1+0.637\lambda) = 1+\lambda$, and the result should be $\lambda = 0.592$, the results are presented in Table 9.

TABLE 9. Fuzzy density and coefficient λ of different decision layers

Decision layer	**fuzzy density g_i**	λ
A	$g_\lambda(\{B_1\}) = 0.560 , g_\lambda(\{B_2\}) = 0.096$ ， $g_\lambda(\{B_3\}) = 0.304$	0.156
B_1	$g_\lambda(\{C_{11}\}) = 0.040 , g_\lambda(\{C_{12}\}) = 0.330 , g_\lambda(\{C_{13}\}) = 0.114 , g_\lambda(\{C_{14}\}) = 0.114$ $g_\lambda(\{C_{15}\}) = 0.114 , g_\lambda(\{C_{16}\}) = 0.045 , g_\lambda(\{C_{17}\}) = 0.200$	0.116
B_2	$g_\lambda(\{C_{11}\}) = 0.267 , g_\lambda(\{C_{12}\}) = 0.637$	0592
B_3	$g_\lambda(\{C_{11}\}) = 0.188 , g_\lambda(\{C_{12}\}) = 0.713$	0.749

Table 10. The evaluation by fuzzy integral of six herbs

Vegetation / Fuzzy integral	**1**	**2**	**3**	**4**	**5**	**6**
Sugeno integral	0.748	0.667	0.749	0.780	0.780	0.713
Choquet integral	0.756	0.680	0.756	0.791	0.825	0.705

3.3. The Evaluation of Slope Vegetation Based on Fuzzy Integral

The overall evaluation of slope vegetation from different common herbs can be provided by comparison of the value of fuzzy integral. The satisfaction function and fuzzy measure can be calculated from the methods above. Then the index i of element x_i is permuted, then the new set $\overline{X} = \{x_{(1)}, x_{(2)}, \cdots, x_{(n)}\}$ is constructed, where $0 \le f(x_{(1)}) \le f(x_{(2)}) \le \cdots \le f(x_{(n)}) \le 1$. Sugeno fuzzy integral and Choquet fuzzy integral of factors can calculated from Eq.8 and Eq.10 based on the satisfaction function value of sub-factors respectively, then fuzzy integral of three factors are permuted again as the satisfaction function value, the fuzzy integral reflecting overall performance of slope vegetation is obtained finally, which is presented in Table 10.

3.4. The Range of Levels for the Evaluation by Fuzzy Integral

The criteria for ranking levels of the slope vegetation are determined by the fuzzy integral of *supremum* and *infinum* of different levels. As for the standards of level Ⅰ (excellent), the satisfaction function value of *supremum* and *infinum* can be divided into 3 types:1. the *supremum* is 1 and the infimum is minimum percentage of level Ⅰ (excellent) for indices expressed as a percentage; 2.the *supremum* is 1 and the *infinum* is 8/9 for indices expressed by grade from 1 to 9; 3.the *supremum* and the *infinum* are both 1 when there is only *infinum* of indices like stretching resistance. The satisfaction function value of *supremum* and *infinum* for the other levels can be calculated by Eq.11 where x_i is the *maximum* and *minimum* value of indices respectively. The criteria of fuzzy integral for different levels can be obtained by the integral calculation of the satisfaction function value for corresponding levels, which is presented in Table 11.

TABLE 11. The range of five levels for the evaluation by fuzzy integral

Level / Fuzzy integral	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
Sugeno integral	(0.800,1.000]	(0.667,0.800]	(0.600,0.667]	(0.500,0.600]	(0.000,0.500]
Choquet integral	(0.870,1.000]	(0.717,0.870]	(0.620,0.717]	(0.508,0.620]	(0.000,0.508]

4. Discussion

An attempt has been made in this paper to provide evaluation for the slope vegetation of highway by AHP and fuzzy math methods, and two evaluation systems are presented respectively: 1. Evaluation system based on AHP and fuzzy comprehensive evaluation, 2. Evaluation system based on fuzzy measure and fuzzy integral.

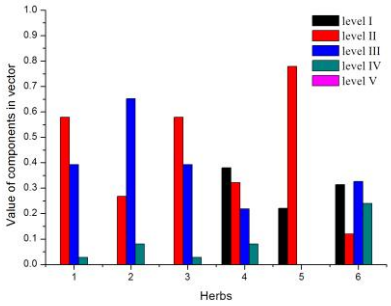


Figure 1. fuzzy comprehensive evaluation for herbs

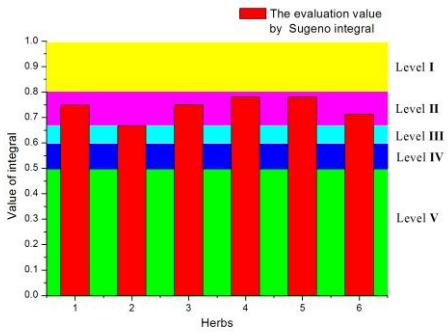


Figure 2. Evaluation based on Sugeno fuzzy integral for herbs

The evaluation of the slope vegetation is a multi-criteria decision making problem, while factors of ecology, landscape and stability should be considered comprehensively. In the first evaluation model, the weight of factors and sub-factors is determined by AHP method, which implies the comprehensive effect of different factors and sub-factors can be calculated by traditional weighted arithmetic mean [17]. The evaluation vectors obtained by fuzzy comprehensive evaluation are expected to determine the performance of different herbs, and the integrative level of slope vegetation is decided by the maximum component in the evaluation vectors. In this model, the interaction between different indices cannot be achieved because of the inherent faults in the weighted arithmetic mean method [12].

According to the evaluating vectors in Table 6, the comprehensive performance of different herbs by first evaluation system can be analysed. The result of *Trifolium pretense Linn* and *Trifolium repens L* is between "good" and "general", and closer to "good", so they belong to level Ⅱ (good); The quality levels of the other herbs can be determined in same way, and the result is : *Agrostis stolonifera* (Ⅲ), *Bromus inermis Leyss* (Ⅰ), *Lolium perenne L* (Ⅱ) and *Elymus tranchycaulum* (Ⅲ). Confusion arises when we review the evaluation of *Bromus inermis Leyss* whose evaluation vector is (0.380, 0.322, 0.218, 0.080, 0.000). As the magnitude of components reflects the probability of belonging to corresponding level, it is arbitrary to conclude that the performance of *Bromus inermis Leyss* can be relegated to level Ⅰ (excellent) because the second component of the vector indicates that it has a considerable probability to

be relegated to level Ⅱ (good). In fact, it is hard to argue the validity of judgment based on the evaluation vector when there is no significant difference between magnitudes of components in the vector. The evaluation vector of *Elymus tranchycaulum* is (0.314, 0.120, 0.326, 0.240, 0.000), which indicates that the magnitude of probabilities for level Ⅰ and that for level Ⅲ is close. It is unreasonable that the probability for level Ⅱ is smaller than probabilities for level Ⅰ and level Ⅲ at the same time, (Figure 1 and Figure 2).

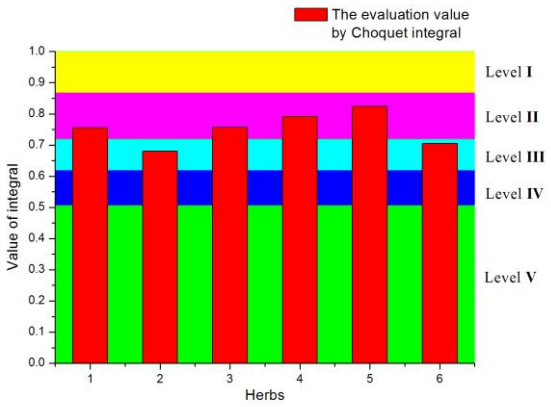


Figure 3. Evaluation based on Choquet fuzzy integral for herbs

In the evaluation, system based fuzzy measure and fuzzy integral, the coordination of factors can be considered [12]. The comprehensive effect of several factors or sub-factors is not the simply calculated by weighted arithmetic mean method, and superadditivity of indices can be achieved by the introduction of -fuzzy measure. For the evaluation based on fuzzy measure and fuzzy integral, the confusion of ranking does not exist. The Sugeno fuzzy integral and Choquet fuzzy integral are calculated respectively to evaluate the performance of slope vegetation and the range of five levels can be obtained by the fuzzy integral of corresponding supremum and infinum. According to result of Sugeno integral, the performance of level for six herbs is Trifolium pretense Linn (Ⅱ), Agrostis stolonifera Linn (Ⅱ), Trifolium repens L (Ⅲ), Bromus inermis Leyss (Ⅱ), Lolium perenne L. (Ⅱ), Elymus tranchycaulum (Ⅱ). For the result of Choquet integral, the evaluation of six herbs is Trifolium pretense Linn (Ⅱ), Agrostis stolonifera Linn (Ⅲ), Trifolium repens L (Ⅱ), Bromus inermis Leyss (Ⅱ), Lolium perenne L. (Ⅱ), Elymus tranchycaulum (Ⅲ). As for the result of Sugeno integral (Figure 3), the evaluation values of Bromus inermis Leyss and Lolium perenne L. are the same and their relative performance is beyond comparison, which is because that there are heavy information loss in the operation of and for the calculation of Seguno integral [18]. Considerable deviations occur in the evaluation result based on Seguno integral for the same reason. The ranking of different herbs and their relative performance can be obtained on the basis of the evaluation provided by Choquet integral, and the result is basically consistent with the previous research [16].

5. Conclusions

Good quality of slope vegetation is an important guarantee for the operation of highway and the environment of surrounding areas. In this paper, three evaluation models are constructed respectively to assess the quality of slope vegetation.

In the first model, the weights of main factors and sub-factors are calculated by Analytic Hierarchy Process (AHP), and the fuzzy comprehensive evaluation is employed to determine the performance rank; in second model, the λ-fuzzy measure and Sugeno fuzzy integral are introduced for the quality evaluation; in third model, Sugeno fuzzy integral in the second model is replaced by Choquet fuzzy integral. The research data of six herbs is used to examine the validity of these models and the reasons for difference of third models are investigated. The analysis of the result shows that the third model can obtain reasonable evaluation result and avoid the confusion of ranking in the first model and overcome heavy information loss in the second model at the same time.

It can be well-recognized that in the financial models of human behaviour outlined above the company acts as an active player. Thus, the active system under consideration is a single-level one.

Acknowledgments

"Project 40962005 supported by National Natural Science Foundation of China." And this work was supported by Opening Fund of State Key Laboratory of Geohazard Prevention and Geoenvironment Protection (Chengdu University of Technology) (Grant No. GZ2007-09). * Jiu-Ming Li, corresponding author, Email: jxjaljm156@163.com. *Hai-Feng Dai, corresponding author, Email:daihaif@163.com.

References

1. Taylor, B.R., Roff, J.C. (1986) Long-term effects of highway construction on the ecology of a Southern Ontario stream. *Environmental Pollution Series A, Ecological and Biological* **40**(4), 317-344.
2. Reijnen, R., Foppen, R., Veenbaas, G. (1997) Disturbance by traffic of breeding birds: evaluation of the effect and considerations in planning and managing road corridors. *Biodiversity and Conservation* **6**, 567-581.
3. Ullmann, I. (1998) Lateral differentiation and the role of exotic species in roadside vegetation in southern New Zealand. *Flora,* **193**, 149–164.
4. Forman, R.T.T., Reineking, B., Hersperger, A.M. (2002) Road traffic and nearby grassland bird patterns in a suburbanizing landscape. *Environmental Management,* **29**, 782-800.
5. Palmer, J.F., Hoffman, R.E. (2001) Rating reliability and representation validity in scenic landscape assessments. *Landscape Urban Plan,* **54**, 149-161.
6. Viles, R.L., Rosier, D.J. (2001) How to use roads in the creation of greenways: case studies in three New Zealand landscapes. *Landscape Urban Plan,* **55**, 15–27.
7. Clay, G.R., Smidt, R.K. (2004) Assessing the validity and reliability of descriptor variables used in scenic highway analysis. *Landscape Urban Plan,* **66**(4), 239–255.
8. Xie, M. (1990) A study on soil mechanical role of tree roots in the stability of slopes. *Bulletin of Soil and Water Conservation,* **4**(3) 7-15. (*in Chinese*)
9. Osman N., Barakbah S.S. (2006) Parameters to predict slope stability—Soil water and root profiles. *Ecological Engineering,* **28**(1), 90-95. (*in Chinese*)
10. Zheng, G., Zhu, N., Tian, Zh., et.al. (2012) Application of a trapezoidal fuzzy AHP method for work safety evaluation and early warning rating of hot and humid environments. *Safety Science,* **50**, 228-239. (*in Chinese*)
11. Cengiz, K., Ufuk, C., Ziya, U. (2003) Multi-criteria supplier selection using fuzzy AHP. *Logistics Information Management* **16**(6), 382-394.
12. Grabisch, M. (1996) The application of fuzzy integrals in multi-criteria decision making. *European Journal of Operational Research* **89**, 445-456.
13. Chen, S.M., Tan, J.M. (1994) Handling multicriteria fuzzy decision-making problems based on vague set theory. *Fuzzy Sets and Systems* **67**, 163–172.
14. Murofushi, T., Sugeno, M. (1992) Non-additivity of fuzzy measures representing preferential dependence. *The 2nd International Conference on Fuzzy Systems and Neural Networks*, Iizuka, Japan, 617-620.
15. Viles, R.L., Rosier, D.J. (2001) How to use roads in the creation of greenways: case studies in three New Zealand landscapes. *Landscape Urban Plan* **55**, 15–27
16. Haohao, J. (2009) Effect Evaluation of Slope Protection for Herbs Based on the Grey Connection and Principal Component Analysis. *The Master Degree Thesis of Northeast Forestry University*, 54–61. (*in Chinese*)
17. Barzilai, J., Golany, B (1994) AHP rank resersal, normalization and aggregation rules. *INFOR* **32**(2), 57-64
18. Zhang, Ya., Yuan, H., Pan, J., Li, Y., Xi, R., Yao, L. (2006) A Fuzzy Integral Method of Applying Support Vector Machine for Multi-class Problem. *ICNC 2006, Part II, LNCS 4222*, 839 –846

Received on the 21st of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 102-111
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

CALCULATION OF SHAFT TEMPERATURE FIELD IN DEEP-WATER TESTING SYSTEM

J. Liu[1,2*], X. He[1], X. Wei[3,*], C. Huang[1]

[1]School of Civil Engineering and Architecture
Southwest Petroleum University, Chengdu 610500, China
[2]Mechanical Engineering Post-Doctoral Research Station
Chengdu 610500, China
[3]College of Mechatronic Engineering
Southwest Petroleum University, Chengdu 610500, China
[*]Corresponding author. Phone: 0086- 028-83037631. E-mail: liujun-888888@163.com

The determination of shaft temperature field in the deep-water testing system is important to the strength analysis on pipe string. Based on the three conservation functions and gas state equation, a mathematical model for the shaft temperature field of deep-water testing is proposed. In the model, testing string system is divided into two parts, formation section and seawater section. Based on the analysis on the heat transfer paths of the two sections in radius direction, the heat transfer coefficients for the model are determined. The correctness of the model is proved by means of comparing the calculation result with that tested in practice production well.

Keywords: deep-water testing system, temperature field, mathematical model, heat transfer coefficients

1. Introduction

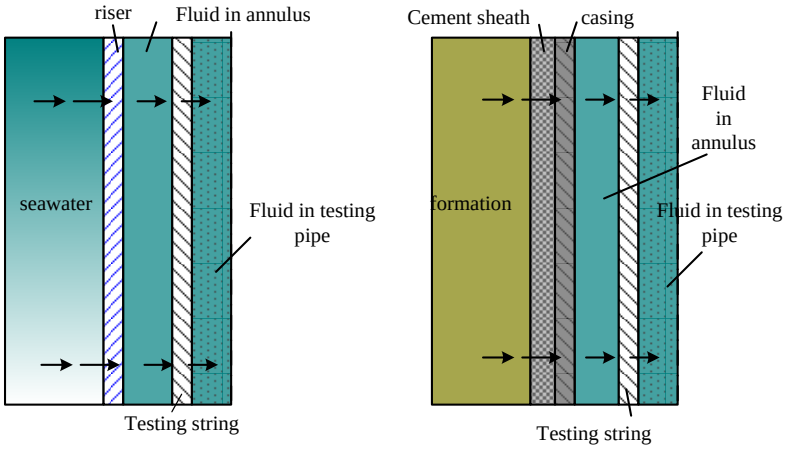
In deep-water gas well, the shaft temperature field has great influence on the stress and strain of pipe string, dynamic analysis on product system and optimization design [1-4]. The effective prediction on the shaft temperature of gas well can not only prevent the brittle fracture and fatigue failure of pipe string, but also optimize the gas well structure and improve production efficiency [5].

In 1964, Ramey [6] proposed an approximate calculation model for the temperature field of the single-phase ideal gas flowing in shaft. Subsequently, the relaxation distance, specific heat at constant pressure and fluid temperature on the well bottom were treated as constant by Shiu&Beggs [7], in order to simplify the calculation formula. In 1991, transient heat transfer formula for formation was given by Hasan&Kabir [8]. In 1999, the temperature test to the donghai sea continent shelf was carried out by institute of oceanology, Chinese academy of sciences, and a fitting equation for sea water temperature in four seasons was obtained [9], based on which, Gao & Sun [10] analysis temperature field distribution of drilling fluid in the annulus and drill string in the process of deep-water drilling. Guo [11] analysis the temperature and pressure of a gas well, taking into account the gas state equation. But there are no comprehensive models for the temperature field prediction of deep-water gas well due to its complication.

This paper outlines an attempt to a mathematical model for the shaft temperature field of deep-water testing, based on the three conservation functions and gas state equation. In the model, testing string system is divided into two parts, formation section and seawater section. The heat transfer coefficients are determined according to the analysis on heat transfer paths of the two sections in radius direction. The correctness of the model is proved by means of comparing the calculation result with that tested from practice production well.

2. Theoretical Model

In the determination of temperature field, the deep-water test system can be divided into two parts: deep-water section and formation section, shown as figure 1, since the working environments of the two sections are very different. The paths of heat transfer in the two sections are respectively simplified as figure 2, 3.



(a) deep-water section (b) formation section

Figure 1. Physical model of wellbore in deep-water test.

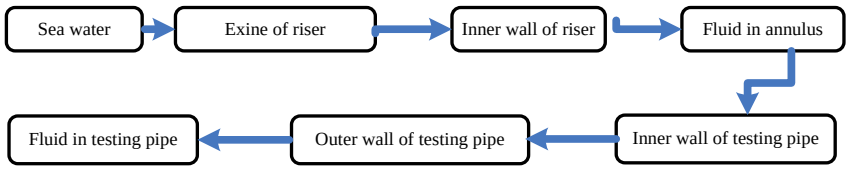


Figure 2. Schematic diagram of energy transfer along the radial direction of seawater section

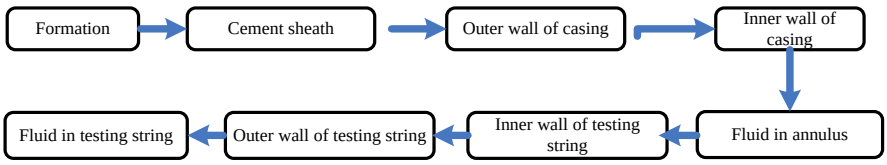


Figure 3. Schematic diagram of energy transfer along the radial direction of formation section

For convenience of derivation, the following hypothesises are made: 1) The gas state in the testing string is stable and uniflow. 2) The heat transfer in the deep-water shaft is stable. 3) The heat transfer in the formation is unstable and obeys no-dimension time function proposed by Remay. 4) The axises of testing pipe and casing (or riser) are coincidence.

2.1. Mathematical Model for Formation Section

Well head on the midline is set as the origin of coordinate z with the forward direction along testing string and downward. Micro unit of formation section and its local coordinate system are shown as figure 4.

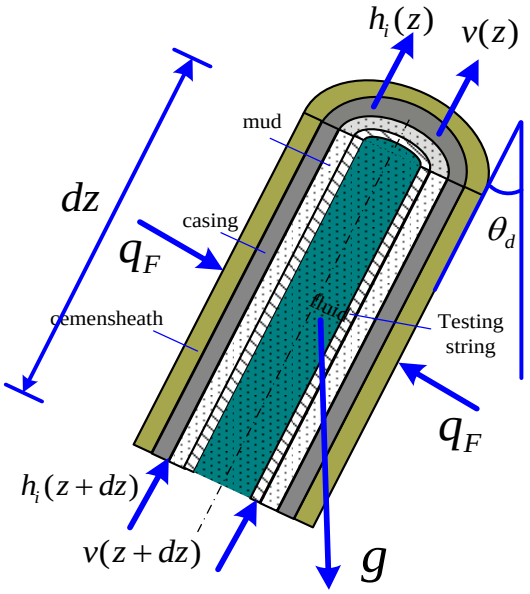


Figure 4. Micro unit of formation section in heat transfer analysis

The mass, moment and energy equations for the gas flowing in a pipe may be written as formula (1) ~ (3)

$$-\frac{dp}{dz}+\left(\frac{dp}{dz}\right)_{fr}+\rho g\cos\theta=\rho\frac{vdv}{dz}, \tag{1}$$

$$h_i(\mathrm{z}+\mathrm{dz})-\mathrm{h}_i(\mathrm{z})+\frac{1}{2}w_i v^2(\mathrm{z}+\mathrm{dz})-\frac{1}{2}w_i v^2(\mathrm{z})-\mathrm{w}_i\, g\cos\theta_d dz+q_F=0, \tag{2}$$

$$\rho\frac{dv}{dz}+v\frac{d\rho}{dz}=0, \tag{3}$$

where

$h_i(z+dz)$ =the enthalpy flowing into the micro unit in the unit time, including internal energy and pressure energy (J/s),

$h_i(z)$ = the enthalpy flowing out of the micro unit in the unit time, including internal energy and pressure energy (J/s),

w_i =fluid quantity in the unit time (kg/s),

q_F =heat transferred from formation into shaft in the unit time (J/t),

θ_d = deviation angle (°).

The moment conservation equation (1) can be written as [12]

$$\frac{dp}{dz}=\rho g\cos\theta+\left(\frac{dp}{dz}\right)_{fr}-\frac{\rho vdv}{dz}, \tag{4}$$

$$\left(\frac{dp}{dz}\right)_{fr}=\frac{f\rho v^2}{D},$$

$$\frac{1}{\sqrt{f}}=-2\log\left[\frac{\varepsilon/D}{3.715}+\left(\frac{6.943}{\mathrm{Re}}\right)^{0.9}\right],$$

where $\left(\frac{dp}{dz}\right)_{fr}$ is friction pressure gradient, f is friction resistance coefficient, D is inner diameter of testing pipe, Re is Reynolds number, ε is pipe absolute roughness.

Substituting the gas equation of state $\rho = \frac{Mp}{RT}$ into equation (4), yields,

$$\frac{dv}{dz} = -\frac{vMdp}{\rho RTdz}, \tag{5}$$

where R is gas constant, M is molar mass, v is the velocity of the fluid in shaft (m/s), T is the temperature of the fluid in shaft, ρ is gas density.

The energy conservation equation (2) can be transformed into the following form,

$$\frac{dh_i}{dz} + w_i \frac{vd_v}{dz} - w_i g\cos\theta_d + \frac{dq_F}{dz} = 0. \tag{6}$$

Considering the following relations

$$q_F = \frac{2\pi r_{co} k_e U_a (T_{ei} - T_i)\,\mathrm{dz}}{(k_e + \mathrm{r}_{co} U_a f_t)},$$

where T_{wb}, T_i and T_{ei} respectively denote the temperatures of shaft, formation and the interface between them (℃), r_{co} is the radius of casing, k_e is thermal conductivity ($W/m\cdot℃$), U_a is heat transfer coefficient, f_t is the time function of transient heat transfer.

The energy conservation equation (6) can be written as the following form

$$\frac{dh_a}{dz} + w_i \frac{vdv}{dz} - w_i g\cos\theta_d + \frac{2\pi r_{co} k_e U_a (T_{ei} - T_i)}{(k_e + r_{co} U_a f_t)} = 0. \tag{7}$$

For $10^{-10} \le t_D \le 1.5$, $f_t = 1.1281\sqrt{t_D}(1 - 0.3\sqrt{t_D})$,

For $t_D > 1.5$, $f_t = (0.4036 + 0.5\ln t_D)(1 + \frac{0.6}{t_D})$,

where $t_D = \frac{\alpha t}{r_{wb}}$, $\alpha = \frac{k_e}{c_e \rho_e}$, α is thermal conductivity of formation (m2/s), r_{wb} is radius of wellbore (mm), ρ_e is formation density (kg/m3), c_e is formation specific heat ($J/kg\cdot℃$).

In the energy conservation equation (7) for the fluid in shaft, the specific enthalpy of gas is the function of temperature and pressure, shown as following formula

$$dh = c_p w_a dT - c_p w_a \alpha_J dp, \tag{8}$$

where c_p is specific heat at constant pressure, α_J is Joule-Thomson coefficient of gas.

Formula (8) may be simplified as follow, since the gas flowing in the testing string and the radius of the pipe change little along pipe length

$$dh = c_p w_a dT. \tag{9}$$

Substituting formula (9) into formula (7), yields

$$c_p w_a \frac{dT}{dz} + w_a \frac{vdv}{dz} - w_a g\cos\theta_d + \frac{2\pi r_{co} k_e U_a (T_{ei} - T)}{(k_e + r_{co} U_a f_t)} = 0. \tag{10}$$

So that the temperature of shaft can be expressed as

$$\frac{dT}{dz} = \frac{g\cos\theta_d - \frac{vdv}{dz} + \frac{2\pi r_{co} k_e U_a (T - T_{ei})}{w_a (k_e + r_{co} U_a f_t)}}{c_p}. \tag{11}$$

Simultaneously solving the following equations, the temperature field of formation section can be obtained

$$\begin{cases} \dfrac{dv}{dz} = -\dfrac{vMdp}{\rho RTdz} \\ \dfrac{d\rho}{dz} = \dfrac{T\rho g\cos\theta_d + T(\dfrac{dp}{dz})_{fr} - \dfrac{\rho g\cos\theta_d}{c_p} - \dfrac{p}{c_p}(\dfrac{2\pi r_{co} k_e U_a (T - T_{ei})}{w_a (k_e + r_{co} U_a f_t)})}{\dfrac{RT^2}{M} - Tv^2 + \dfrac{Pv^2}{c_p \rho}} \\ \dfrac{dp}{dz} = \rho g\cos\theta_d + \left(\dfrac{dp}{dz}\right)_{fr} - \dfrac{\rho vdv}{dz} \\ \dfrac{dT}{dz} = \dfrac{g\cos\theta_d - \dfrac{vdv}{dz} + \dfrac{2\pi r_{co} k_e U_a (T - T_{ei})}{w_i (k_e + r_{co} U_a f_t)}}{c_p} \end{cases} \tag{12}$$

Boundary condition:

$$\begin{cases} v(\mathrm{z}_b) = \dfrac{w_i}{A\rho} \\ T(\mathrm{z}_b) = T_b \end{cases}, \tag{13}$$

where z_b is coordinate of well bottom, T_b is temperature of well bottom.

2.2. Mathematical Model for Deep-water Section

The derive process for the mathematical model of deep-water section is the same as above [13]. Simultaneously solving the following equations, the temperature field of deep-water section can be obtained

$$\begin{cases} \dfrac{dv}{dz} = -\dfrac{vMdp}{\rho RTdz} \\ \dfrac{d\rho}{dz} = \dfrac{T\rho g\cos\theta_d + T(\dfrac{dp}{dz})_{fr} - \dfrac{\rho g\cos\theta_d}{c_p} - \dfrac{p}{c_p}(\dfrac{2\pi r_{co}k_e U_a(T-T_{ei})}{w_a(k_e + r_{co}U_a f_t)})}{\dfrac{RT^2}{M} - Tv^2 + \dfrac{Pv^2}{c_p\rho}} \\ \dfrac{dp}{dz} = \rho g\cos\theta_d + \left(\dfrac{dp}{dz}\right)_{fr} - \dfrac{\rho v dv}{dz} \\ \dfrac{dT}{dz} = \dfrac{g\cos\theta_d - \dfrac{vdv}{dz} + \dfrac{2\pi r_{hw}k_{se}U_t(T-T_{ei})}{w_i(k_{se} + r_{hw}U_t f_t)}}{c_p} \end{cases}, \tag{14}$$

where U_t is heat transfer coefficient of deep-water section, k_{se} is heat transfer coefficient of sea water, r_{hw} is the radius of borehole, namely the inside diameter of riser.

Boundary condition:

$$\begin{cases} v(\mathrm{z}_{ml}) = \dfrac{w_i}{A\rho} \\ T(\mathrm{z}_{ml}) = T_{ml} \end{cases}, \tag{15}$$

where z_{ml} is coordinate of mudline, T_{ml} is temperature of the mudline.

3. Heat Transfer Coefficient in Deep-water Testing System

In the formula (12), (14), *Ua*, *Ut* are overall heat transfer coefficients of seawater section and formation section. For the former, the heat transfer coefficient is relative to the resistance of the fluid in shaft, casing material and cement sheath, whereas for the latter, the heat transfer coefficient is relative to the resistance of the fluid in shaft, riser material. The two heat transfer coefficients can be determined using the fluid energy conservation law from formation section (or seawater) to shaft.

3.1. Heat Transfer Coefficient for Formation Section

In the deep-water testing, the temperatures of product casing, cement sheath and other casings out of cement sheath are assumed identical to that of formation. Therefore, the heat transfer model for formation section can be simplified as figure 1(b). According to energy conservation law, the overall heat transfer coefficient from cement sheath to the inner wall of testing string can be written as

$$\frac{1}{U} = \frac{1}{h_{ce}} + \frac{d_h In(d_{tso}/d_{tsi})}{2\lambda_{ts}} + \frac{d_h In(d_{cro}/d_{cni})}{2\lambda_{can}} + \frac{d_h In(d_{cao}/d_{cai})}{2\lambda_{cao}} + \frac{d_h In(d_h/d_{cai})}{2\lambda_{cem}} \tag{16}$$

where d_h is the diameter of well bore (mm), h_{ce} is the thermal conductivity coefficient of convective heat transfer of the fluid in the testing pipe, d_{tsi} and d_{tso} are the inner and outer diameters of testing pipe, d_{cro} and d_{cni} are the outer and inner diameters of the annulus mud, and d_{cao} are the inner and outer diameters of casing (mm), λ_{ts}, λ_{cao}, λ_{can}, λ_{cem} are respectively the heat conduction coefficients of testing pipe, casing, mud and cement sheath.

In order to obtain overall heat transfer coefficient, the convective heat transfer coefficient should be determined firstly using the following techniques.

1) If Re<2100, the process of convective heat transfer is laminar heat transfer, so that

$$h_1 = \frac{4.12k_m}{D}, \tag{17}$$

where h_1 is the convective heat transfer coefficient for the heat transfer between the shaft and the fluid in it, k_m is the thermal conductivity coefficient of the fluid in shaft, D is the inner diameter of shaft.

2) If Re>2100, the process of convective heat transfer is turbulent heat transfer, so that

$$\begin{cases} h_1 = \dfrac{0.027(\mathrm{Re})^{0.8}(\mathrm{Pr})^{0.333}k_m}{D} \\ \mathrm{Pr} = \dfrac{c_p\mu_d}{k_m} \end{cases}. \tag{18}$$

3.2. Heat Transfer Coefficient for Seawater Section

The heat conduction process along the radial direction of seawater section can be represented as figure 2. According to energy conservation law, the energy equation for seawater section can be written as

$$Q_{apw} = \frac{\left[T_{st}(z,t) - T_w(z,t)\right]p\Delta z\Delta t}{\dfrac{1}{h_{ce}d_{hw}} + \dfrac{In(d_{tso}/d_{tsi})}{2l_{ts}} + \dfrac{In(d_{cro}/d_{cni})}{2l_{can}} + \dfrac{In(d_{rio}/d_{rii})}{2l_{rin}}}. \tag{19}$$

The overall heat transfer coefficient from riser to the inner wall of testing pipe is assumed as U_t, the inner wall of the testing pipe is set as reference surface, so that overall heat transferred into the fluid in testing pipe Q_{ap} can be expressed as

$$Q_{apw} = U_t\pi d_{hw}\left[T_{st}(z,t) - T_w(z,t)\right]\Delta z\Delta t. \tag{20}$$

The overall heat transfer coefficient of seawater can be written as

$$\frac{1}{U_t} = \frac{1}{h_{ce}} + \frac{d_{hw}In(d_{tso}/d_{tsi})}{2\lambda_{ts}} + \frac{d_{hw}In(d_{cro}/d_{cni})}{2\lambda_{can}} + \frac{d_{hw}In(d_{rio}/d_{rii})}{2\lambda_{rin}}, \tag{21}$$

where h_{ce} is the thermal conductivity coefficient of convective heat transfer of the fluid in the testing pipe, d_{tei} and d_{tso} are the inner and outer diameters of testing pipe, d_{cro} and d_{cni} are the outer and inner diameters of the annulus mud, d_{hw} is the diameter of the well bore in the seawater section, d_{rio} and d_{rii} are the outer and inner diameters of the riser, λ_{ts}, λ_{rin}, λ_{can} are respectively the heat conduction coefficients of testing pipe, riser and annulus mud.

4. Solution for Temperature Field

In this section, numerical method, fourth-order Runge-Kutta method is used to solve the temperature field, since the mathematical models established in the section 2 is too complicate to be solved by analytical method. For the calculation of the temperature field of formation shaft, since the fluid velocity and temperature is the function of deep, the right parts of the equations (12) can be written

as F_1 and F_2, So that the equations (12) can be expressed as

$$\begin{cases} \dfrac{dv}{dz} = F_1(\mathrm{z},\mathrm{T}) \\ \dfrac{dT}{dz} = F_2(\mathrm{z},\mathrm{T}) \end{cases} \tag{22}$$

If the function value $y_i(z_0)$ at the start position z_0 and step length h are known, the solution of node $z = z_0 + h$ can be written as follow using fourth-order Runge-Kutta method.

$$y_i^0 = y_i^0 + \frac{h}{6}(a_i + 2b_i + 2c_i + d_i), \tag{23}$$

where

$$a_i = F_i(z_0, y_1^0, y_2^0),$$
$$b_i = F_i(z_0 + \frac{h}{2}, y_1^0 + \frac{h}{2}a_1, y_2^0 + \frac{h}{2}a_2),$$
$$c_i = F_i(z_0 + \frac{h}{2}, y_1^0 + \frac{h}{2}b_1, y_2^0 + \frac{h}{2}b_2),$$
$$d_i = F_i(z_0 + h, y_1^0 + hc_1, y_2^0 + hc_2).$$

If z does not reach the destine depth, the computational value of current node is set as the value of the start position in the next step. Repeating the above process, the temperature values on the discretize node are obtained.

5. Example

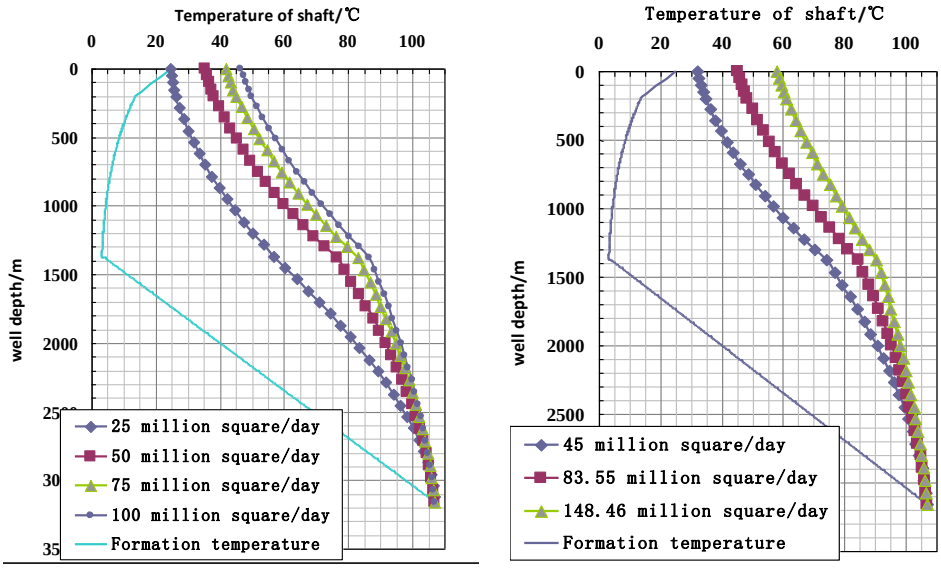
Temperature field of a practical deep-water gas well # is investigated using the method presented in this paper. The basic parameters are listed in table 1.

TABLE1. Basic parameters of deep-water gas well #.

Parameter name	value	Parameter name	value
Inner diameter of testing pipe(m)	0.095	Depth of sea water(m)	1378.6
Outer diameter of testing pipe(m)	0.1143	Specific heat of sea water(J/g·℃)	4.182
Heat conduction coefficient of formation(W/m·℃)	2.06	Inner diameter of casing(m)	0.2168
Specific heat of formation rock(J/kg·℃)	0.837	Outer diameter of casing(m)	0.24448
Density of formation rock(kg/m^3)	2640	Heat conduction coefficient of casing(W/m·℃)	43.26
Temperature of formation(℃)	107	Inner diameter of cement sheath(m)	0.24448
Depth of formation(m)	3160	Outer diameter of cement sheath (m)	0.31115
Temperature Gradient of formation(℃/100m)	6	Heat conduction coefficient of cement sheath(W/m·℃)	0.52
Specific heat of natural gas(J/kg·℃)	2227	Inner diameter of cement sheath (m)	0.4698
Density of natural gas（g/cm^3）	0.65	Outer diameter of cement sheath (m)	0.5334
Heat conduction coefficient of natural gas(W/m·℃)	0.03	Heat conduction coefficient of riser(W/m·℃)	43.26
Apparent viscosity of natural gas(Pa*s)	2.555e-5	Molar mass of natural gas(g/mol)	16
Density of sea water(kg/m^3)	1023	Gas constant, J/（mol·k）	8.314
Heat conduction coefficient of sea water(W/m·℃)	0.57	Pressure gradient of gas (pa/m)	22638

The calculation of temperature field under four kinds of gas yields is carried out and the results are

compared with those tested in practice, see figure 5. It can be found in the figure that temperature of formation shaft decrease with the reduction of well depth. For the temperature-well depth curve, there is an evident inflection point on the mudline well head. The temperature variation trend from well head on the mudline to deep-water region is closed to that in the formation section, except for that temperature gradient of the sea water is greater than that of the formation section. In addition, the calculated temperature result shown in figure 5(b) approaches the tested value in practice. The comparison of the two kinds of results is listed in table 2. It can be found that the error between the calculated value and the tested value is small, the model proposed in this paper is effective for the temperature calculation of deep-water testing shaft.



(a) Temperature calculation of shaft under different yields

(b) Temperature calculation of shaft under practical yields

Figure 5. Shaft temperature distribution of deep-water gas well # under different yields

TABLE 2. Temperature comparison between the practice test and calculation of well #

Gas yield ($10^4m^3/d$)	Temperature of well head on seawater section		Temperature of the well head on mudline	
	test(℃)	calculation(℃)	test(℃)	calculation(℃)
45.00	34	32.099	74.3	74.2
83.55	47	45.08	84.9	84.4
148.46	55	57.817	87.9	90.5

6. Conclusion

According to the characteristics of deep-water testing technology, the temperature fields of seawater section and formation section have been respectively established. In the models, the heat transfer characteristics of sea water, riser, fluid in annulus, casing, cement sheath, testing pipe and the fluid in it have been fully taken into account. Using the model, the temperature field of an actual well has been investigated. For the two sections of the testing string, the temperature-well depth curve decrease with the reduction of well depth, there is an evident inflection point on the mudline well head, the temperature gradient of the two sections are different. The effective of the model is proved because the error between the calculated result and the value tested in practice is little.

Acknowledgements

The authors gratefully acknowledge the financial support of National Natural Science Foundation of China (51274171) and National Significant Science and Technology Special Sub-project （2011ZX05026-001-07）.

References

1. Song, X.C., Guan, Z.C. (2011) Full transient analysis of heat transfer during drilling fluid circulation in deep-water wells. *Acta Petrolei Sinica,* **32**(4), 704-708 (*In Chinese*)
2. Ma, C.H., Wu, X.D. (2006) Theoretical research for improving temperature distribution of fluid in wellbore using beat pipe. *Acta Petrolei Sinica,* **27**(1),114-118 (*In Chinese*)
3. Kabir, C.S., Hasan, A.R. (2006) Simplified wellbore flow modeling in gas condensate system. *SPE Production & Operation,* **21**(1), 89-97
4. Wu, H., Wu, X.D. (2011) A wellbore flow model of CO_2 separate injection with concentric dual tubes and its affecting factors. *Acta Petrolei Sinica,* **32**(4), 722-727 (*In Chinese*).
5. Li, Q.D., Wang, T.C. (2012) Risk control of deep-water well blowout considering wellbore temperature field and pressure field. *SPE,* **12**(25), 6310-6313 (*In Chinese*).
6. Ramey, H.J.Jr. (1995) Wellbore heat flow of steam down a wellbore. *JPY,* **17**(7), 845-851
7. Shiu, K.C., Beggs, H. D. (1980) Predicting temperatures in flowing oil well. *Resource Tech. Trans. ASMF*, **102**(1), 2-11.
8. Hasan, A.R., Kabir, H.D. (1991) Predicting temperatures in flowing oil wellbor; Part Ⅱ-wellbore Fluid Temperature. *SPE22948*, 695-708
9. Zheng, W.P., Zhou, D. (2003) GIS aided estimation of resource amount of natural gas hydrate in southern South China Sea. *Journal of Tropical Oceanography,* 22(6), 35-45 (*In Chinese*)
10. Gao, Y.H., Sun, B.J. (2008) Calculation and analysis on the shaft temperature field of deepwater drilling. *Journal of China University of Petroleum,* **32**(2), 47-51 (*In Chinese*)
11. Guo, C.Q., Li, Y.C. (2001) Numerical simulation prediction on the pressure and temperature of gas well. *Acta Petrolei Sinica*, 22(3),100-105 (*In Chinese*)
12. Liang, Y.T., Hao, F.F., Liu, J., Wang, Z.J. (2011) Explicit formula determination on friction factor for southwest products pipeline. *Oil & Gas Storage and Transportation*, 30(4), 316-319 (*In Chinese*)
13. Liu, G.X., Ling, W.X. (2008) Calculation of wellbore temperature in offshore drilling. *Offshore Oil,* **28**(4), 83-86 (*In Chinese*).

Received on the 21st of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 112-120
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

RESEARCH ON SPHERICAL VOXEL-BASED MACHINING SIMULATION

Ch.-Ju. Chen[1,2], Ju. Hong[2], L. Yan[3]

[1]School of Mechanical Engineering
Qingdao Technological University, Qingdao 266033, China
[2]School of Mechanical Engineering
Xi'an Jiaotong University, Xi'an 710049, China
[3]Shandong Institute of Business and Technology
Yantai 26400, China
Phone: (+86)-15269218816.
E-mail: Chenchengjun79@163.com

To improve simulation speed and accuracy of virtual machining simulation system, a spherical voxel-based 3D machining simulation method was proposed in this paper. In the initialization stage, the spherical voxel-based octree model of virtual blank is constructed and the adjacent map of any voxel is modelled. In the stage of virtual machining simulation, the collision detection between virtual blank and virtual tool will be conducted from the voxels machined in last collision detection loop to their adjacent voxels calculated using the adjacent map. If some test voxel of virtual blank overlays with virtual tool, the overlap of this voxel will be removed and the parameters such as the coordinates of centre, radius and machining points are revised. By defining machining point in spherical voxel of virtual blank, the surface of virtual blank can be rendered smoothly and efficiently using triangular faces. The presented method was realized and an augmented reality-based milling simulation system was conducted, which showed that the presented algorithm was feasible and effective.

Keywords: virtual machining; spherical voxel model; adjacent map; machining point

1. Introduction

Virtual machining, which uses computer-generated 3D virtual scene to simulate physical machining process, can be used to test and improve machining process before machining physical blank. It enables users to model, program and test machining process in an intuitive virtual environment. So the duration and cost of both design and manufacturing can be greatly reduced. Since virtual machining can verify and check correctness of machining processes and collision between tool and blank, it has been widely used in production and education.

Simulation speed and accuracy are key indicators for virtual machining algorithms. Virtual blank modelling is not only a prerequisite for machining simulation, but also the fundamental reason for speed and accuracy of simulation. Traditional techniques of virtual blank modelling mainly include B-rep, CSG, Z-buffer, dexel and so on. B-rep method [1] can include geometric topology information in detail, but it cannot provide original information of entity. CSG method was proposed and adopted by Voelcker and Hunt [2]. Its model can be expressed simply, but in virtual machining simulation, it involves a large number of Boolean operations, so the simulation speed is greatly deteriorated, especially for complicated models. In Z-buffer methods [3], the Boolean operations is conducted by comparing the z values of virtual blank with those of tool instead of 3D collision detection. Therefore, they can reduce 3D Boolean operation to one-dimensional comparison, and the simulation efficiency can be greatly enhanced. But it is not suitable for more than three-axis virtual machining. Dexel model based methods [4] split virtual blank and virtual tool into a collection of dexels, then use collision detection among dexels to simulate material removal. It has very fast simulation speed and fairly simulation accuracy, but it applies only to three-axis milling simulation too. Voxel-based methods [5-6] use a collection of voxels to approximate the representation of the virtual blank and virtual tool. During machining simulation, they use collision detection between voxels of virtual blank and virtual tool to simulated material removal. The main disadvantage of Voxel-based methods is that they occupy more memory and involve a large number of Boolean operations. The virtual machining accuracy and speed of voxel-based methods mainly depend on the voxel size of virtual blank. The smaller the voxel size, the lower the virtual machining speed but the more precise the machining. Triangular model-based methods [7] use a collection of triangular faces to represent virtual blank surface. Then the virtual machining is realized by intersection testing between triangular faces of virtual blank and those of virtual tool. So methods belonging to this kind are simple,

efficient, and easy to implement, but they are hard to meet the simulation speed, simulation accuracy and other requirements for virtual machining with more than three degrees

To improve simulation speed and accuracy of virtual machining, this research combines octree model with spherical voxel model to represent virtual blank and tool. Octree is used to subdivide virtual blank and tool, and spherical voxel is used to represent subdivision elements of virtual blank and virtual tool. By encoding voxels reasonably, the adjacent map of virtual blank is modelled, and the six adjacent voxels of any voxel on can be quickly accessed. In stage of virtual machining simulation, the collision detection of virtual blank will be conducted from the voxels machined in last collision detection loop, and then spread to their adjacent voxels calculated using the adjacent map. So the times of collision detection between voxels are reduced and virtual machining efficiency can be enhanced. In addition, according to collision states between voxels of virtual blank and those of virtual tool, the centres and radiuses of collided voxels in virtual blank are changed dynamically instead of removing the whole voxel, so the simulation accuracy is supposed to be improved.

The reminder of this paper is organized as followings. The section 2 gives virtual blank modelling using spherical voxel-based octree model, and the adjacent map is proposed. Section 3 presents virtual machining simulation method using spherical voxel models. Section 4 gives a simple example to verify the presented methods. The last section gives a conclusion of this paper.

2. Virtual Blank Modelling Using Spherical Voxel –Based Octree Model

2.1. Spherical Voxel Octree Model

As shown in Fig. 1, most existing voxel-based virtual machining algorithms divide the minimum bounding cube which surrounded the virtual blank recursively into eight smaller sub-cubes, until the minimum size of sub-cube or division depth is met, and sub-cubes at the bottom of octree model is named as voxel. All cubes and voxels are saved in octree data structure [8]. The cubic voxel can enclose virtual blank tightly. However, these methods do not distinguish empty cubes from solid cubes. In addition, it needs large memory space and the collision detection between cubes is inefficient. To improve the virtual machining speed and efficiency, this paper uses spherical voxels instead of cubic voxel to model virtual blank and virtual tool. As shown in Fig. 2, only two parameters (namely centre and radius) are needed to represent a spherical voxels, so the memory space can be significantly reduced and collision detection between spheres are more efficient. In additional, after triangulation of adjacent spheres on surface of virtual blank, triangles faces can be used for display surface of virtual blank. So the can graphics rendering load of virtual machining scene can be greatly decreased.

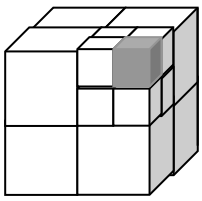


Figure 1. Cubic voxel based octree model

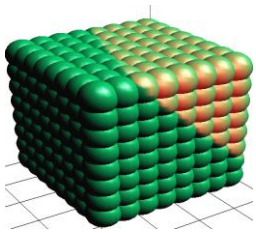


Figure 2. Spherical voxel based octree model

After modelling virtual blank using spherical voxel-based octree model, each voxel needs to be encoded. As shown in Fig. 3, this paper encodes the voxel in sequential order along Y and X axis from negative to positive direction then along *Z* axis from positive to negative direction. Assume that the depth or level of octree model is n (n>0), then the octree model has $\sum_{l=0}^{l=n} 8^l$ nodes or voxels, and the indices of these nodes are 0,1,2……$((8^n-1)/7)-1$.

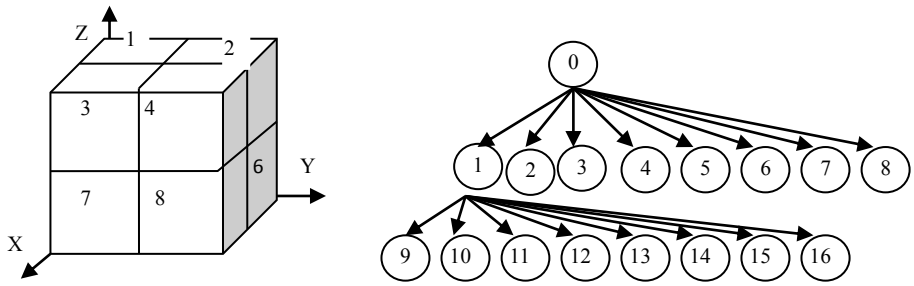


Figure 3. Octree encoding rules

Then given a node with an index i (i>0), the depth and other information of this node can be accessed using follow expressions.

The index of its upper parent node:(i-1)/8；

The indices of its eight children node :$8\times i+t$, (t = 1,2,……8)；

The octree level of this node: $\log_8{}^{(7i+1)}$；

2.2. Adjacent Map Model

Using spherical voxel-based octree models of virtual blank and virtual tool, virtual machining can be regarded as a process that repeatedly conducts collision detection between voxels of virtual blank and those of virtual tool [9]. In this process, the voxels of virtual blank, which collided with those of virtual tool will be removed or revised from virtual blank, so they will not be displayed in virtual scene. Due to time-space consistency of relative motion of tool with respect to blank, collision between virtual blank and virtual tool will most likely to occur around the collision regions detected in last collision detection loop. Based on this character, the collision detection in virtual machining can be directly conducted around the collision regions of last loop. By this way, it is hoped to enhance the virtual machining speed and efficiency. Therefore, this paper tries to build the adjacent map of spherical voxel-based octree model, which can provide the six–adjacency of any voxel of virtual blank.

According to the above-mentioned encoding rules, the graph for XOY plane of octree can be drawn in Fig.4. In this graph, the depth of octree is three, that is to say the virtual blank is split from level 0 to level 3. The circles denote the nodes or voxels of octree, and the numbers in circles denote indices of the nodes or voxels. By analysing and summing up, the incremental index of adjacent node along the X, Y, Z axis can be solved using the following iterative calculation. In the followings formula, l is the level of nodes, and is great than 1. $\Delta h(i)$, $\Delta v(i)$, $\Delta w(i)$ represent incremental index of adjacent voxels for level i along X, Y and Z axis respectively.

$$\text{X axis:}\quad \begin{cases} \Delta h(1)=[1] \\ \Delta h(l)=[\Delta h(l-1),\frac{6}{7}8^{l-1}+\frac{1}{7},\Delta h(l-1)] \end{cases}$$

$$\text{Y axis:}\quad \begin{cases} \Delta v(1)=2\Delta h(1)=[2] \\ \Delta v(l)=2\Delta h(l) \end{cases}$$

$$\text{Z axis:}\quad \begin{cases} \Delta w(1)=4\Delta h(1)=[4] \\ \Delta w(l)=4\Delta h(l) \end{cases}$$

For example, in Fig.4： $\Delta h(1)$=[1], $\Delta v(1)$=[2], $\Delta w(1)$=[4]; $\Delta h(2)$=[1,7,1],$\Delta v(2)$=[2,14,2], $\Delta w(2)$=[4,28,4]; $\Delta h(3)$=[1,7,1,55,1,7,1], $\Delta v(3)$=[2,14,2,110,2,14,2], $\Delta w(3)$=[4,28,4,220,4,28,4]。

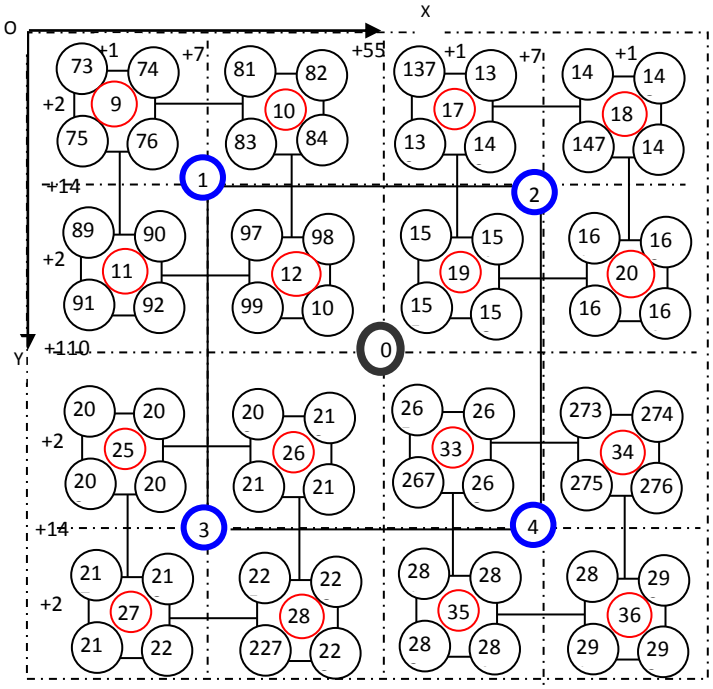


Figure 4. XOY plane graph of octree

From abovementioned iterative equations, we can see that with increment of l, the adjacent relationship between voxels becomes increasingly complicated. In addition, the iterative equations are difficult to be programmed. Therefore, this paper proposes a parameterized adjacency model, which can access any voxel's six adjacent voxels along X, Y and Z by looking up tables using its index as parameter.

As shown in Fig. 5, each voxel has six adjacent voxels along X, Y, Z axis. In programming, the initial numbers of the six adjacent voxels for any voxel is set to be -1, which means there is no adjacent voxel in corresponding direction. If there is no adjacent voxel in certain direction or the adjacent voxel has been removed, the corresponding initial value is still -1. Otherwise, the number of adjacent voxels along the X, Z, Y axis can be obtained by the rules listed in Table 1, Table 2 and Table 3. Suppose the index of the voxel to be searched is m, and the index of its parent voxel is k, the parameters a, b, and c in tables can be calculated using the following equations.

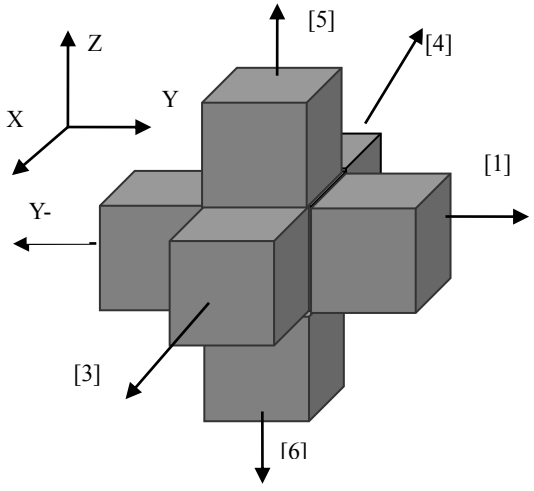


Figure 5. Six–adjacency voxel

$$a = \frac{6}{7}8^{PP} + \frac{1}{7}, b = 2(\frac{6}{7}8^{QQ} + \frac{1}{7}), c = 4(\frac{6}{7}8^{ZZ} + \frac{1}{7}) \tag{1}$$

In which, all value of *PP*, *QQ*, *ZZ* can be calculated as follows.

(1) Calculating *PP*: The initial value of *PP* is set to be one. If m is odd (even), then find all its parent nodes from the bottom level to the top level. Once the index of its one parent node is odd (even), the value of *PP* will be added with 1. For example, the voxel index 74 is even, the indices of its parents are nine and one respectively, but they are both odd, so the corresponding value of *PP* is still one.

(2) Calculating *QQ*: The initial value of *QQ* is set to be one. This method firstly finds all its parent nodes in different levels. Supposing that the remainder of m divided by 8 is n, once the remainder of its parent index divided by 8 is also n, the value of *QQ* will be added with 1. For example, all the parent nodes of voxel 91 are 11 and 1 respectively. Both the remainders of 91 and 11 divided by 8 are three, so the value of *QQ* is two.

(3) Calculating *ZZ*: The initial value of *ZZ* is also set to be1. Find all its parent nodes in different levels. Supposing current voxel lies at the upper layer (bottom layer) of its upper parent voxel, once one of its parent node is also lies at the upper layer (bottom layer), the value of *ZZ* will be added with 1. For example, the voxel index is 96, all its parent nodes are 11 and 1, 96 lies at the lower layer, but 11 and 1 are at the upper layer in their corresponding level, so the value of *ZZ* is till 1.

Table 1: Adjacent voxels along X axis

conditions		**results**	
m	***k***	***X+*:**	***X-*:**
odd	odd	*m +1*	*m-a*
odd	even	*m+1*	*m-a*
even	odd	*m+a*	*m-1*
even	even	*m+a*	*m-1*

Table 2: Adjacent voxels along Z axis

conditions		**Adjacent voxels along**	
		Z+:	**Z-:**
remainder of *m* divided by 8 is:	1, 2, 3 or 4 (top level)	*m-c*	*m+4*
	5, 6, 7 or 0 (bottom level)	*m-4*	*m+c*

Table 3: Adjacent voxels along Y axis

conditions	**Adjacent voxels along**	
	Y+:	**Y-:**
The remainder of *m* divided by 8 is 1, 2, 5 or 6 The remainder of *k* divided by 8 is 1, 2, 5 or 6	m+2	m-b
The remainder of *m* divided by 8 is 1, 2, 5 or 6 The remainder of *k* divided by 8 is 3, 4, 7 or 0	m+2	m-b
The remainder of *m* divided by 8 is 3, 4, 7 or 0 The remainder of *k* divided by 8 is 1, 2, 5, 6	m+b	m-2
The remainder of *m* divided by 8 is 3, 4, 7 or 0 The remainder of *k* divided by 8 is 3, 4, 7 or 0	m+b	m-2

For example, If we need to find the six adjacent voxels of the voxel indexed with 82, according to the adjacent map model proposed in this paper, *PP* = 2, *QQ* = 2, *ZZ* = 2. So the adjacent voxels along Y+, Y-, X+, X-, Z+, and Z- are 137, 81, 84, -1, -1 and 86 respectively. The adjacent voxels calculated using the proposed model was proved to be right using different test data.

To speed of calculating six adjacent voxels of any voxel, the formulation (1) can be saved as table, so the parameters a, b, and c can be calculated by looking up tables. So the speed of virtual machining can be enhanced.

3. Virtual Machining Simulation

3.1. Collision Detection

The key to the virtual machining algorithm is to test the collisions between voxels of virtual blank and those of virtual tool. As spherical voxel-based octree model is used to model both virtual blank and virtual tool in this paper, the virtual machining process is a recursive process that conducts voxels collision detection from top level to bottom level of virtual blank octree with virtual tool. This research used time-space consistency of the relative motion of tool with respect to blank. Using the adjacent map

of virtual blank, the collision detection in virtual machining simulation is directly conducted around the collision regions detected in last collision detection loop. The algorithm is depicted in Fig. 6.

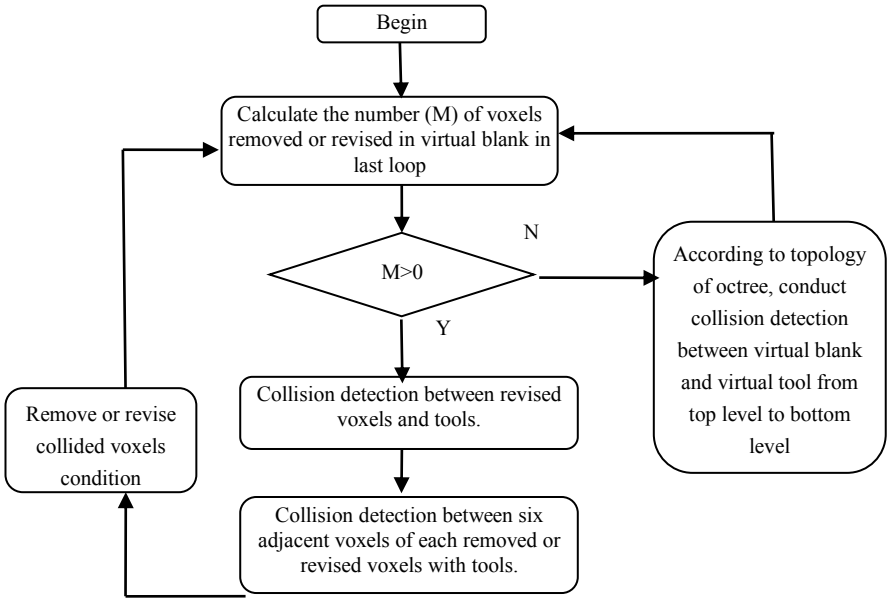


Figure 6. Diagram of collision detection using adjacent map

As shown in Fig.7, during the collision detection of virtual machining simulation, first the number (*M*) of voxels removed totally or partly in last loop is calculated. When *N* is great than 0, the collision detection will be conducted around the collision region of virtual blank by testing whether removed voxels and their six adjacent voxels collide with virtual tool until there is no new adjacent voxel overlapping with virtual tool. The collided voxels in current frame is store in special arrays, which will be access in the next collision detection loop. If *N* is equal to zero, the collision detection will be conducted as usual from the top level to the bottom level of octrees of virtual blank and virtual tool.

3.2. Material Removal and Triangulation Patch

Most existing voxel-based virtual machining methods simulate material removal process by simply removing collided voxels totally. The main disadvantage of these methods is that the simulation accuracy is affected by the voxel size. The smaller the voxels, the higher simulation accuracy these methods can obtain. Nevertheless, with the voxel size smaller, more and more memory spaces are needed to store large amount of voxels. This paper presents a method to revise the parameters (such as radius and centre coordinates) of spherical voxels dynamically according to the collision between spherical voxels of virtual blank and virtual tool. Therefore, the overlap of voxels can be partly removed, and the simulation accuracy can be enhanced in condition of fixed voxel size.

As shown in Fig.7, Suppose that the voxel radiuses of virtual blank and virtual tool are *R1* and *R2* respectively, $P(x_p,y_p,z_p)$ is the centre of ball-end mill, $Q(x_q,y_q,z_q)$ is the centre of voxel of virtual blank,. The collision detection problem is transformed into central distance comparison, and the material removal can be simulated by removing or revising spherical voxels of virtual blank.

(1) If $|PQ|$ is greater than $R1+R2$ and the voxel of virtual blank lies under the virtual tool, there is no collision, so no voxel should be removed from virtual blank.
(2) If $|PQ|$ is less than the absolute value of $R1-R$, in other words, the whole voxel of virtual blank lies within virtual tool, then the voxel of virtual blank is cut off. So the voxel is totally removed from virtual blank;
(3) As shown in Fig.7 If $|PQ|$ is less than $R1+R2$ and greater than the absolute value of $R1-R2$, and the voxel of virtual blank lies under the virtual tool, the blank voxel should be cut partly. In this case, the voxel of virtual blank can be simplified as a residual spherical voxel with a smaller radius, which just contacts with surface of tool. The centre $W(x_w,y_w,z_w)$ and radius r of the residual spherical voxel can be calculated using diagram depicted in Fig. 7.

To enhance the 3D display effectiveness, the 3D virtual machining scene is expected to be displayed using triangle instead of spherical voxel. How to get the vertexes of triangles becomes a key problem. In this paper, a concept of machining point is proposed. Machining point in this paper refers to the last contact point of virtual blank voxel with virtual tool in single pass. For example, point A is the machining point of the residual voxel in this machining pass. By connecting adjacent contact points of all virtual blank voxels, the surface of virtual blank can be displayed using triangle or other polygon.

A voxel may have more than one contact point. In addition, the coordinates of contact point of voxel is changed with the dynamic machining simulation. As shown in Fig. 8, supposing point B is the contact point of the first machining pass, and point B is the contact point of the second machining pass, when the angle α between OB and OC is larger than a threshold value, both the point B and A are the contact points of the voxel, otherwise contact point B will be replaced by A, In other words, the voxel will only one contact point. In this way, the display precision can be controlled especially for complicated surface machining.

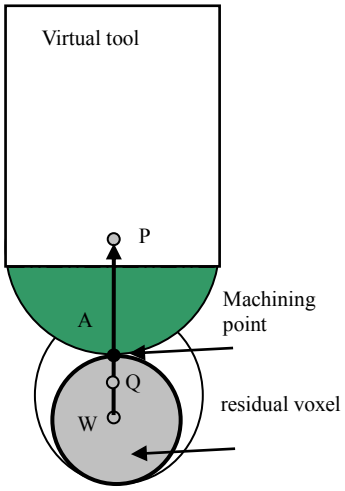


Figure 7. Collision detection between virtual blank voxels and virtual tool

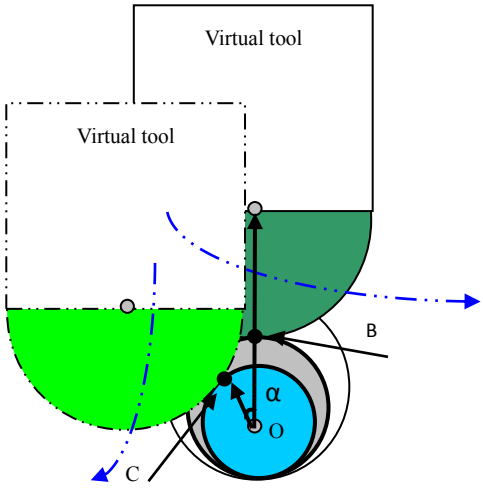


Figure 8. Change of contact point in machining simulation

Above all, according to time-space consistency of the relative motion of tool with respect to virtual blank, this paper uses the proposed adjacent map to speed the virtual machining simulation. And by dynamically revising parameters of collided voxels and the contact points, the simulation accuracy can also be enhanced. The virtual machining simulation speed and accuracy is little affected by voxel size.

4. Experiments

To verify the methods proposed in this paper, we implemented related algorithms to simulate ball-end milling process on platform of Intel Pentium Dual-Core E5800@3.20GHz processor, 2.0GB memory, ATI Radeon HD 5500 Series (1 GB) card and VC++ development environment. As shown in Fig.8, the virtual blank is divided into the 7 levels according using the spherical voxel-based Octree model presented in this paper. In addition, simulation speed of proposed method using adjacent map model was compared with method unused adjacent map. The experimental snapshot is shown in Fig.9. From the figure, we can see that simulation precision is acceptable.

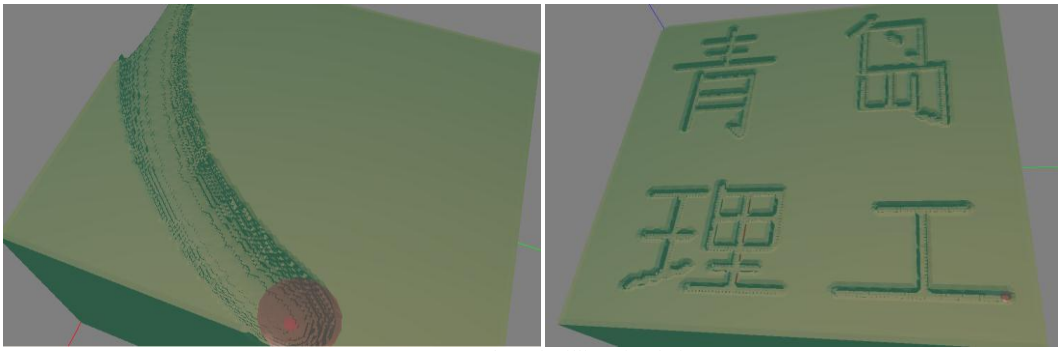
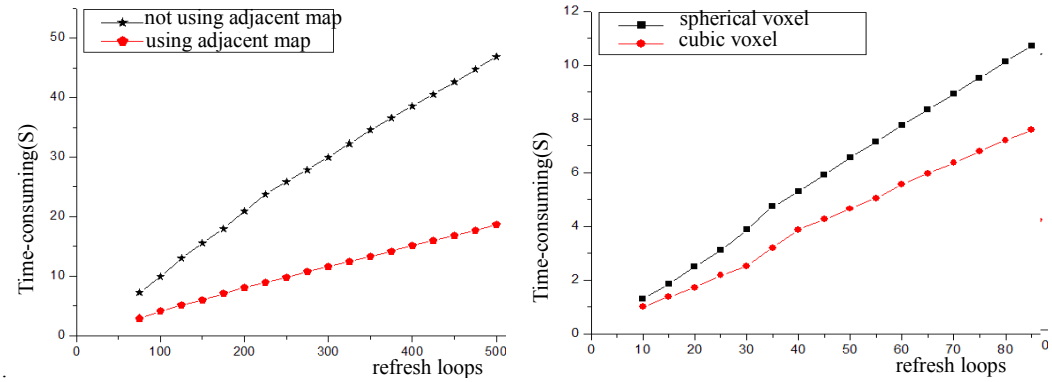


Figure 9. Snapshot of milling simulation

As shown in Fig. 10, the influences of voxel adjacent map and voxel type on machining simulation speed are also compared. From the Fig. 10(a), we can see that after using adjacent map the simulation speed is significant improved. This is because the test of collision detection is greatly reduced by looking up voxel adjacent map. From the Fig. 10(b) we can see that the spherical voxel-based method proposed in this paper has advantages in simulation speed comparing with cubic voxel method.



(a) Influence of adjacent map (b) Influence of voxel type

Figure 10. Comparison of simulation speed

The proposed spherical voxel-based Machining simulation method was applied in an anti-collision machining simulation system of CNC (Computer numerical control) machine. In this system, the virtual prototype of physical CNC machine is built. When the system is running, the position of physical machine table is collected in real-time from a PLC, which connected with encoders amounted on lead screw of physical machine table. The virtual machining simulation is driven by the real-time positions of physical machine table and the position predicted based on CNC program. Moreover, the function of anti-collision is realized by the collision detection between virtual blank and virtual prototype of physical CNC machine. The snapshot of the anti-collision is depicted in Fig. 11. The yellow part means the collision between virtual blank with the physical machining environments.

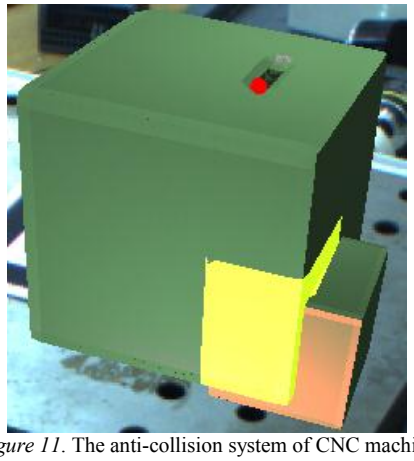


Figure 11. The anti-collision system of CNC machine

5. Conclusions

To improve simulation speed and accuracy of existing virtual machining methods, we proposed a new partial rendering method for spherical voxel-based machining simulation. The advantages of presented method include: ① The adjacent map was model according to encoding rules of spherical voxels-based octree. In addition, collision detection of virtual machining simulation is conducted from voxels that collided with virtual tool in last frame to their adjacent voxels calculated using adjacent map, so the speed of virtual machine simulation can be greatly improved. ② By dynamically revising parameters such as center, radius and contact points of collided spherical voxels of virtual blank, the simulation accuracy was enhanced. It should be pointed out that the simulation accuracy and speed is little affected by voxel size of virtual blank.

Acknowledgements

This research work is co-supported by the promotive research fund for excellent young and middle-aged scientists of Shandong Province (Grant NO. BS2012ZZ004), the natural science foundation of Shandong Province (Grant NO. ZR2012FL09), and NSFC project (Grant NO.51105215).

References

1. Fleisig, R.V., Spence, A.D. (2005) Techniques for accelerating B-rep based parallel machining simulation Computer-Aided Design, 37(21), 1229-1240.
2. Voelcker, H., Hunt, W. (1981) The Role of Solid Modelling in Machining-Process Modelling and NC Verification," SAE Technical Paper 810195, doi:10.4271/810195.
3. Wang, D., Li, T., Chen, E.Y. (2010) Machining of Composite Surface Based on Z-buffer Method with Torus Tools [J]. Journal of Mechanical Engineering, 46(7), 193-198.
4. Van Hook, T. (1986) Real Time Shaded NC-Milling Display. ACM SIGGRAPH'86 Computer Graphics Proceedings, 20(4), 15-20.
5. Jang, D, Kim, K, Jung, J. (2000) Voxel-Based Virtual Multi-Axis Machining. International Journal of Advanced Manufacturing Technology, 16(10), 709-713.
6. Wu, X.J., Liu, W.J., Wang, T.R. (2005) Octree-based three-dimensional mesh voxelization method. Engineering Graphics, (4), 1-7.
7. Liu, S.Q., Ong, S.K., Chen, Y.P., Nee, A.Y.C. (2006) Real-time, dynamic level-of-detail Management for three-axis NC milling simulation. Computer-Aided Design, 38(4), 378-391.
8. Roy, U., Xu, Y.X. (1998) 3-D object decomposition with extended octree model and its application in geometric simulation of NC machining. Robotics and Computer-Integrated Manufacturing, 14(4), 317-327.
9. Zhou, J.W., Wan, Y., Wan, W.G. (2009) Octree based on bounding boxes OBB collision detection method. Computer Applications and Software, 4(26), 75-77.

Received on the 21st of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 121-129
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

MECHANISM RESEARCH OF VISCOELASTIC MAGNETIC ABRASIVE TOOLS FINISHING

W. H. Li, S. Q. Yang and X. H. Li

[1]College of Mechanical Engineering, Taiyuan University of Technology
Taiyuan, Shanxi, 030024, P. R. China
E-mail: wenhui_li7190@126.com

Based on the characteristics of solid magnetic abrasive and abrasive flow finishing, a new technology of viscoelastic magnetic abrasive finishing is presented. This paper introduces the composition of viscoelastic magnetic abrasives and realization mode of finishing, establishing the theoretical micro model of viscoelastic magnetic abrasive tools, determining influencing factors of this technology on a positive pressure on the workpiece surface by force analysis, analyzing theoretically and verifying experimentally the finishing mechanism of micro-grinding, slipping, rolling and so on, thus providing the theoretical basis for the preparation of viscoelastic magnetic abrasive and experimental study of this process..

Keywords: finishing, magnetic abrasive tools, viscoelastic, finishing mechanism

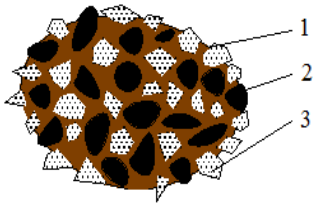
1. Introduction

An important development direction of modern manufacturing technology is precision and ultra precision, fine and ultra-fine processing technology. Surface finishing technology is the main methods to improve the surface quality and surface integrity of the part as a very important part of the precision and ultra-precision processing technology, being widely studied and applied. There are many surface finishing problems such as pipeline of engine, oil and gas transmission pipeline, various valves of the hydraulic components, bearings retainer and other parts in actual production, as well as the development of new composite materials presents new subjects and challenges for surface finishing technology. Developing practical surface finishing technology will provide a more efficient means for solving the finishing problems of composite materials, difficult to cut materials and special-shaped surface [1].

Solid magnetic abrasive has special advantages in finishing effect, finishing efficiency and application range, etc., occupying the important position, having the huge potential application value, where the magnetic abrasive grains directly influence the finishing effect and efficiency. The existing problems of solid magnetic abrasive are easy to disperse and spatter during finishing, complicated preparing processes, high cost, etc.[2-5]. Because the abrasive flow has good fluidity and viscoelasticity, it is helpful to realize the deburring and finishing of special-shaped hole surface, but the equipment requires high pressure (approximately 20MPa), and the special machine and fixture must be designed according to the different parts [6-8]. For this reason, integrating effectively the advantages of solid magnetic abrasive and abrasive flow finishing, a new technology of viscoelastic magnetic abrasive finishing is presented, which provides a more effective method for solving the surface finishing problems of long-thin tube, narrow groove surface and special-shaped surface [9-10].

2. Viscoelastic Magnetic Abrasive Finishing

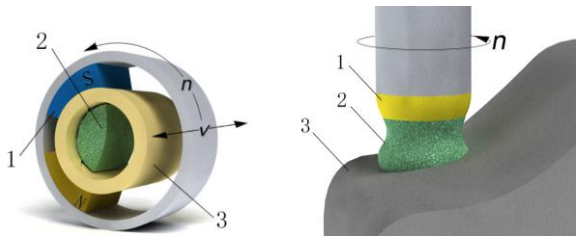
As Figure 1 shows that viscoelastic magnetic abrasive tools are prepared from matrix, magnetic medium phase and abrasive phase by the certain process. Matrix is prepared from carrier (viscoelastic polymer) and many auxiliary additives, and magnetic medium and abrasive phase disperse evenly in the matrix. The abrasive phase surrounded by the magnetic medium phase would be pressed against the workpiece surface because of the magnetic force of magnetic medium phase in magnetic field, thus realizing the finishing of the workpiece. The abrasive phase play a major role in finishing, the magnetic medium providing force for the finishing of abrasive phase, and improving significantly the finishing effect of abrasives under rheological property of the matrix.



1. Matrix. 2. Magnetic medium phase. 3. Abrasive phase
Figure 1. Viscoelastic magnetic abrasive tools schematic diagram

The abrasive tool is semi-solid in appearance, having a certain viscoelasticity. It can adapt to different workpiece surfaces or any part of the workpiece surface because of its own fluidity. It has a certain magnetic, producing a certain force on the workpiece surface in magnetic field. The viscoelastic magnetic abrasive tool would form complex motion with respect to the surface of workpiece under the action of relative rotational motion, reciprocating linear motion and vibration, etc., realizing the finishing of the workpiece surface by micro-grinding, collision, rolling, electrochemical corrosion, etc., under the action of force and relative motion.

The viscoelastic magnetic abrasive tools finishing for cavity surface such as grooves, etc., as shown in Figure 2 (a). The amount of viscoelastic magnetic abrasive was placed on the inner cavity of workpiece, realizing the finishing and deburring of the whole surface under the action of relative rotation and axial rectilinear motion between the magnetic pole and the workpiece. When finishing free surface as shown in Figure 2 (b), the amount of viscoelastic magnetic abrasive was placed between the workpiece and the magnetic pole, the abrasive tools forming a "fretting flexible tools" under the control action of the magnetic field, realizing super-smooth or flawless sub-surface finishing of the whole by rotary motion of the pole and the shaping motion of the free surface (by means of numerical controlled machine used in the former surface shaping process, using viscoelastic magnetic abrasive as cutting tools).



(a) groove surface (b) free surface
1.magnetic pole 2. viscoelastic magnetic abrasives 3. workpiece
Figure 2. Viscoelastic magnetic abrasive tools finishing schematic diagram

3. Viscoelastic Magnetic Abrasive Tools Microcosmic Model

The flow morphology of the viscoelastic abrasive tools depends on the solid phase volume fraction in abrasive tools. Because they are uncompatible between the magnetic medium phase, abrasive phase and the matrix, without considering the porosity, according to the principle of volume additivity:

$$\frac{V_{\mathrm{x}}}{m_{\mathrm{x}}}=\frac{V_{\mathrm{g}}}{m_{\mathrm{x}}}+\frac{V_{\mathrm{y}}}{m_{\mathrm{x}}} \tag{1}$$

Where, V_{x} is the volume of suspension (L), m_{x} is the mass of suspension(kg), V_{g} is the volume of solid phase (L), V_{y} is the volume of the fluid phase (L).

Two representation methods of the solid content of suspension are mass fraction ω_g and volume fraction ϕ_g of the solid phase, respectively:

$$\omega_g = m_g/m_x \quad \phi_g = V_g/V_x \tag{2}$$

When m_x=1kg, there is:

$$\frac{\omega_g}{\rho_g \phi_g} = \frac{\omega_g}{\rho_g} + \frac{1-\omega_g}{\rho_y} \tag{3}$$

Where, ρ_g is density of the solid phase (kg/m^3), ρ_y is density of the fluid phase (kg/m^3).

Thus the relationship between the mass fraction and the volume fraction is:

$$\phi_g = \frac{\omega_g / \rho_g}{\omega_g / \rho_g + (1-\omega_g)/\rho_y} \tag{4}$$

There are magnetic medium phase and abrasive phase in the viscoelastic abrasive tools, referring to the calculation method of mixed slurry solid phase volume fraction in the sticky slurry flow [11], there is:

$$\frac{\omega_c}{\rho_c \phi_c} = \frac{\omega_m}{\rho_m \phi_m} = \frac{\omega_c}{\rho_c} + \frac{\omega_m}{\rho_m} + \frac{1-\omega_c-\omega_m}{\rho_j} \tag{5}$$

Where ω_c is mass fraction of the magnetic medium phase, ω_m is mass fraction of the abrasive phase, ρ_c is density of the magnetic medium phase (kg/m^3), ρ_m is density of the abrasive medium phase (kg/m^3), ϕ_c is volume fraction of the magnetic medium phase, ρ_j is density of the matrix (kg/m^3).

The volume fraction of magnetic medium phase and abrasive phase can be derived from formula (5):

$$\phi_c = \frac{\omega_c}{\omega_c + \frac{\omega_m \rho_c}{\rho_m} + \frac{(1-\omega_c-\omega_m)\rho_c}{\rho_j}} \tag{6}$$

$$\phi_m = \frac{\omega_m}{\omega_m + \frac{\omega_c \rho_m}{\rho_c} + \frac{(1-\omega_c-\omega_m)\rho_m}{\rho_j}} \tag{7}$$

According to the suspension viscosity calculation formula with wider application range put forward by Chong [12]:

$$\eta = \eta_0 \left[1 + 0.75 \frac{\phi_g / \phi_{vc}}{1 - \phi_g / \phi_{vc}} \right]^2 \tag{8}$$

Where η is viscosity of the suspension (cp), η_0 is viscosity of the matrix (cp), ϕ_g is volume fraction of the solid fraction, ϕ_{vc} is maximum volume fraction of the solid fraction in fluid.

The viscosity of viscoelastic magnetic abrasive tools may be derived by substituting the variables of the formula (6) and (7) into the formula (8):

$$\eta_1 = \eta_0 \left[\left[1 + 0.75 \frac{\phi_c / \phi_{vc}}{1 - \phi_c / \phi_{vc}} \right]^2 + \left[1 + 0.75 \frac{\phi_m / \phi_{vc}}{1 - \phi_m / \phi_{vc}} \right]^2 \right] \tag{9}$$

Where η_1 is viscosity of the viscoelastic magnetic abrasive tools (cp).

Although the shear stress increases non-linearly with the shear, deformation rate in the non-Newtonian fluid experiment, but can be expressed by a power function. The viscoelastic magnetic abrasive tools is the power-law fluid, its constitutive equation is [13]:

$$F_\tau = K \dot{\gamma}^n \tag{10}$$

Where F_τ is shear stress of the viscoelastic magnetic abrasive tools, K is the viscosity, $\dot{\gamma}$ is the shear rate, n is the value of power exponent (Dilatant fluid when n>1, Newtonian fluid when n=1, pseudoplastic fluid when n<1).

Above all, the shear stress of the viscoelastic magnetic abrasive tools is:

$$F_\tau = \eta_0 \left[\left[1 + 0.75 \frac{\phi_c / \phi_{vc}}{1 - \phi_c / \phi_{vc}} \right]^2 + \left[1 + 0.75 \frac{\phi_m / \phi_{vc}}{1 - \phi_m / \phi_{vc}} \right]^2 \right] \dot{\gamma}^n \tag{11}$$

It can be found from formula(11) that the shear stress of the viscoelastic magnetic abrasive tools is not only related to the viscosity of the matrix, but also the maximum volume fraction, the volume fraction of the magnetic medium and abrasive medium phase. Figure 3 shows the variation in shear stress with the volume fraction of magnetic medium and abrasive medium phase under certain conditions.

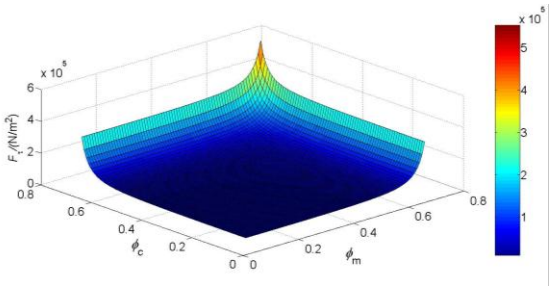


Figure 3. Variation in shear stress with the volume fraction of magnetic medium and abrasive medium phase

4. Force Analysis

The force analysis of the viscoelastic magnetic abrasive tools is shown in Fig.4. In Figure 4, F_N is the normal force of the workpiece on the abrasive tools, F_m is the magnetic force, F_z is the finishing resistance, m is the mass of abrasive, θ is the angle between F_m and normal direction, α is the angle between gravity and normal direction, β is the angle between F_z and tangential direction, M is the viscous torque of the abrasive subjected to the matrix.

The tangential direction:

$$F_1 = F_m \sin\theta - F_z \cos\beta - mg\sin\alpha \tag{12}$$

The normal direction:

$$F_2 = F_{\mathrm{m}}\cos\theta\text{-}F_{\mathrm{z}}\sin\beta + m\ g\cos\alpha + F_C - F_{\mathrm{N}} = 0 \tag{13}$$

$$F_{\mathrm{z}} = fF_{\mathrm{N}} / \cos\beta \tag{14}$$

$$F_{\mathrm{N}} = \frac{F_{\mathrm{m}}\cos\theta + mg\cos(2\pi nt) + F_C}{1 + f\tan\beta} \tag{15}$$

Where F_1 is the resultant force in tangential direction (N), F_2 is the resultant force in normal direction (N), f is the friction coefficient between magnetic medium phase and workpiece (r/min), t is the finishing time (min).

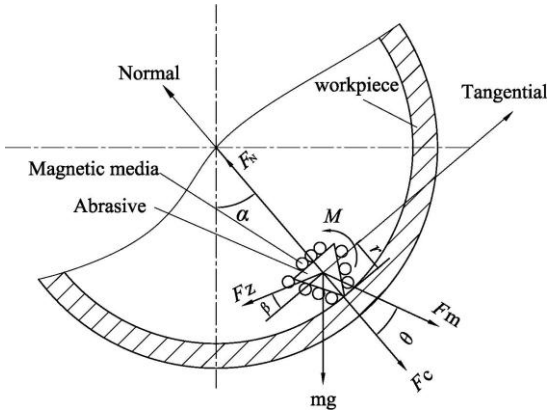


Figure 4. The force analysis diagram of the viscoelastic abrasive tools

Assuming that the magnetic fields distribute evenly in the area of magnetic permeable material, according to the circuit design and practical calculation, the formula of the magnetic field force can be simplified as follows [14]:

$$F_{\mathrm{m}} = \frac{\pi B^2 r^2}{2\mu_0} \tag{16}$$

Then, F_{N} can be calculated as:

$$F_{\mathrm{N}} = \frac{\pi B^2 r^2\cos\theta + 2gm\mu_0\cos(2\pi nt) + 2\mu_0 F_C}{2\mu_0(1 + f\tan\beta)} \tag{17}$$

Where μ_0 is the permeability of vacuum (H/m), B is the magnetic induction intensity (T), r is the simplified radius of single abrasive (mm).

The overturning moment of workpiece acting on the abrasive in the tangential direction can be calculated as:

$$T = F_1 \cdot r = (F_{\mathrm{m}}\sin\theta\text{-}F_{\mathrm{z}}\cos\beta\text{-}mg\sin\alpha)\cdot r \tag{18}$$

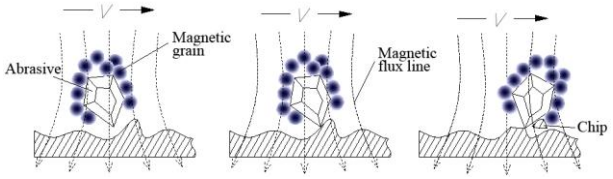
Where T is the overturning moment (N·mm).

When the overturning moment M is larger than the viscous torque T, the viscoelasticity of the abrasive is relatively large, thus abrasive does not roll, but sliding relatively to the workpiece surface, the major roles are grinding and sliding at this time; when M<T, the abrasive would roll, thus producing collision and rolling action; when M≈T, the abrasive would produce grinding, rolling and slipping effect.

5. Finishing Mechanism

5.1. Micro- Grinding and Sliding

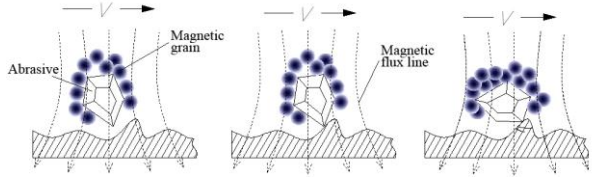
Without considering the gravity, magnetic abrasive grain in machining region could be pressed into the workpiece surface under the action of the centrifugal force and normal component force of magnetic force. Because the magnetic fields drive abrasives moving, thus producing the relative sliding between the abrasive and the workpiece surface, the surface atoms of the workpiece would be removed if thrust force of magnetic force and interaction force of magnetic particles is greater than the machining resistance (mainly of the sum of the binding force between surface atoms and sub-surface atoms of the workpiece and the friction between abrasives and the workpiece). The action process is shown in Figure 5.



(a) Abrasives approaching to the surface peak (b)Abrasives grinding the surface peak (c) The chip peeling off
Figure 5. Surface grinding and sliding of viscoelastic magnetic abrasive tools acting on the surface of workpiece

5.2. Rolling

The abrasives could be pressed to the workpiece surface under the action of the centrifugal force and normal component force of magnetic force after entering into the finishing region. If the tangential component force of the magnetic field force and the interaction force between particles were less than the finishing resistance, the abrasives would not produce grinding action on the workpiece surface. According to mechanical analysis of abrasives in the finishing region, there is relative rolling besides relative sliding. If the kinetic energy of abrasives rolling on the workpiece surface were greater than the energy for elastic deformation and plastic deformation of the workpiece, the micro peak of the workpiece surface would produce plastic deformation and flat under the rolling action of the abrasives, thus realizing rolling finishing of the workpiece surface. When the kinetic energy of abrasives acting on the workpiece surface is not enough to make the micro peak producing plastic deformation, but only elastic deformation, under the alternate effect of a large number of abrasives, micro peak of the surface would harden or break because of fatigue, forming chip, thus realizing the finishing of the workpiece surface. The action process is shown in Figure 6.



(a) Abrasives approaching to the surface peak (b)Abrasives colliding , rolling and sliding the surface peak (c) Plastic deformation of the peak stacking
Figure 6. Collision, rolling and sliding process of abrasive magnetic tools acting on the surface of workpiece

It could be described mathematically as follows:

$$\frac{1}{2} J_0 \omega^2 \geq E_\varepsilon + E_P \tag{19}$$

Where J_0 is the moment of inertia of abrasives (kg.m^2), ω is the rotation angular velocity of abrasives (rad/s), E_ε is elastic deformation energy of the workpiece (J), E_P is the plastic deformation energy of the workpiece (J).

6. Preparation of the Abrasive Tools

The preparation flow of the viscoelastic magnetic abrasive tools includes the preparation of matrix, uniform mixture of the abrasive and magnetic medium phase, the preparation of abrasive. The matrix is mainly prepared form low viscosity silicone or PVC or PVC paste resin as the carrier, and then fully stirring by JJ - 1 precise timing power electric blender and mixing with the loose abrasive uniformly mixed. The plasticizer, liquid paraffin and stearic acid were added to the abrasive tools in the process of mixing, the preparation process is shown in Figure 7. The photo of prepared abrasive tools is shown in Figure 8. Figure 8 (b) shows the abrasive distribution morphology of the three abrasive tools under binocular inverted microscope 4 XB. The shiny parts in the Figure 8 (b) are SiC particle and Fe powder, the big size particle is SiC and the small is the iron powder. As can be seen from Figure 8 that the aggregate phenomenon is less and the carrier has certain dispersion.

The plastic circle tube of 20mm in diameter and 12mm in height filled with the viscoelastic magnetic abrasive tools were placed on a piece of transparent glass plate; Then gently removing the plastic tube (gently shaking the plastic pipe to ensure that the deformation of the abrasive tools is less), the strength of the abrasive liquidity can be reflected by the increasing amount of the contact area of the abrasive and glass plate placing it on a glass plate after 2h. The contrast diagram of the contact area before and after the flow of the abrasive tools is shown in Figure 9.

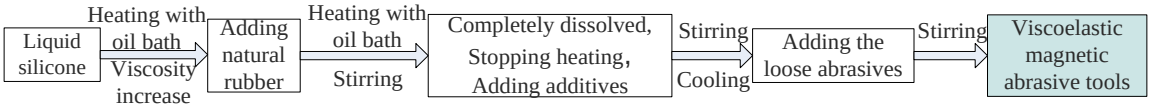
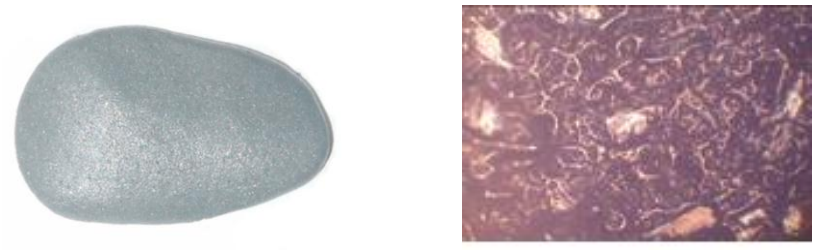
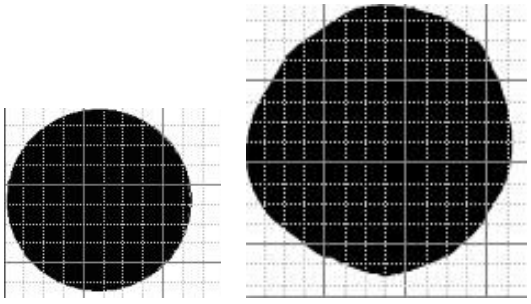


Figure 7. The preparation process flow of the low viscosity silicone viscoelastic magnetic abrasive tools



(a) Real photos of the abrasive tools (b) The Microscope photo of the abrasive distribution morphology (200X)
Figure 8. Photo of the low viscosity silicone viscoelastic magnetic abrasive tools



(a) Before flow (b) after flow
Figure 9. The contrast of the liquidity

7. Experiment Study

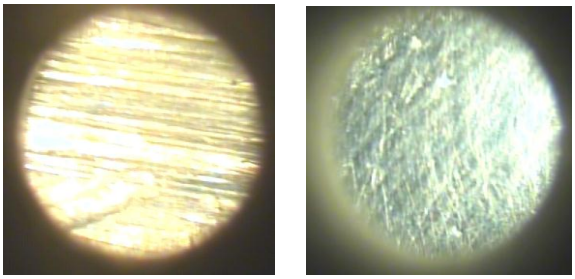
The magnetic field generating device of viscoelastic magnetic abrasive tools finishing is mainly composed of a pair of permanent magnets, where the workpiece is stationary, the magnetic poles rotating and moving along the axis. The experiment device is shown in Figure 10[15]. The concrete experimental parameters are as shown in Table 1.

TABLE 1. Experimental Parameters

Parameters	Magnetic poles			Viscoelastic magnetic abrasive tools			Work-piece			Finishing time
	Arrangement angle	Rotational speed	Feed speed	Abrasive phase	Magnetic medium phase	Filling amount	Inner diameter	External diameter	Material	
Value	90°	0.55m/s	10mm/s	Green SiC, Size 240#	Iron power	20g	32mm	36mm	YL12	20min

Figure 10. The viscoelastic magnetic abrasive tools finishing experimental device

The image contrast of hole surface acquired by SW01 keyhole video instrument before and after finishing is shown in Figure 11.



（a）Before finishing （b）After finishing 20min

Figure 11. Microscopy of hole surface before and after finishing (40X)

It can be seen from Figure 11 that cutting markers are visible on the workpiece surface before finishing, exhibiting isotropy, the surface being rough. After finishing 20min, the surface exhibits anisotropy due to the random motion of abrasives under the action of rotation and moving of the magnetic poles, the surface quality improving greatly. It can be seen from the finishing texture that the trajectory of abrasives is crossing reticulate clay, mainly caused by micro-grinding and sliding of abrasives, and micro-pits of the workpiece surface are mainly caused by rolling of abrasives.

8. Conclusions

(1) The viscoelastic magnetic abrasive tools finishing technology is advanced on the basis of analysing synthetically the advantages and disadvantages of the solid magnetic abrasive tools and abrasive flow finishing, and the composition of viscoelastic composition and realization way of the technology are analysed.

(2) The micro mathematical model has been developed for viscoelastic magnetic abrasive tools, the shear stress formula having been deduced, the simulation having been carried out, the influencing factors having been determined for the shear stress, which was not only related to the viscosity of the matrix, but also the maximum volume fraction of solid phase in the matrix and the volume fraction of magnetic medium phase and abrasive phase. This provide theoretical basis for the preparation of viscoelastic magnetic abrasive tools.

(3) The mechanism of viscoelastic magnetic abrasive tools finishing has been analysed in theory; the formula of normal force and overturning moment having been deduced; the finishing mechanism of micro-grinding, sliding and rolling having been determined, which has been validated experimentally. This provide theoretical basis for experimental study of the technology.

Acknowledgments

This project is supported by national natural science foundation of China (51175365). New Teacher Research Fund for the Doctoral Program of Higher Education of China（20111402120003）and provincial natural science foundation of Shanxi (2011021021-4).

References

1. Qingpu, M. (2009) The Development of NC Magnetic Grinding Machine Tool, *Science & Technology Information*, 36, 81-83.
2. Jain, V.K. (2009) Magnetic Field Assisted Abrasive Based Micro-/nano-finishing, *Journal of Materials Processing Technology*, 20(209), 6022-6038.
3. Xuequan, L., Renmao, L., Dejin, H., et.al. (1999) The Preparation Technology of Magnetic Abrasives, *Diamond & Abrasive Engineering*, 5, 20-22.
4. Wangyan, Dejin, H. (2003) Grinding Mechanism and Preparation Technology of Magnetic Abrasives, *Tool Technology*, 6, 32-33.
5. Jiancheng, F., Zhuji, J., Wenji, X., et.al. (2001) Research on Magnetic Abrasive Machining by Applying A Rotating Magnetic Field", *China Mechnical Engineering*, 11, 1304-1307.
6. Lisheng, W., Yuanzong, L. (2005) Progress of Abrasive Flow Machining, *Diamond & Abrasive Engineering*, 145(1), 69-73.
7. Walia, R.S., Shan, H.S., Kumar, P. (2009) Enhancing AFM Process Productivity Through Improved Fixturing, *The International Journal of Advanced Manufacturing Technology*, 44, 700-709.
8. Sankar, M.R., Jain, V.K., Ramkumar, J., (2009) Experimental Investigations into Rotating Workpiece Abrasive Flow Finishing, *Wear*, 267, 43-51.
9. Li, W.H., Li, X.H. (2012) Technique Project and Experimental Study of Viscoelastic Magnetic Abrasive Finishing, *Advanced Science Letters*, 12, 30-33.
10. Li, W.H., Li, X.H., (2011) New Viscoelastic Magnetic Abrasive and Its Surface Finishing Process, *Advanced Materials Research*, 314-316, 300-303.
11. Weijun, J., Youyuan, D., Huijun, G. (2003) The Principle of Chemical Engineering (the second edition), *Tsinghua University press*.
12. Shouci, L. (2003) Industry Suspension Performance, Modulation and Processing, *Chemical Industry Press*.
13. Zhaoguo, L., Guoyan, C. (1998) Non-Newtonian Fluid Mechanics, *Petroleum University Press*.
14. Wangyu (2007) Calculation of Magnetic Force of Permanent Magnet Devices, *Journal of Magnetic Materials and Devices*, 5, 49-52.
15. Li, W.H., Li, X.H., Yang, S.Q., Sun, C.H. (2012) A Surface Finishing Device and Method of Viscoelastic Magnetic Abrasive Tools, China, ZL201010234181.8, 2012-10-03.

Received on the 21st of June 2013

Mathematical and Computer Modelling

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 130-135
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

THE FULL ARRAY COUNT WITHOUT FIXED POINT K

Zh. Li[1], L. Ding[2]

[1]Dalian Jiaotong University, Dalian, China
Phone(+86)-041184106999. E-mail: lizhibinky@163.com,
[2]Dalian Jiaotong University, Dalian, China
Phone(+86)-041186223616. E-mail: dinglijiaky@163.com

This paper mainly studies the full array count without fixed point K. It aims to get the display expressions and recurrence relations when K is prime number by discussing the nature of function generating from the index of this count, and provide each count for odd-even arrangement in the full array count without fixed point K.

Keywords: fixed point K, full array, generating function.

1. Introduction

A variety of combination representations of full array for set [n] have been discussed in [1], one of which is presenting arrangement as function. As for any full array $a_1a_2...a_n$, define $\pi(i)=a_i$ to get bijective $\pi:[n]\to[n]$. The fixed point of function π is the solution for equation $\pi(x)=x$. Full Derangement is the full array whose i position is not i. This problem was first solved by the famous mathematician Leonhard Euler. This paper regards full derangement as full array count excluding fixed point in order to promote its theory.

If function π is a bijective one, define the complex of function π is $\pi^2(x)=\pi\circ\pi(x)=\pi(\pi(x))$, then it can be defined recursively that the k complex of π is $\pi^k(x)=\pi(\pi^{k-1}(x))$. Define the fixed point k of function π is the fixed point of k Composite function, that is the solution for $\pi^2(x)=x$. When K=1, fixed point k degradates to original concept of fixed point.

The main results of this article is getting from the method of generating functions [2, 9, 10]. Generating function is a method that uses power series to present counting results. If a_n is counting function, its general generating function and exponential generating function is

$$f(x)=\sum_{n_\infty=0}^{\infty}a_nx^n$$

$$g(x)=\sum_{n=0}^{\infty}a_n\frac{x^n}{n!}$$

The generating function method can transform a lot of combinatorial enumeration problems into unified power series computational problems, which is a powerful tool for solving combinatorial counting problems.

This paper studies the combinatorial enumeration for full array excluding fixed point k. Its structure is as follows: first part introduces the circle indicates for arrangement, the second part provides generating function of full array excluding fixed point k by circle indicates, the third part studies count for odd-even arrangement respectively in the full array count excluding fixed point K , and the fourth part is conclusion.

2. Circle Indicates for Arrangement

Definition symbols used here are referenced from *Enumerative Combinatorics* [1] by Richard P.Stanley. Use symbol ς_n to indicate a full array collection for $[n]$. Since an arrangement can be expressed as a bijective π, we can say arrangement π equals bijective π, which is not indiscriminated

in this paper. For an arbitrary integer, consider the sequence, is a finite integer, so there is the smallest integer such that. Is defined as the arrangement of long circle, and shows that according to the definition of each element is only part of a circle, so the order can be expressed as a number of disjoint circle [8].

To any arrangement π in ς_n, define the type of π is $(c_1, c_2, ..., c_n)$, in which c_i means the number of long circle, it is clearly that there is an equation $1c_1 + 2c_2 + ... + nc_n = n$.

Proposition 1 [1]. The number of $\pi \in \varsigma_n$ for arrangement $(c_1, c_2, ..., c_n)$ is

$$n! \Big/ 1^{c_1} c_1! 2^{c_2} c_2! .. n^{c_n} c_n! .$$

The index generation function of the full array $[n]$ is trivial [6], because the total number of full array is $n!$, so its index generating function can be expressed as

$$\sum_{n=0}^{\infty} n! \frac{x^n}{n!} = \frac{1}{1-x}.$$

According to proposition 1, we can derive the equation, and it is helpful for later derivation,

$$\sum_{n=0}^{\infty} \sum_{1c_1 + 2c_2 + ... + nc_n = n} n! \Big/ 1^{c_1} c_1! 2^{c_2} c_2! .. n^{c_n} c_n! \frac{x^n}{n!}$$

$$= \sum_{n=0}^{\infty} \sum_{1c_1 + 2c_2 + ... + nc_n = n} 1 \Big/ 1^{c_1} c_1! 2^{c_2} c_2! .. n^{c_n} c_n! x^{1c_1 + 2c_2 + ... + nc_n}$$

$$= \sum_{n=0}^{\infty} \sum_{1c_1 + 2c_2 + ... + nc_n = n} \prod_{i=1}^{n} \frac{x^{ic_i}}{i^{c_i} c_i!} = \prod_{i=1}^{n} \left(\sum_{n=0}^{\infty} \frac{\left(\frac{x^i}{i} \right)^n}{n!} \right) = \prod_{i=1}^{\infty} \exp\left(\frac{x^i}{i} \right).$$

$$= \exp\left(\sum_{i=1}^{\infty} \frac{x^i}{i} \right) = \exp(-\ln(1-x)) = \frac{1}{1-x}$$

According to previous derivation we can get followings easily, as for $(c_1, c_2, ..., c_n)$, satisfy the rest are equal to 0 Except c_k, the Ordering counting index generation function is $\exp\left(\frac{x^k}{k} \right)$.

3. The Count for Array Excluding Fixed Point K

For the circle indicates for any arrangement π, if x is in the circle with l long, define the order for x is l, according to the knowledge of the cyclic group in abstract algebra, we can know that $\pi^m(x) = x$ when $l | m$. Thus, if no k-order fixed point if and only if the ring in any one cycle length is not divisible by k, i.e., excluding the circle of length k factor.

Use T_n to represent the number of arrangement excluding fixed point k in ς_n, thus,

$$\sum_{n=0}^{\infty} T_n \frac{x^n}{n!} = \prod_{i \neq k} \exp\left(\frac{x^i}{i} \right) = \prod_{i=1}^{\infty} \exp\left(\frac{x^i}{i} \right) \Big/ \prod_{i|k} \exp\left(\frac{x^i}{i} \right) = 1 \Big/ (1-x) \prod_{i|k} \exp\left(\frac{x^i}{i} \right).$$

When k=1, we can get index generating function for derangement [7] $\sum_{n=0}^{\infty} D_n \frac{x^n}{n!} = e^{-x}/(1-x)$.

When k is a prime number, we can get index generating function for arrangement without k-order fixed point

$$\sum_{n=0}^{\infty} T_n \frac{x^n}{n!} = e^{-x} \Bigg/ \left((1-x)\exp\left(\frac{x^k}{k}\right)\right) = \left(\sum_{n=0}^{\infty} D_n \frac{x^n}{n!}\right)\left(\sum_{n=0}^{\infty} (-1)^n \frac{x^{kn}}{k^n n!}\right).$$

According to the multiplication rule, it can be that

$$T_n = \sum_{ki+j=n} (-1)^i \frac{n!}{k^i i!\, j!} D_j \ .$$

In addition,

$$\left(\sum_{n=0}^{\infty} T_n \frac{x^n}{n!}\right)(1-x) = \exp\left(-x - \frac{x^k}{k}\right).$$

Through the derivation for both sides it can be

$$\left(\sum_{n=0}^{\infty} T_{n+1} \frac{x^n}{n!}\right)(1-x) - \left(\sum_{n=0}^{\infty} T_n \frac{x^n}{n!}\right) = \exp\left(-x - \frac{x^k}{k}\right)\left(-1 - x^{k-1}\right)$$

$$\left(\sum_{n=0}^{\infty} T_{n+1} \frac{x^n}{n!}\right)(1-x) - \left(\sum_{n=0}^{\infty} T_n \frac{x^n}{n!}\right) = \left(\sum_{n=0}^{\infty} T_n \frac{x^n}{n!}\right)(1-x)\left(-1 - x^{k-1}\right)$$

$$\sum_{n=0}^{\infty} T_{n+1} \frac{x^n}{n!} - \sum_{n=1}^{\infty} n T_n \frac{x^n}{n!} =$$

$$\sum_{n=1}^{\infty} n T_{n-1} \frac{x^n}{n!} + \sum_{n=k}^{\infty} n(n-1)\ldots(n-k+1) T_{n-k} \frac{x^n}{n!} - \sum_{n=k-1}^{\infty} n(n-1)\ldots(n-k+2) T_{n-k+1} \frac{x^n}{n!} \quad .$$

$\Rightarrow$

$$T_n = \begin{cases} (n-1)(T_{n-1} + T_{n-2}) + (n-1)\ldots(n-k+1)(T_{n-k} - (n-k)T_{n-k-1}), & n \geq k+1 \\ (n-1)(T_{n-1} + T_{n-2}) + (k-1)!, & n = k \\ (n-1)(T_{n-1} + T_{n-2}), & 2 \leq n \leq k-1 \\ 0, & n = 1 \\ 1, & n = 0 \end{cases}$$

It can be seen easily that when $n \leq k-1$, the recurrence relations of T_n and D_n is the same, and it satisfies the same initial value. Thus when $n \leq k-1$, $T_n = D_n$, because there is no possible k circle for any arrangement when $n \leq k-1$, which can be seen from principle.

4. The Count for Odd-Even Arrangement in the Full Array Count Excluding Fixed Point K

We discuss parity arrangement count in derangement, which need to be discussed first the matrix form for arrangement. For any arrangement, it can be expressed that for each column and row there is a 1, and the remaining elements are the $n \times n$ square for 0.

According to the definition for determinant [3], as for any determinant of matrix $(a_{ij})_{n\times n}$, it can be

represented that
$\sum_{i_1 i_2 \ldots i_n \in \varsigma_n} (-1)^{\tau(i_1 i_2 \ldots i_n)} a_{1i_1} a_{2i_2} \ldots a_{ni_n}$.

$\tau(i_1 i_2 \ldots i_n)$ is the inverse number of arrangement $i_1 i_2 \ldots i_n$. For any arrangement which has forbidden position, define conditional matrix A satisfies the i position can be j, let $a_{ij} = 1$, otherwise $a_{ij} = 0$. In this forbidden ranking even in the collection are arranged number of conditions minus the number of odd arrangement is equal to the corresponding determinant of a matrix.

The conditional matrix for derangement is the matrix that diagonal is zero, and remaining elements is one. Its determinant is calculated as follows:

$$\det\begin{bmatrix} 0 & 1 & \cdots & 1 & 1 \\ 1 & 0 & & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 & 1 \\ 1 & 1 & & 1 & 0 \end{bmatrix} = (n-1) \cdot \det\begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & 0 & & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 & 1 \\ 1 & 1 & & 1 & 0 \end{bmatrix}$$

$$= (n-1) \cdot \det\begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 0 & -1 & & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 0 \\ 0 & 0 & & 0 & -1 \end{bmatrix} = (n-1)(-1)^{n-1}$$

Thus, make D_n^e, D_n° represent the number of odd arrangement and even arrangement respectively in derangement, then $D_n^e - D_n^\circ = (n-1)(-1)^{n-1}$.

In addition $D_n^e + D_n^\circ = D_n$

Then we can get D_n^e, D_n°.

For most odd-even arrangement count of the array without fixed point k, make T_n^e, T_n° represent the number of odd arrangement and even arrangement. Since in the circle indication for any arrangement in ς_n, odd circle can be decomposed into even exchanges, and even circle can be decomposed into odd exchanges. Therefore,

$$\sum_{n=0}^{\infty} (T_n^e - T_n^\circ) \frac{x^n}{n!}$$

$$= \prod_{i \neq k} \exp\left((-1)^{i-1} \frac{x^i}{i} \right) = \prod_{i=1}^{\infty} \exp\left((-1)^{i-1} \frac{x^i}{i} \right) \Big/ \prod_{i|k} \exp\left((-1)^{i-1} \frac{x^i}{i} \right).$$

$$= (1+x) \Big/ \prod_{i|k} \exp\left((-1)^{i-1} \frac{x^i}{i} \right) = (1+x) \prod_{i|k} \exp\left((-1)^{i} \frac{x^i}{i} \right)$$

When k=1, we can get

$$\sum_{n=0}^{\infty} (D_n^e - D_n^\circ) \frac{x^n}{n!} = (1+x) e^{-x} .$$

We can get the same conclusion in accordance with Taylor expansion

$D_n^e - D_n^\circ = (n-1)(-1)^{n-1}$.

When k is prime number,

$$\sum_{n=0}^{\infty}\left(T_n^e - T_n^\circ\right)\frac{x^n}{n!} = (1+x)e^{-x}\exp\left((-1)^k\frac{x^k}{k}\right)$$

$$= \left(\sum_{n=0}^{\infty}(n-1)(-1)^{n-1}\right)\left(\sum_{n=0}^{\infty}\frac{(-1)^{kn}}{k^n}\frac{x^{kn}}{n!}\right).$$

Thus, we can get the following formula:

$$T_n^e - T_n^\circ = \sum_{ki+j=n}\frac{n!(-1)^{n-1}(j-1)}{k^i i!\, j!}.$$

5. Conclusion

This paper studies the combinatorial enumeration and odd-even arrangement count for full array without fixed-point k. It gets the display expression by calculating index generation function when k is prime number.

$$T_n = \sum_{ki+j=n}(-1)^i\frac{n!}{k^i i!\, j!}D_j$$

$$T_n^e - T_n^\circ = \sum_{ki+j=n}\frac{n!(-1)^{n-1}(j-1)}{k^i i!\, j!}.$$

When k is prime number, the recurrence relation of T_n is

$$T_n = \begin{cases}(n-1)(T_{n-1}+T_{n-2})+(n-1)\ldots(n-k+1)(T_{n-k}-(n-k)T_{n-k-1}), & n \ge k+1\\ (n-1)(T_{n-1}+T_{n-2})+(k-1)!, & n=k\\ (n-1)(T_{n-1}+T_{n-2}), & 2 \le n \le k-1\\ 0, & n=1\\ 1, & n=0\end{cases}$$

It gets good result for the combinatorial enumeration for the full array without fixed point k by the formula, which greatly facilitates the probability calculation for related problems.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No.61273022.

References

1. Stanley, R.P. (2011) *Enumerative Combinatorics*, 1 [M]. NewYork: Cambridge University Press.
2. Wilf, H.S. (2006) *Generatingfunctionology*, third ed. [M]. A K Peters, Ltd., Wellesley, MA.
3. Jacobson, N. (1953) *Lectures in Abstract Algebra*, vol. II—Linear Algebra [M].D. van Nostrand Co., Princeton, NJ.
4. Knuth, D.E. (1998) *The Art of Computer Programming*,Vol.3 [M] Jersey: Addison-Wesley.

5. Andrews, G.E. (1976) *The Theory of Partitions* [M]. Jersey: Addison-Wesley.
6. Bona, M. (2004) *Combinatorics of Permutations*[M].Boca Raton:Chapman and Hall/CRC.
7. Richard, A. (2004) *Brualdi.Introductory Combinatorics*.fourthed [M]. Pearson Education.
8. Kung, J.P.S. (1981) *The cycle structure of a linear transformation over a finite field* [J]. Linear Algebra Appl. 36, 141–155.
9. Flajolet, P., Sedgewick, R. (2009) *Analytic Combinatorics* [M]. Cambridge University Press,Cambridge.
10. MacMahon, P.A. (1913) *The indices of permutations and the derivation therefrom of functions of a single variable associated with the permutations of any assemblage of objects* [J]. Amer. J. Math. 35, 281–322; reprinted in [1.3], 508–549.

Received on the 21st of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 136-141
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

EIGENFACES AND IDENTIFICATION OF COLOR FACE IMAGES

L. Liu, Sh. Li, H. Zhang, Ch.Zhang, Zh. Jia*

School of Mathematics and Statistics
Jiangsu Normal University, Xuzhou 221116, China
Phone: (+86)-15162227511. E-mail: jiazhigang1980@126.com

To preserve the relation of three-primary colours of colour images, this paper represents a colour image by a quaternion matrix. Colour eigenfaces are defined by eigenvectors of the covariance matrix of a set of faces. Several main colour eigenfaces form the eigenface subspace, which is used to calculate the distance between new image and known colour faces. That helps us to judge whether an input colour image is a face and further more whether it is a known one.

Keywords: colour face, quaternion matrix, colour eigenface, face subspace

1. Introduction

Three basic colours for colon images are red, green, and blue (RGB for short). Quaternion has been introduced in accordance with the relation of RGB to represent a set of colon face images and compute the eigenvalues and eigenvectors, which are used to span the face subspace. The identification of colour face images can be operated in the assistance of such face subspace.

Irish mathematician, physicist, and astronomer -William Rowan Hamilton in 1843 invented quaternion [1], which is his greatest achievement. When he was considering expanding complex to higher dimension (a complex can be seen as one point of the plane), he found four-dimensional space could create quaternion. In recent years, quaternion has been applied to many advanced fields [2, 3, 4]. The position and rotation of three-dimensional objects can be expressed by it when we draw pictures by computer and do some related graphic analysis. At the same time, we can find its appearance in cybernetic, signal processing, attitude control, physics and orbital mechanics and it is also used to express position and rotation. Explicitly, quaternion is the non-commutative expansion of complex. It represents a four-dimensional space if the set of quaternions is considered as multi-dimensional real space relative to a two-dimensional space of complex. That is why quaternion plays an important role in the facet of colour image processing [5, 6].

A quaternion q including four components-one real part and three imaginary parts is represented as:

$$q = q_r + q_i i + q_j j + q_k k$$

where q_r, q_i, q_j and q_k are four real numbers, i, j, k are three imaginary units and they are related to each other as follows:

$$i^2 = j^2 = k^2 = ijk = -1, \tag{1}$$

$$ij = -ji = k,\ jk = -kj = i,\ ki = -ik = j. \tag{2}$$

A quaternion is called pure quaternion when the real part is set as zero. One pure quaternion can represent one pixel of the colour image [5] by the expression: $Ri + Gj + Bk$, in which *R*, *G*, and *B* each stands for the value of RGB.

For a quaternion matrix $X_{(q)} = X_r + X_i i + X_j j + X_k k \times Q^{m\times n}, X_n \times R^{m\times n}(n = r, i, j, k)$, then we define its real representation [7, 8]:

$$X_{e(r)} = \begin{matrix} X_r & X_i & X_j & X_k \\ -X_i & X_r & -X_k & X_j \\ -X_j & X_k & X_r & -X_i \\ -X_k & -X_j & X_i & X_r \end{matrix} \times R^{4m\times 4n}.$$

We have $X_{(q)}$, $Y_{(q)} \in Q^{m\times n}$ and make the following conclusion by upper definition:

$$\left(X_{(q)} + Y_{(q)}\right)_{e(r)} = X_{e(r)} + Y_{e(r)},\ \left(X_{(q)}Y_{(q)}\right)_{e(r)} = X_{e(r)}Y_{e(r)},\ \left(X_{(q)}^H\right)_{e(r)} = \left(X_{e(r)}\right)^T,$$

where $X_{(q)}^H$ means the conjugate and transpose of $X_{(q)}$.

In this paper, we use quaternion matrices to deal with colour faces, that is, a quaternion matrix can express a colour image. By making use of its equivalent real matrix, it is possible to calculate the colour eigenfaces. Then, we choose several eigenvectors to span the colour face subspace. In practice, we input one new colour image and compute its distance from the subspace and the face classes respectively to figure out whether it belongs to the known individual or not.

2. Eigenfaces and Face Identification

A computational model of face recognition [9, 10, 11] will have huge influence both in theory and in practice. In this paper, we develop a rapid, reasonable and simple one suitable for a constrained environment.

Colour eigenfaces can be considered as a group of standardized human faces. Each colour face image can be represented by the linear combination of eigenfaces. The recognition of new colour face image is conducted in the subspace spanned by colour eigenfaces. We represent a $N\times N$ colour face image by an $N^2\times 1$ order pure quaternion column vector. Let the training set of colour face images be the column vectors $F1, F2, F3, \cdots, F_M$ (as it shows in Figure 1, $M = 16, N = 128$). Then,

$$\Psi = \frac{1}{M}\sum_{n=1}^{M} F_n$$

is defined as the average face vector. Each face differs from Ψ by $\Phi_i = F_i - \Psi$. The orthogonal vectors u_k are chosen such that

$$\lambda_k = \frac{1}{M}\sum_{n=1}^{M}\left(u_k^T\Phi_n\right)^2$$

is the k -th maximal eigenvalue. The vectors u_k and scalars λ_k are respectively eigenvectors and eigenvalues of the covariance matrix

$$C = \frac{1}{M}\sum_{n=1}^{M}\Phi_n\Phi_n^T = AA^T,$$

where $A = \left[\Phi_1\Phi_2 ... \Phi_M\right]$, $A \in Q^{N^2\times M}$. Due to $C \in Q^{N^2\times N^2}$, there are N^2 pairs of eigenvalues and eigenvectors. If N is huge, working out the eigenfaces of such a big matrix is tough. Generally, M is far smaller than N^2. So, can we figure out the eigenvectors and eigenvalues of a small $M\times M$ order matrix to directly obtain those of C ? We find that:

$$A^T A v_i = \mu_i v_i , \tag{3}$$

$$AA^T A v_i = \mu_i A v_i . \tag{4}$$

From Eq. (3) and Eq. (4), we know that if we want the eigenvectors of $C = AA^T$, we can firstly compute the eigenvector v_i of $A^T A$. Because $A^T A$ is an $M \times M$ order quaternion matrix and M is much smaller than N^2, we can work out v_i easier.

In fact, we just need to use M' eigenfaces corresponding to M' biggest eigenvalues to span the subspace Z. Set $M'=8$ then we can get the subspace. Figure 4 shows the eight main eigenfaces. We assume the new face image Γ (in Figure 3, including the face image and animal's face image), represented by column vector F.

Let the colour image projected onto the subspace based on the following formula:

$$\omega_k = u_k^T (F - \Psi), k = 1,...,M' .$$

Vector $\Omega^T = [\omega_1, \omega_2, ..., \omega_{M'}]$ denotes the weights of new input face image that describe the distribution of eigenfaces spanned the subspace. Ω_k is the average weights obtained from a face class, which can be used to check out which face class the new face image belongs to or which is the best suitable face class. We have three images of each individual here, for example, in Figure 2. To find the right face class, we define

$$\varepsilon_k = \left\| (\Omega - \Omega_k) \right\| / \left\| \Omega \right\| .$$

The simple way to find out the exact face class is to compare the minimal ε_k. If ε_k is less than a certain value, we can tell that the new face belongs to the k-th face class, otherwise it's the face of an unknown individual. To conclusion, ε_k means the distance between new face image and the face classes.

Set $wi = \Phi_i^{'} Z$, $weight = (F - \Psi)' Z$ and $WM = \sum_{i=1}^{M} norm(wi)$ so that $\delta = abs((WM - norm(weight))/norm(weight))$. If δ is less than a certain value, we can judge that the new face image belongs to a human being. To conclusion, δ means the distance between new face image and the face subspace.

Figure 1. Some colour faces in the training set from [12]

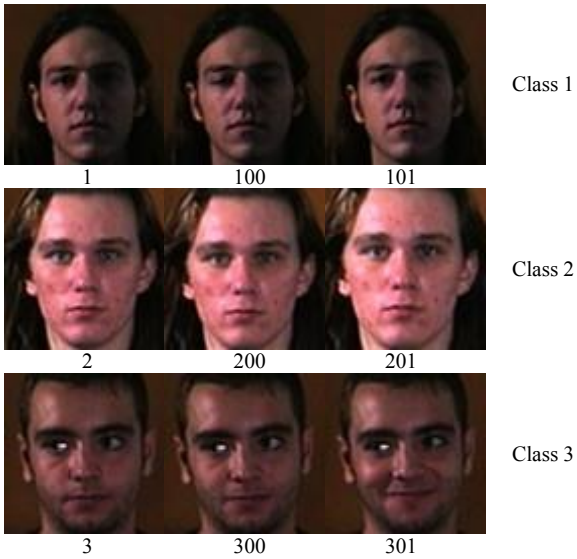


Figure 2. Three face classes

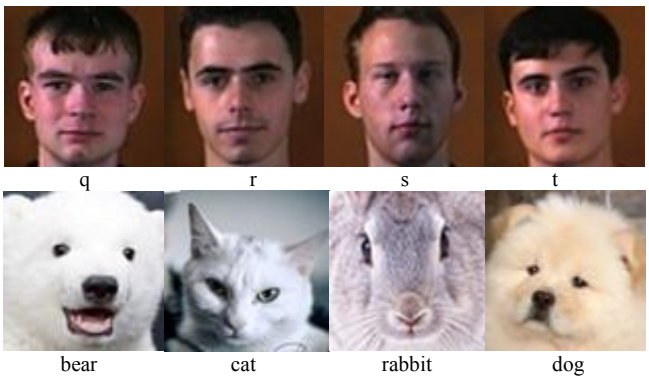


Figure 3. The new face images of both human and animals

3. Result Analysis

3.1. Computation of Eigenfaces

Figure 4. Eight main eigenfaces calculated from the input image

From the above figure, we find that the last eigenface with the biggest eigenvalue containing the most information of the training set.

3.2. Judgment

TABLE 1. Data of δ and ε_k of the faces belonging to the input individual

figure	**1**	**2**	**3**	**4**
δ	0.2154	0.1140	0.6098	0.8640
ε_k	.0356	0.1135	0.0568	0.2413
figure	**5**	**6**	**7**	**8**
δ	0.0262	0.4051	0.2124	0.5567
ε_k	0.1329	0.0223	0.0408	0.0276
figure	**9**	**10**	**11**	**12**
δ	0.0416	0.5881	0.3742	0.5730
ε_k	0.0940	0.1496	0.0299	0.1020
figure	**13**	**14**	**15**	**16**
δ	0.3168	0.6512	0.9345	0.2762
ε_k	0.1242	0.1294	0.0633	0.0865
figure	**17**	**18**	**19**	**20**
δ	0.1860	0.0707	0.4349	0.3255
ε_k	0.0598	0.1348	0.1491	0.2903
figure	**21**	**22**	**23**	**24**
δ	0.3042	0.6519	0.1522	0.5441
ε_k	0.3238	0.2328	0.2422	0.0308
figure	**25**	**26**	**27**	**28**
δ	0.1093	0.3649	0.3356	0.6175
ε_k	0.2142	0.0634	0.1604	0.1130
figure	**29**	**30**	**31**	**32**
δ	0.4300	0.0750	0.6270	0.2557
ε_k	0.0460	0.1116	0.0917	0.0553

TABLE 2. Data of δ and ε_k of the new face of both human and animals

figure	**q**	**r**	**s**	**t**
δ	0.6160	0.4588	0.3181	0.6801
ε_k	0.5205	0.3728	0.5125	0.6576
figure	**bear**	**cat**	**rabbit**	**dog**
δ	0.8180	0.7383	0.7883	0.7926
ε_k	0.6394	0.5915	0.5958	0.5765

4. Conclusion

Situation 1: Generally, if $\delta < 0.7$ and $\varepsilon_k < 0.3$, that is to say, the new face is near to one of the face classes and the subspace, we can judge that the new face belongs to certain individual and we can recognize him or her.

Situation 2: Generally, if $\delta < 0.7$ and $\varepsilon_k < 0.3$, that is to say, the new face is near to the subspace but far away from any face class, we can judge that the new face doesn't belong to any individual and it's a new human face.

Situation 3: Generally, if $\delta < 0.7$ and $\varepsilon_k < 0.3$, that is to say, the new face is far away from the subspace and face classes, we can judge that the new face doesn't belong to any individual and it's also not a human face.

Because of light and shooting angle, the distance δ between a human face and the subspace may be bigger than 0.7 and the distance between an animal face and the subspace may be less than 0.7. These

cases are rare. People can also make a conclusion depending on the second parameter ε_k. For example, the fourth known face have a big δ, but its ε_k is smaller than 0.3. We can say that it is a known face.

References

1. Hamilton, W.R. (1969) *Elements of Quaternions*, New York: Chelsea.
2. Zhang, F. (1997) *Quaternions and matrices of quaternions*, Linear Algebra Appl., 251: 21-57.
3. Adler, S.L. (1995) *Quaternionic Quantum Mechanics and Quantum Fields*, Oxford University Press, New York.
4. Leo, S. D., Rotelli, P. (1992) Quaternion scalar field, Phys. Rev. D 45, 575-579.
5. Bihan, N.L., Sangwine, S.J (2003) Quaternion principal component analysis of colour images. Image Processing, 809-810.
6. Luo, Y., Chen, D. (2006) Face recognition based on colour Gabor features. *J. Image Graph.*, 13(2), 242-243.
7. Jiang, T. (2004) An algorithm for eigenvalues and eigenvectors of quaternion matrices in quaternionic quantum mechanics, *J. Math. Phys.*, 45, 3334-3338.
8. Wang, M., Wei, M., Feng, Y. (2008) An iterative algorithm for least squares problem in quaternionic quanrum theory*, Comp. Phys. Commun.*, 179, 203-207.
9. Turk, M., Pentland, A. (1991) Eigenfaces for recognition. *J. Cognitive Neurosc.*, 3(1), 71-76.
10. Bihan, N.L., Sangwine, S.J. (2003) Quaternion principal component analysis of colour image. In: Proceedings of International Conference on Image Processing. Barcelona, Spain: IEEE, 809-812.
11. Xing, Y., Tan, J. (2011) Colour Image decomposition based on the SVD of quaternion matrix. *J. Engineering Graph.*, 94-96.
12. http://cswww.essex.ac.uk/mv/allfaces/faces95.html

Received on the 21st of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 142-152
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

AN EVALUATION MODEL FOR PASSENGER HUB UNDER FUZZY ENVIRONMENT

***L. Yun*[1,2], *Ch. Ji*[1*], *H. Fan*[3]**

[1]Rail Traffic Control and Safety State Key Laboratory, Beijing Jiaotong University, Shangyuancun3, Beijing, P.R. China,
Phone: (+01) 6626179649. E-mail: yunlifen@gmail.com
[*] Phone: (+86)13701127279. E-mail: chxji@bjtu.edu.cn
[2] Civil and Environmental Engineering, Mississippi State University, 235L Walker Hall, 501 Hardy Road, Starkville, Mississippi State, USA
[3]MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Shangyuancun3, Beijing, P.R. China,
Phone: (+86) 13811752653. E-mail: fanhq1985@126.com

This paper aims to propose a modelling framework for service quality evaluation of passenger hub under fuzzy environment, i.e., when valuators cannot quantify all criteria accurately. The research problem is to evaluate the operation status of passenger hub that maximizes the service level, including not only quantitative index but also qualitative index. A mixed-model combined the Grey Relational Analysis (GRA) and VIseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) has been developed to solve the problem in a fuzzy environment where both criteria and weights could be fuzzy sets. An empirical study for evaluating service quality of three passenger hubs in different times was put forth to illustrate an application of the proposed model. The results reveal a number of interesting insights into service level evaluation, namely, this study has instructive significance in management practice.

Keywords: fuzzy environment; service quality; evaluate; combined-model

1. Introduction

A critical factor in evaluating the performance of urban traffic network is the service quality in the transport hub, which essentially contains not only quantitative index but also qualitative index. With the rapid development of urban traffic network, passenger hub is becoming a distribution centre with different transport modes and large-scale passenger flow, namely, the service level and efficiency of hub directly determine the performance of urban traffic network bring into play.

The transfer is the core service supplied by passenger hub; that is why so many existing studies focus on how to evaluate and improved its efficiency. In fact, there are three main bodies of literatures to the research presented in this paper. The first is the literature on planning program, most often applied to find the best design for hubs in a macro perspective. Demetsky et al. [1] focused on the procedures, which can be used to establish policy for station features, to provide performance measures for subsystems, and to give cost estimates. Alexandre and Wirasinghe [2] did an analysis with a special case, which performed individually for a single pier, several types of pier-satellites, and a set of remote parallel piers connected by an automated people mover (APM).

The second concerns the transfer model in a mid-scale view, which are frequently focus on the efficiency measurement, improved lines, and other things to optimal the passenger organization process. Lee and Schonefld [3] proposed optimization of various transportation modes' slack time to enhance transfer efficiency by researching the transfer relationship of modes. Some scholars [4, 5] used simulate analysis to discuss passenger flow in hubs and found the efficiency bottleneck.

Finally, further research discussed the approaches for comprehensive service quality of hubs since they just considered one kind of equipment before [6, 7]. Yan et al. [8] developed an integer programming model to assist airport authorities to assign common use check-in counters. Raik [9] addressed operation models for workforce planning for check-in systems at airport and developed a binary linear programming formulation. As we can see, the research for hubs has become an important issue in each city [10-15].Some researchers tried to evaluate the service of the hub [16, 17]. Tsaur et al. invited fuzzy set theory into the measurement of performance [18], while one new method of measuring perceived service quality based on triangular fuzzy numbers were proposed by Chien and Tsai [19].

In fact, our mixed model can be seen as an outgrowth of these literatures. However, they did not propose comprehensive evaluation for the hubs transfer service, not to mention the method combined the

subjective and objective. Scientific evaluation has important realistic significance for transfer service quality, which could reflect the operation status directly. To bridge this research gap, this paper proposes a new comprehensive evaluation model under fuzzy environment.

The remainder of the paper is organized as follows. Section 2 proposes a mixed-model combined VIKOR and GRA under fuzzy environment, which considers as a decision matrix to carry out program optimization. Section 3 conducts case studies to test the model, and we get the final comprehensive evaluation index value and ranking order. Compared the SIV orders and service orders with the results, the applicability of the evaluation model can be proved. Section four makes concluding remarks and briefly discusses future research directions.

2. Model Formulation

2.1. GRA Methods

GRA is a grey correlation order to describe the relationship between the order of the strength, size, sequence and some other factors. Compared with the traditional multi-factor analysis (correlation, regression, etc.), the grey correlation analysis demands lower and calculate smaller, especially under the fuzzy environment, and that is why it is easy to be widely used.

GRA method was originally developed in 1982. This method is successfully applied in solving a variety of multiple attribute decision-making problems, such as the hiring decision, the restoration planning for power distribution systems, the inspection of integrated-circuit marking process, the modelling of quality function deployment, etc. GRA decides the best and ideal one by defining the similarity among each option. The option is more preferred if it has more similarity. The procedure of GRA is shown in Figure 1.

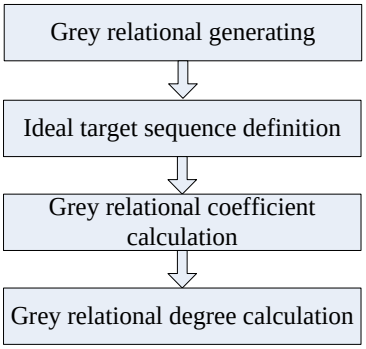


Figure 1. Procedure of grey relational analysis

Let $X = \{x_1, x_2, ...x_j, ...x_n, j = 1,2,3,..n\}$ denote the set of given various alternatives. For an alternative x_j, the rating of the i-th aspect is denoted by f_{ij}, that is, f_{ij} is the value of i-th criterion function for the alternative x_j. m is the number of criteria, $i = 1,2,3,...,m$. Then we can calculated the grey relation coefficient $\gamma(f_{i1}, f_{ij})$ of these alternatives at point i as follows

$$\gamma(f_{i1}, f_{ij}) = \frac{\min_i \min_j \Delta_{ij} + \zeta \max_i \max_j \Delta_{ij}}{\Delta_{ij} + \zeta \max_i \max_j \Delta_{ij}}, \tag{1}$$

where $\Delta_{ij} = |f_{i1} - f_{ij}|$, and ζ is the resolving coefficient $\zeta \in [0,1]$.

The grade of grey relation $\gamma(f_{i1}, f_{ij})$ between x_1 and x_j can be calculated:

$$\gamma(f_{i1}, f_{ij}) = \sum_{i=1}^{m} w_i \gamma(f_{i1}, f_{ij}), \tag{2}$$

where w_i denotes the weight of point/factor i, and $\sum_{i=1}^{m} w_i = 1$.

2.2. VIKOR Methods

VIKOR method tries to maximize the community benefits and minimize the negative effects in the multi-attribute decision calculation, and that is why it could be accepted by compromise solution, which is an optimization method suitable for the comprehensive evaluation calculation process.

The VIKOR method was proposed by Opricovic [20], who introduced the multicriteria ranking index based on the particular measure of "closeness" to the "ideal" solution. The basic concept of VIKOR is to define the ideal solution (positive) and negative ideal solution (negative) at first. Here, the ideal solution is the best value of the alternative criteria in the assessment while the negative ideal solution means the worst.

We use the same assumption, which mentioned in the previous section, and the VIKOR method started with the following form of L_p -metric:

$$L_{pj} = \left\{ \sum_{i=1}^{m} \left[w_i \left(f_i^* - f_{ij} \right) / \left(f_i^* - f_i^- \right) \right]^p \right\}^{1/p}, \quad 1 \le p \le \infty; \quad j = 1,2,...,n. \tag{3}$$

Let F^* be the ideal solution, F^- be the negative ideal solution, and F^c be the feasible solution which is "closest" to F^*. Each value $f_i^* \in F^*$ is the best one in all alternatives, and $f_i^- \in F^-$ is the worst one in all alternatives. We can get the compromise as an agreement established by mutual concessions (Figure 1): $\Delta f_1 = f_1^* - f_1^c$, $\Delta f_2 = f_2^* - f_2^c$.

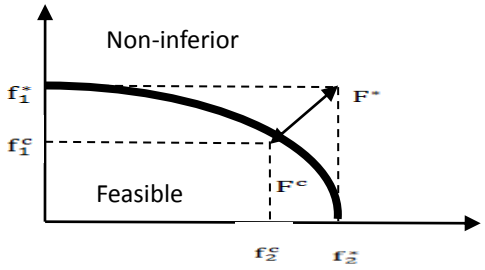


Figure 2. Ideal and compromise solution

We will first briefly review the steps of VIKOR method as follows:

(1) Determine the best f_i^* and the worst f_i^- values of all criterion functions. Assuming that i -th criterion function represents a benefit:

$f_i^* = \max_j f_{ij}$, $f_i^- = \min_j f_{ij}$, $i = 1,2,..m$.

(2) Compute the values S_j and R_j, by the relations

$$S_j = \sum_{i=1}^{m} w_i \left(f_i^* - f_{ij} \right) / \left(f_i^* - f_i^- \right), \quad j = 1,2,...,n, \tag{4}$$

$$R_j = \max_i \left[\left(f_i^* - f_{ij} \right) / \left(f_i^* - f_i^- \right) \right], \quad j = 1,2,...,n, \tag{5}$$

here w_i are the weights of criteria, expressing the relative importance of each criterion.

(3) Compute the values Q_j, which are defined as

$$Q_j = v\left(S_j - S^*\right)/\left(S^- - S^*\right) + (1-v)\left(R_j - R^*\right)/\left(R^- - R^*\right),\ \ j = 1,2,\ldots,n, \tag{6}$$

here $S^* = \min_j S_j$, $S^- = \max_j S_j$, $R^* = \min_j R_j$, $R^- = \max_j R_j$. v is the weight of the strategy of "the majority of criteria", whereas $(1-v)$ is the weight of the individual regret.

2.3. The Combined-Model

Based on the previous introduction of VIKOR and GRA, we will propose the mixed-model combined these methods in this section. As we have discussed before, the most important aspect in the complex decision is the vague information under our fuzzy environment. Therefore, our mixed-model employs a major technique of GRA theory to deal with the vague information as the first step.

Here, some grades of qualitative criteria will be considered as linguistic terms, which can be expressed in fuzzy numbers, as shown in Table 1. Instead of the given weight, the Buckley's fuzzy AHP method is used to calculate them, as this method can allow fuzzy numbers for pairwise comparisons verified and satisfy fundamental consistency requirements.

TABLE 1. Linguistic terms for the fuzzy ratings

Very poor (VP)	(0,0,1)
Poor (P)	(0,1,3)
Medium poor (MP)	(1,3,5)
Fair (F)	(3,5,7)
Medium good (MG)	(5,7,9)
Good (G)	(7,9,10)
Very good (VG)	(9,10,10)

A fuzzy positive reciprocal matrix $\tilde{C} = \left[\tilde{c}_{ij}\right]$ is considered in Buckley's fuzzy AHP method, which defines the fuzzy geometric mean of each row $\tilde{g}_i$ and fuzzy weight $\tilde{w}_i$ corresponding to each criterion.

It is assumed that an evaluation problem contains m possible alternatives and n criteria as before, and the formulae can be written as follows:

$$\tilde{g}_i = \left(\tilde{c}_{i1}(\cdot)\tilde{c}_{i2}(\cdot)\ldots(\cdot)\tilde{c}_{in}\right)^{1/n},\ \ \tilde{w}_i = \tilde{g}_i(\cdot)\left(\tilde{g}_1(+)\tilde{g}_2(+)\ldots(+)\tilde{g}_m\right)^{-1} \tag{7}$$

where $\tilde{w}_i$ can be indicated by a triangular fuzzy number 2.4. $\tilde{w}_i = \left(w_i^m, w_i^n\right)$

A decision group has K judges and each rating (performance value) of alternatives can be calculated as follows:

$$\tilde{f}_{ij} = \frac{1}{K}\left[\tilde{f}_{ij}^1(+)\tilde{f}_{ij}^2(+)\ldots(+)\tilde{f}_{ij}^K\right] = \frac{1}{K}\sum_{t=1}^{K}\tilde{f}_{ij}^t \tag{8}$$

where $\tilde{f}_{ij}^t$ is the fuzzy rating assigned by the t-th judge.

Then we get the fuzzy judgment matrix, which can be described as follows:

$$\tilde{D} = \begin{pmatrix} \tilde{f}_{11} & \cdots & \tilde{f}_{1n} \\ \vdots & \ddots & \vdots \\ \tilde{f}_{m1} & \cdots & \tilde{f}_{mn} \end{pmatrix} = \left[\tilde{f}_{ij}\right]_{m\times n},\ \ \tilde{w} = \left[\tilde{w}_1\tilde{w}_2\ldots\tilde{w}_m\right], \tag{9}$$

where $\tilde{f}_{ij}$ is the fuzzy rating of possible alternative, and $\tilde{w}_i$ is the fuzzy weight of criterion $\tilde{C}_i$. In order to easy to calculate the matrix, we can establish the normalized fuzzy decision matrix which denoted by $\tilde{R}$.

$$\widetilde{R} = \left[\widetilde{r}_{ij}\right]_{m\times n},$$
$$\widetilde{r}_{ij} = \left(\frac{f_j^l}{r_j^*}, \frac{f_j^0}{r_j^*}, \frac{f_j^u}{r_j^*}\right), \ if \ x_j \in B,$$
$$\widetilde{r}_{ij} = \left(\frac{f_j^-}{f_j^u}, \frac{f_j^-}{f_j^0}, \frac{f_j^-}{f_j^l}\right), \ if \ x_j \in C, \tag{10}$$

here, $r_j^* = \max_j f_{ij}$, f_j^- is the lower bound of F, f_j^0 is the median of F, f_j^u is the upper bound of F, and f_j^- is the minimum value of F. B is the abbreviation of benefit criteria set, and C means cost criteria set, namely, $F = B \cup C$.

After completing the performance normalization of various criteria scale, we can define the positive ideal solution F^* and negative ideal solution F^- as follows:

$$F^* = \left[\widetilde{r}_{10}^*, \widetilde{r}_{20}^*, ..., \widetilde{r}_{n0}^*\right], \ F^- = \left[\widetilde{r}_{10}^-, \widetilde{r}_{20}^-, ..., \widetilde{r}_{n0}^-\right],$$

where $\widetilde{r}_{i0}^* = \max_j\left(\widetilde{r}_{ij}\right)$, $\widetilde{r}_{i0}^- = \min_j\left(\widetilde{r}_{ij}\right)$, $i = 1,2,...,n$.

Then we can use the GRA method to get the grey relation coefficient $\gamma\left(\widetilde{r}_{i0}^u, \widetilde{r}_{ij}\right)$ for each alternative as follows:

$$\gamma\left(\widetilde{r}_{i0}^u, \widetilde{r}_{ij}\right) = \frac{\min_i \min_j \Delta_{ij} + \zeta \max_i \max_j \Delta_{ij}}{\Delta_{ij} + \zeta \max_i \max_j \Delta_{ij}} \tag{11}$$

where $u = *, -$, $\Delta_{ij} = \left|\widetilde{r}_{i0}^u - \widetilde{r}_{ij}\right|$, and ζ is the resolving coefficient $\zeta \in [0,1]$.

At last, we must calculate the value of $\widetilde{S}_i$ and $\widetilde{R}_i$ using the method of VIKOR as follows:

$$\widetilde{S}_j = \sum_{i=1}^{n} \gamma\left(\widetilde{r}_{i0}^*, \widetilde{r}_{ij}\right),$$
$$\widetilde{R}_j = \max_i \gamma\left(\widetilde{r}_{i0}^*, \widetilde{r}_{ij}\right), \tag{12}$$

where $i = 1,2,3,...,n$, $j = 1,2,3,...,m$.

Values of $\widetilde{Q}_j$ can be defined as

$$\widetilde{Q}_j = v\left(\frac{\widetilde{S}^* - \widetilde{S}_i}{\widetilde{S}^* - \widetilde{S}^-}\right) + (1-v)\left(\frac{\widetilde{R}_i - \widetilde{R}^*}{\widetilde{R}^- - \widetilde{R}^*}\right), \tag{13}$$

where $\widetilde{S}^* = \max_j \widetilde{S}_j$, $\widetilde{S}^- = \min_j \widetilde{S}_j$, $\widetilde{R}^* = \min_j \widetilde{R}_j$, $\widetilde{R}^- = \max_j \widetilde{R}_j$. v is the weight for the strategy of "the majority of criteria", whereas $(1-v)$ is the weight of the individual regret.

Based on what we discussed in section 2.2, the alternatives can be raked and our results are three ranking lists by the values $\widetilde{S}$, $\widetilde{R}$ and $\widetilde{Q}$. If an alternative x_j has the better value of $\widetilde{Q}$, which approaches to 0, it means that it is closer to the positive ideal solution and farther from the negative ideal solution. That is what we are trying to select from a set of feasible alternatives.

Noted that, our rank of $\tilde{Q}$ will work only when they satisfied the condition below:

$$\tilde{Q}'' - \tilde{Q}' \geq \frac{1}{(n-1)}, \tag{14}$$

here, $\tilde{Q}', \tilde{Q}''$ are the first and second values in rank of $\tilde{Q}$, and n is the number of alternatives.

3. Case Study

In this section, we will illustrate how the proposed model is applied to a real-world problem by using case study. As the capital of China, Beijing is the centre of National politics, culture, transport, tourism and international exchanges. What's more, this city has the largest number of passenger hubs, that's the reason why we choose this city as the study case. Beijing South Railway Station (BRS), Beijing Xizhimen Subway Station (BXZS) and Beijing Xuanwumen Subway Station (BXWS) are selected here as the alternatives, and note that BRS is the largest railway station in Asia. Since the number of passengers affects service quality directly, we choose the different data in both rush and common periods as the different alternatives. Then we get 6 data sets, BRS(rp), BRS(cp), BXZS (rp), BXZS (cp), BXWS (rp), and BXWS (cp).

Five intuitive criteria are listed in Table 2.

TABLE 2. Service quality evaluation criteria

Evaluation on criteria	Descriptions
Efficiency level (C_1)	Cost of processing time, can be measured by average queue length
Unobstructed level (C_2)	Information display for passengers, can be measured by average stagnate time
Coordination level (C_3)	Total number of passenger through
Security level (C_4)	Sense of security, like service in case of fire
Comfort (C_5)	Air/temperature conditioning, seat congestion

3.1. Alternatives Component Quantification

Although C_1, C_2 and C_3 are quantified indexes, simulation model is established to get these values. On the other side, C_4 and C_5 can be obtained by providing survey questionnaires to passengers.

3.1.1. Quantitative Index

There are many existing models to simulate the pedestrian walking behaviour, and we choose social mechanics model, which based on the Newtonian mechanics and pedestrian behaviour propensity as the simulation model here. The method of pedestrian force analysis will be used because of the characteristics of passenger flow in hubs, including high density and self-organization phenomena. The walking parameters can be reset according to different traffic environment, as shown in Figure 3.

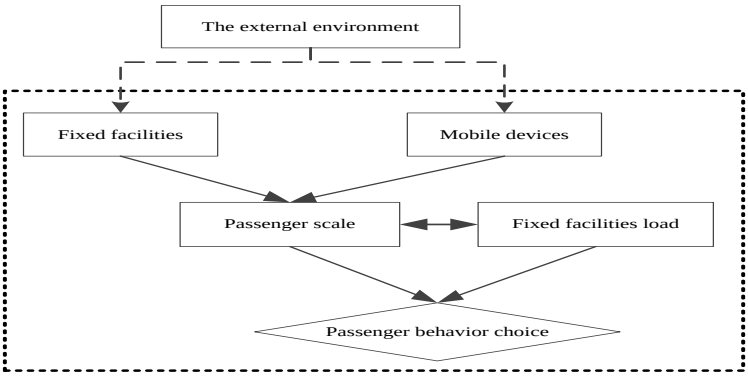


Figure 3. Hub simulation environment

The purpose of our simulation is to reproduce the distribution process, which is actually successive and can be divided into many successive time steps T_i. Since the simulation is not our focus, we will just show the simulation flow in Figure 4.

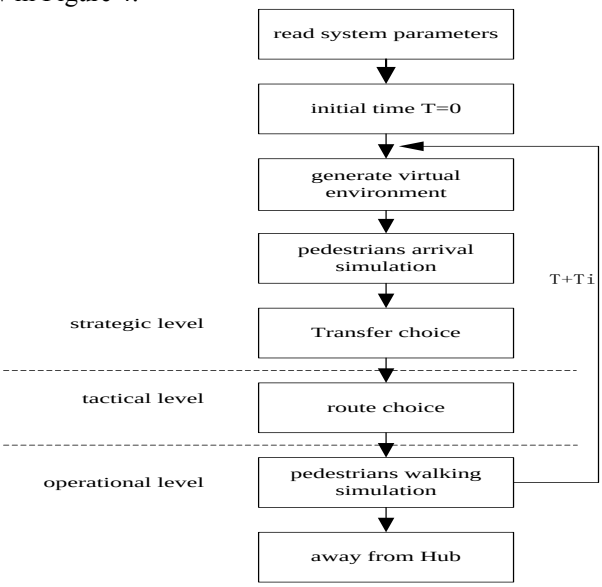


Figure 4. Simulation flow

As the largest passenger transport hub in Asia, BRS integrates many modes, such as high-speed rail, subway, and some other public transport mode. The facility configuration and simulation region design scheme are shown in Figure 5.

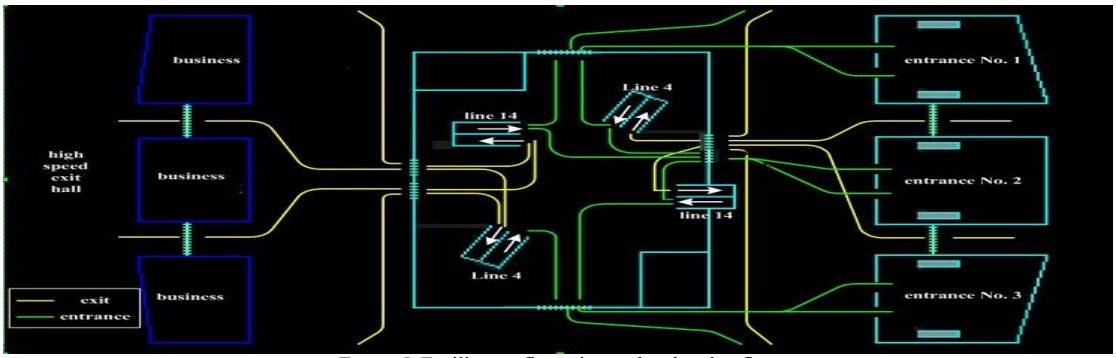


Figure 5. Facility configuration and pedestrian flow

The model is coded in Phython, and all our problem instances are solved at a PC with 3.3 GHz CPU and 4GB RAM by commercial software, Vissim 5.1. The results of the model can be seen in Table 6 and Table 3. Similarly, the simulation results of BXZS and BWXS are listed in Table 4 and Table 5.

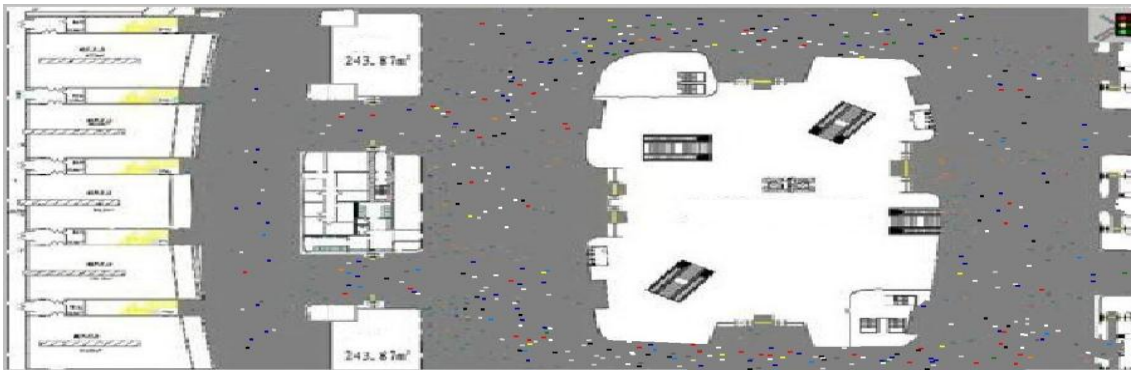


Figure 6. Simulation screenshot

TABLE 3. Simulation results of BRS[1]

Arrive number	Average lost time(s)	Average stagnate time(s)	Average number of times for stopping	Average queue length(m)	Maximum queue length(m)	Total number of queue
20000(cp)	39.3	29.3	7.81	0.6	5.3	67.3
50000(rp)	142.6	117.7	24.41	14.7	28.2	70.3

TABLE 4. Simulation results of BXZS

Arrive number	Average lost time(s)	Average stagnate time(s)	Average number of times for stopping	Average queue length(m)	Maximum queue length(m)	Total number of queue	SIV
10000(cp)	42.3	30.1	6.98	1.1	5.7	69.6	72.4
15000(rp)	152.1	116.7	25.6	15.1	29.8	156.3	268.8

TABLE 5. Simulation results of BXWS

Arrive number	Average lost time(s)	Average stagnate time(s)	Average number of times for stopping	Average queue length(m)	Maximum queue length(m)	Total number of queue	SIV
10000(cp)	52.3	31.6	7.98	1.3	6.9	70.5	83.9
15000(rp)	160.3	120.5	27.8	16.4	30.1	160.4	280.8

Now we get the value of C_1, C_2 and C_3 as describe in Table 2.

In order to do the comparative analysis, we proposed one concept of service invalid value (SIV), which can be calculated as the sum of average lost time and stagnate time. As we all known, the service quality and efficiency can be reflect directly with the waiting time, that's why we will compare the rank order of SIV with our mixed-model results as the authentication.

After calculate all alternatives, we can get the ranking order of the hubs' SIV as follows:

$\mathrm{BRS(cp)} < \mathrm{BXZS(cp)} < \mathrm{BXWS(cp)} < \mathrm{BRS(rp)} < \mathrm{BXZS(rp)} < \mathrm{BXWS(rp)}$.

Although SIV can be mirrored the average service quality of the passenger hub in an easy way, they are in the reverse order. That means the ranking order of service quality is

$\mathrm{BRS(cp)} > \mathrm{BXZS(cp)} > \mathrm{BXWS(cp)} > \mathrm{BRS(rp)} > \mathrm{BXZS(rp)} > \mathrm{BXWS(rp)}$.

3.1.2. Qualitative Index

Since C_4 and C_5 are qualitative index which cannot be obtained with exact numbers, we can describe them with linguistic variables as shown in Table 1.

The Questionnaire Survey method is used to get the description of these two criteria. For the

[1] Here, cp means common period, and rp means rush period.

questionnaires, forms were given randomly to visitors to answer in site, and were taken back to analyse data.

A total of 100 questionnaires were distributed, and 59 were returned (the return rate is 59%), of which 11 were invalid and 48 were valid.

The average fuzzy performance values of our alternatives are listed in the last two columns in Table 6. If we look at these columns and study them for a moment, some interesting insights can be seen.

TABLE 6. Quantitative values of six alternatives

		C_1	C_2	C_3	C_4	C_5
common period	BRS	0.6	29.3	67.3	(7.54,9.01,9.13)	(7.32,8.79,9.11)
	BXZS	1.1	30.1	69.6	(7.03,7.86,8.35)	(6.70,7.89,8.42)
	BXWS	1.3	31.6	70.5	(6.78,7.13,7.99)	(6.54,7.59,8.01)
rush period	BRS	14.7	117.7	70.3	(7.11,7.59,8.41)	(6.98,7.32,8.13)
	BXZS	15.1	116.7	156.3	(6.14,6.84,6.92)	(5.01,5.37,6.32)
	BXWS	16.4	120.5	160.4	(5.97,6.31,6.07)	(5.21,5.73,6.18)

Note that when the values of C_1, C_2 and C_3 are lower, the service quality will get better, namely, they are cost criteria sets in our model. That's inverse to C_4 and C_5, which means the last two criteria are the benefit criteria sets.

As we mentioned in Eq. (10), C_1, C_2 and C_3 affiliated to the set B, while C_4 and C_5 affiliated to the set C. In fact, these criteria have the trade-off relationship, so the reciprocal will be adopted in the simulation process. The ranking order of these six alternatives' service quality is listed as one choice question in the survey, and we get 89% responses ranked as follows:

$$\text{BRS(cp)} > \text{BXZS(cp)} > \text{BRS(rp)} > \text{BXWS(cp)} > \text{BXZS(rp)} > \text{BXWS(rp)}$$

This is one intuitive sort, which is almost conformed to the rank order of SIV. This compare can partly prove that the result of our simulation is accords with the reality. Note that, this intuitive sort is also one counterpart for our evaluation model.

3.2. Evaluation Computing and Analysis

Based on Eq. (7), Eq. (9), the overall weights and BNP values of all criteria can be calculated. The results are shown in Table 7.

To analyse the importance of these criteria, we calculate the BNP values.

The results illustrate that C_4, security level, is the most important evaluation criterion with the highest value. Clearly, the passengers are very concerned about security issue. This is probably because too many traffic accidents happened in recent years. However, the criterion C_1 are not far behind the safety, which illustrate that the key factor of r passengers is still the transfer efficiency. That is easy to understand because the primary function of passenger hub is transfer.

Therefore, if the passenger transport hub wants to improve service quality, it must pay more attention to security level first.

TABLE 7. Criteria weights for five criteria

Criteria	Overall weights	BNP values
C_1	(0.168,0.192,0.251)	0.204
C_2	(0.157,0.181,0.244)	0.194
C_3	(0.099,0.161,0.191)	0.150
C_4	(0.197,0.303,0.342)	0.281
C_5	(0.128,0.169,0.224)	0.174

According to Eq. (9), the fuzzy normalized decision matrix can be calculated with fuzzy performance matrix $\tilde{D}$ and fuzzy weight matrix $\tilde{w}$. Then, F^* and F^- can be obtained using Eq. (11), and so on. The results are listed in Table 8. ($\zeta = 0.5$)

TABLE 8. The ranking results of combined methods

		$\tilde{S}$	$\tilde{R}$	$\tilde{Q}_{(v=1)}$	Rank	$\tilde{Q}_{(v=0.5)}$	Rank	$\tilde{Q}_{(v=0)}$	Rank
common period	BRS	6.3891	0.6831	0.0000	1	0.0000	1	0.0000	1
	BXZS	5.8796	0.7679	0.3984	3	0.3330	2	0.2676	2
	BXWS	5.6428	0.8443	0.5836	4	0.5461	4	0.5087	4
rush period	BRS	6.0184	0.8274	0.2899	2	0.3726	3	0.4553	3
	BXZS	5.4831	0.9301	0.7085	5	0.7440	5	0.7794	5
	BXWS	5.1103	1.0000	1.0000	6	1.0000	6	1.0000	6

When v value is 1, the $\tilde{Q}$ values of 6 alternatives are 0, 0.3984, 0.5836, 0.2899, 0.7085, and 1, respectively. Therefore, the ranking order of the given hubs in different periods is

$$\text{BRS(cp)} > \text{BRS(rp)} > \text{BXZS(cp)} > \text{BXWS(cp)} > \text{BXZS(rp)} > \text{BXWS(rp)}.$$

It means during common period, the best service level of passenger transport hub is evidenced in BRS(cp). Therefore, BRS hub can provide useful suggestions to improve service quality to other hubs.

Similarly, when v value changes to 0.5, it can be considered as produce ranking order of $\tilde{S}$ and $\tilde{R}$ values, while the ranking order of the given hubs is BRS(cp) > BXZS(cp) > BRS(rp) > BXWS(cp) > BXZS(rp) > BXWS(rp).

Note that, no matter how value of v changed, the value of $\tilde{Q}_{BRS(cp)}$ is always zero. Namely, the service quality of this alternative is the best in these hubs. At the same time, BXWS(rp) faces the totally opposite situation, which means BXWS's service quality is the worst.

Finally, we can discuss the practicality of our model by comparing our rank order with rank of SIV and rank of questionnaire survey. They are conformed to each other. Hence, our proposed model is applicable in practice.

More importantly than all of that, these results actually provide some very useful guidance for the management of the passenger hub.

4. Conclusions

In order to optimize the public service management in the passenger transport hub, this paper proposed a mixed evaluation model that considers both qualitative index and quantitative index under the fuzzy environment. Combined with GRA and VIKOR, we formulated a comprehensive model to determine the optimal hub that minimize the total system cost and maximize the system quality.

This model was tested over a number of data sets with different alternatives, and it showed that our model could efficiently solve these instances to the true optimum. Then we analysed the importance of the criteria and conducted a series of sensitivity analysis to show how the rank order changes with parameter values. We have a number of interesting observations from these experiments. For example, no matter how v 's value changed, the ranking order is also kind of the same.

This research can be extended in several directions. The commercial solver, though working fine for the tested cases, may fail to solve this model for larger-scale practical problem instances. It will be worthwhile exploring more efficient customized algorithms. This study also has a potential to connect evaluation model with emergency passenger behaviour research. If further data on both operation patterns and human behaviour in passenger hub is available, this model can be applied to more realistic real-world cases, and the results may help understand and improve the service quality of public service systems for all situations.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China through Grant # 71171015, the State Key Laboratory of Rail Traffic Control and Safety through Contract # RCS2011ZT002, the National High Technology Research and Development Program of China (863 Program) (No. 2012AA112403), and the Fundamental Research Funds for the Central Universities (2013YJS053).

References

1. Demetsky, M.J., Hoel, L.A., Virkler, M.R. (1978) A transit station design process, *Transport Research Record*, 81-85.
2. Barros, A.G., Wirasinghe, S.C. (2003) Optimal terminal configurations for new large aircraft operations, *Transport Research Part A*, 37, 315-331.
3. Lee, K.K.T., Schonefld, P. (1991) Optimal slack time for timed transfers at a transit terminal, *Journal of Advanced Transportation*, 25, 281-308.
4. Seti, J.R., Hutchinson, B.G. (1994) Passenger-terminal simulation model, *Journal of transportation engineering*, 120, 517-535.
5. Lin, Y-D., Trani, A.A. (2000) Airport automated people mover systems: analysis with a hybrid computer simulation model, *Transport Research Record*, 45-57.
6. Weihua, Liu. (2003) Analysis of evaluation indexes quantum used in the yard layout of high way hub", *China Journal of Highway and Transport*, 16, 86-89.
7. Mansel, D.M., Menaker, P.J., Hartnett, G. (1998) High-capacity light rail transit: balancing station side and rail side capacities, *Transport Research Record*, 170-178.
8. Yan, S., Tang C.-H., Chen, M. (2004) A model and a solution algorithm for airport common use check-in counter assignments, *Transport Research Part A,* 38, 101-125
9. Stolletz, R. (2010) Operational workforce planning for check-in counters at airports, *Transport Research Part E*, 46, 414-425.
10. Perk, V.A., Foreman, C. (2003) Florida metropolitan planning organization reports on transit capacity and quality of service: First-year evaluation, *Transport Research Record*, 128-134.
11. Jia, H.F., Yang, L.L., Tang, M. (2009) Pedestrian Flow Characteristics Analysis and Model Parameter Calibration in Comprehensive Transport Terminal, *Journal of Transportation Systems Engineering and Information Technology*, 9, 117–123.
12. Stone, B., Wang, Y. (2011) Airport Logic: Usability Testing, Prototyping, and Analysis of an Airport Way finding Application, *HCI International 2011 Posters' Extended Abstracts*, 81-84.
13. Mastoris, A., Mouzakitis, S., Askounis, D. (2011) A Distributed Online Service-Oriented Platform for the Airport Community, *International Journal of Computer and Communication Engineering*, 2, 343-347.
14. Sohn, C.S., Kim, W.K., Lee, C. (2012) IT Framework and User Requirement Analysis for Smart Airports, *Springer Netherlands: Computer Science and its Applications*.
15. Shi, C.L., Zhong, M.H., Nong, X.Z., He, L., Shi, J.H., Feng, G.G. (2012) Modelling and safety strategy of passenger evacuation in a metro station in China, *Safety Science*, 50, 1319-1332.
16. Truitt, L.J., Haynes, R. (1994) Evaluating service quality and productivity in the regional airline industry", *Transportation Journal*, 33, 21-32.
17. Yang, H., Bell, M.G.H., Meng, Q. (2000) Modelling the capacity and level of service of urban transportation networks. *Transport Research Part B*, 34, 255–275.
18. Tsaura, S.-H., Changb, T.-Y., Yena, C.-H. (2002) The evaluation of airline service quality by fuzzy MCDM, *Tourism Management*, 23, 107–115.
19. Chien, C.-J., Tsai, H.-H. (2000) Using fuzzy numbers to evaluate perceived service quality, *Fuzzy Sets and Systems*, 116, 289-300.
20. Opricovic, S. (1998) Multicriteria, Optimization of civil engineering systems. *Belgrade: Faculty of Civil Engineering*.

Received on the 20th of December 2013

Mathematical and Computer Modelling

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 153-161
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

SUBWAY BEARING FAULT DIAGNOSIS METHODOLOGY RESEARCH BASED ON WAVELET PACKAGE AND NEURAL NETWORK

Xi Li[1] Yu. Zhang[2], L. M.Jia[3]

[1]State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, No.3 Shangyuan Village, Haidian District, Beijing, 100044,China
Subway Operation Technology Centre, Mass Transit Railway Operation Corporation LTD, Suburban railway depot, Huojiaying Village, Changping District, Beijing,102208, China
Phone:(+86)-10-82225713-8202.Email: bjtulx@163.com
[2]Subway Operation Technology Centre, Mass Transit Railway Operation Corporation LTD, Suburban railway depot, Huojiaying Village, Changping District, Beijing,102208, China
Email: beijingdt@sohu.com
[3]State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, No.3 Shangyuan Village, Haidian District, Beijing, 100044,China
Email: jialm@vip.sina.com

In order to diagnose different kinds of subway rolling bearing fault, a new method of fault diagnosis methodology based on wavelet packet and neural network was put forward. Vibration signal of subway vehicle rolling bearing was collected by Piezoelectric Accelerometer. The collected signal was de-noised by wavelet, and then decomposed by wavelet packet, construct the eigenvector. The signal was taken as fault samples to train the Elman neural network and BP neural network. The whole process finally recognizes fault types and realizes intelligent fault diagnosis. Test results show that the application of fault diagnosis method can effectively diagnose rolling bearing faults such as fatigue, peeling and crack, which occurred in inner ring, outer ring and the rolling body surface during subway vehicle operating. The fault diagnosis method has high application value in subway operation process.

Keywords: subway vehicle, rolling bearing, fault diagnosis, wavelet analysis, neural network

1. Introduction

Recent years, with rapid development of subway all over the world, the safety situation of subway is not optimistic, such as train safety problems are constantly emerging, accidents often happen. In addition, the above fault is not only a serious impact to the vehicle operation, once the accident occurs it will bring huge economic losses and casualties.

The foundation of subway safety operation is the key systems of subway vehicle, and the running system is the most important. Therefore, Subway vehicle safety monitoring, early warning and fault diagnosis technology research is taking more and more attention to different kinds of researchers [1, 2]. In subway vehicle operation process, the major reason of wheels, bearing fault is caused by the vibration impact of infrastructure such as wheel and rail, and the generated damage will cause greater fault impact. The vibration in subway actual operation process such as equipment failure, fatigue fault, bearing fault and frame crack are potential hazards of subway operation.

Signal processing is an effective method for subway vehicle early-warning and fault diagnosis [3, 4, 5, 6, 7, 8]. In order to clearly demonstrate the fault characteristic frequency, filtering and resonance demodulation are always used to analysis the collected signal. However, the effect of these diagnosis methods will be affect by resonant and filter band selection [11].

Combining the wavelet packet transform and artificial neural network is a powerful means to identify fault signal from nonlinear and unsteady signal, which can effectively solve the problem of adaptive window and fault feature extraction problem. Therefore, this paper put forward wavelet packet transform and neural network method for subway vehicle rolling bearing fault diagnosis. Firstly, de-noise to the collected signal, then analysed the signal by wavelet packet transform combined with Elman and BP neural network, extract the fault feature, and finally recognized fault of the subway vehicle rolling bearing.

2. Wavelet Analysis

The realization of this work supposes the availability of a great number of repetitions of samples responding to the same known theoretical model. In practice, as the theoretical model is unknown, we use the Monte-Carlo method based on the generation of the data by computer according to a fixed theoretical model.

Wavelet analysis includes continuous wavelet transform and discrete wavelet transform. It is a powerful tool for nonlinear and unsteady signal processing. Wavelet packet analysis is an extended method with more detailed analysis and reconstruction out of wavelet analysis for signal processing. It decomposed the frequency band into multiple layers, further decomposed the high frequency signal, adaptive to choose corresponding frequency band, make match the signal spectrum according to the analysis of signal characteristics, and improve the time-frequency resolution [9, 10].

In the research, three-layer wavelet packet decomposition method is put forward; the structure is shown in figure 1. Wavelet packet decomposition deposed the signal into low frequency part a_1 and high frequency part d_1. In the next layer of decomposition, both the two parts a_1 and d1 are decomposed into low frequency and high frequency part aa_2, da_2, ad_2 , dd_2, and so on, the signal can be further decomposed. Therefore, wavelet packet decomposition is a finer decomposition method.

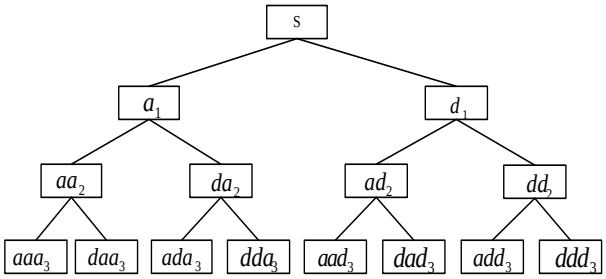


Figure 1. Signal wavelet packet decomposition schemes

Define V_j as Dimension subspace, W_j as Wavelet subspace, U_j^n as new subspace [6]

$$\begin{cases} U_j^0 = V_j \\ U_j^1 = W_j \end{cases}, \ j \in Z$$

And then, orthogonal decomposition of Hilbert space $V_{j+1} = V_j \oplus W_j$ can be unified by decomposition of U_j^n,

$$U_{j+1}^0 = U_j^0 \oplus U_j^1, \ j \in Z$$

Define subspace U_j^n is closure space of the function of, $u_n(t)$ space, and U_j^{2n} is the closure space of function $u_{2n}(t)$, $u_n(t)$ meet the dual dimension equation below:

$$\begin{cases} u_{2n}(t) = \sqrt{2}\sum_{k\in Z} h(k) u_n(2t-k) \\ u_{2n+1}(t) = \sqrt{2}\sum_{k\in Z} g(k) u_n(2t-k) \end{cases} \tag{2.1}$$

In the formula, $g(k) = (-1)^k h(1-k)$, when $n = 0$:

$$\begin{cases} u_0(t) = \sum_{k\in Z} h_k u_0(2t-k) \\ u_1(t) = \sum_{k\in Z} g_k u_0(2t-k) \end{cases} \tag{2.2}$$

In the analysis, $\phi(t)$ and $\psi(t)$ meet the dual dimension equation,

$$\begin{cases} \phi(t)=\sum_{k\in Z} h_k \phi(2t-k) & \{h_k\}_{k\in Z}\in l^2 \\ \psi(t)=\sum_{k\in Z} g_k \phi(2t-k) & \{g_k\}_{k\in Z}\in l^2 \end{cases} \tag{2.3}$$

$$U_{j+1}^n = U_j^n \oplus U_{j+1}^{2n+1}, \quad j\in Z, n\in Z \tag{2.4}$$

Wavelet packet decomposition and reconstruction algorithm: define $g_j^n(t)\in u_j^n$, $g_j^n(t)$ can be represented as:

$$g_j^n(t)=\sum_l d_l^{j-n} u_n\left(2^j t-l\right). \tag{2.5}$$

Wavelet packet decomposition algorithm:

$$\begin{cases} d_l^{j,2n}=\sum_k a_{k-2l} d_k^{j+1,n} \\ d_l^{j,2n+1}=\sum_k b_{k-2l} d_k^{j+1,n} \end{cases} \tag{2.6}$$

Wavelet packet reconstruction algorithm:

$$d_l^{j+1,n}=\sum_k \left[h_{l-2k} d_k^{j,2n} + g_{l-2k} d_k^{j,2n+1}\right] \tag{2.7}$$

3. Neural Network

3.1. Elman Neural Network

Elman recurrent neural network with less train times, quick convergence speed and associative memory features, is widely used in device fault diagnosis.

Elman recurrent neural network includes four layers: input layer, middle layer (hidden layer), undertake layer, and output layer, as shown in Figure 2. Undertake layer also known as context or state layer, is used to remember the former time output value of middle layer, can be considered as a step delay operator.

Output of hidden layer automatic connected to its input layer through the delay and storage of undertake layer. Elman network was sensitive to historical status data and has a strong temporal and spatial learning ability and good dynamic properties.

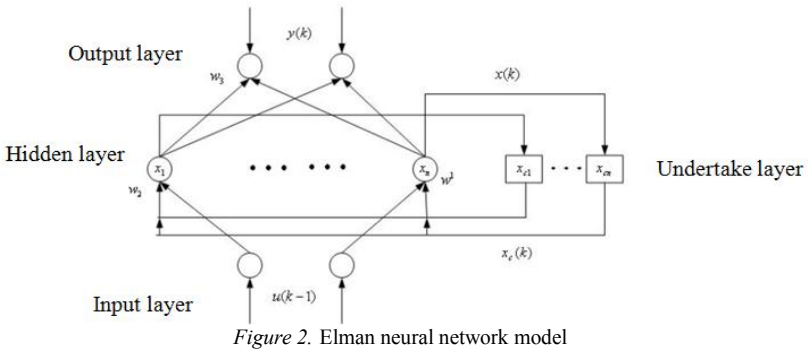


Figure 2. Elman neural network model

State-space of Elman networks was expressed as follows:

$$y(k)=g\left(w^3x(k)\right) \tag{3.1}$$

$$x(k)=f\left(w^1x_c(k)+w^2(u(k-1))\right) \tag{3.2}$$

$$x_c(k)=x(k-1) \tag{3.3}$$

y, x, u, x_c are known as node vector of m-dimensional output, unit vector of n-dimensional middle layer node, r-dimensional input vector and n-dimensional feedback state vector.

w^3, w^2, w^1 are known as the link value among middle layer to output layer, input layer to middle layer and undertake layer to middle layer.

$g(\bullet)$ is transfer function of output neuron, linear combination of middle layer output. $f(\bullet)$ is transfer function of middle neuron.

3.2. BP Neural Network

BP neural network is consists of input layer, hidden layer and output layer, as shown in Figure 3. Input layer neurons number is equal to the number of fault feature vector elements $X=(x_1,x_2,...,x_m)$. The output layer neurons number is equal to the number of fault pattern vector elements $Y=(y_1,y_2,...,y_n)$. Number of the hidden layer neurons is generally determined by the experiment [12, 13].

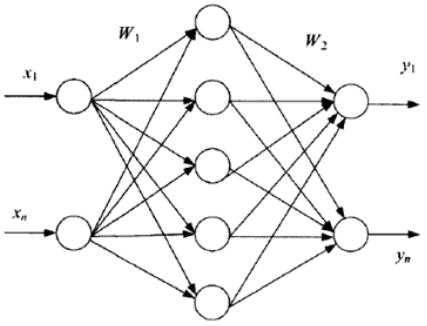


Figure 3. BP neural network model

4. Experiment Data

Through experiment analysis of variable fault bearing in subway operation and maintenance, part of experiment data was shown in table1.

TABLE 1. Part of original data

No fault	Inner Ring Fault	Outer ring fault	Roller fault
0.23406982421875	0.02838134765625	0.02044677734375	-0.0823974609375
0.1239013671875	-0.0177001953125	0.0152587890625	0.03204345703125
0.079345703125	-0.0201416015625	0.0042724609375	0.0390625
0.00823974609375	0.02716064453125	0.03753662109375	0.00244140625
0.015869140625	-0.0164794921875	-0.0347900390625	-0.029296875
-0.0103759765625	0.02197265625	-0.0042724609375	-0.030517578125
0.02166748046875	0.0244140625	0.01983642578125	0.02105712890625
0.03662109375	0.01129150390625	0.00732421875	-0.009765625
0.17608642578125	0.00762939453125	0	0.00457763671875
0.29083251953125	0.03631591796875	0.03753662109375	0.00640869140625
0.39459228515625	0.0775146484375	0.05584716796875	0.00335693359375
0.4217529296875	0.04425048828125	0.00091552734375	0.03082275390625
0.54412841796875	0.00946044921875	-0.0091552734375	0.0250244140625
0.60211181640625	0.01678466796875	0.00335693359375	0.02655029296875
0.64483642578125	-0.0341796875	0.01983642578125	0.01861572265625
0.63934326171875	0.01678466796875	0.0311279296875	0.01220703125

0.64849853515625	0.02227783203125	-0.0091552734375	0.0408935546875
0.526123046875	0.00701904296875	0.00701904296875	-0.0042724609375
0.41351318359375	-0.013427734375	-0.0732421875	-0.068359375
0.31951904296875	0.02471923828125	0.01312255859375	0.03509521484375
0.3155517578125	0.00396728515625	-0.00732421875	0.0872802734375
0.2886962890625	0	0.00579833984375	0.069580078125
0.23406982421875	0.02838134765625	-0.0152587890625	-0.0347900390625
0.1239013671875	-0.0177001953125	0.02716064453125	-0.0823974609375
0.079345703125	-0.0201416015625	0.02044677734375	0.03204345703125
0.00823974609375	0.02716064453125	0.0152587890625	0.0390625
0.015869140625	-0.0164794921875	0.0042724609375	0.00244140625
0.01007080078125	0.0219726562 5	0.03753662109375	-0.029296875
0.03570556640625	0.0244140625	-0.0347900390625	-0.030517578125
0.009765625	0.01129150390625	-0.0042724609375	0.02105712890625
0.02349853515625	0.00762939453125	0.0048828125	-0.009765625
0.0579833984375	0.00579833984375	0.00732421875	0.00457763671875
-0.0048828125	0.01617431640625	0.00274658203125	-0.001220703125

5. Analysis Process

Time-domain waveform after FIR noise reduction to no fault [9], rolling fault, outer ring fault was shown in Figure 4(a)-Figure 6(a). Figure 4(b)-Figure 6(b) was three layers db_1 orthogonal wavelet packet decomposition result.

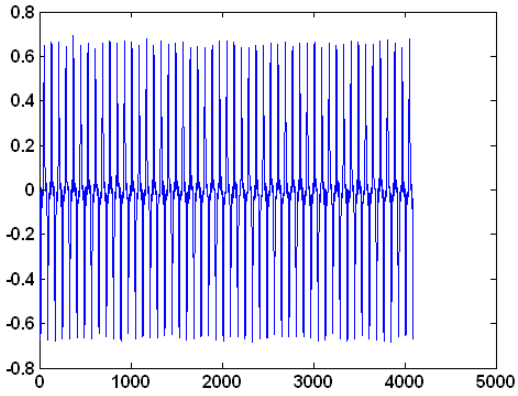


Figure 4 (a). Time domain waveform without fault

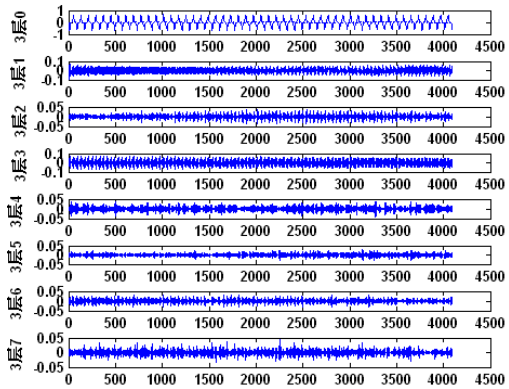


Figure 4 (b). Reconstruction waveform after no fault three layers wavelet packet decomposition

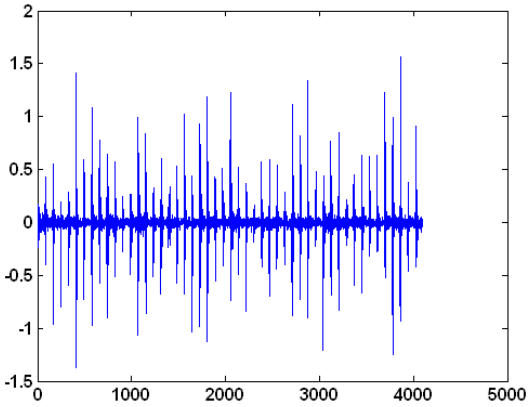


Figure 5 (a). Roller fault time domain waveform

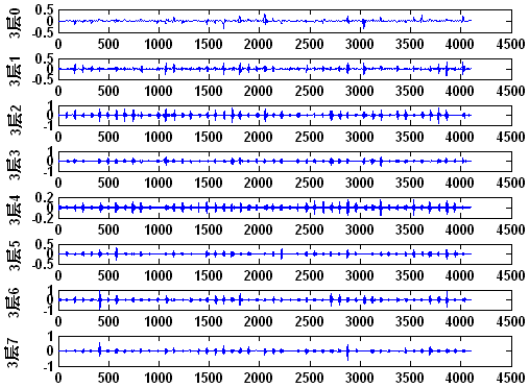


*Figure 5 (b).*Reconstruction waveform after roller fault three layers wavelet packet decomposition

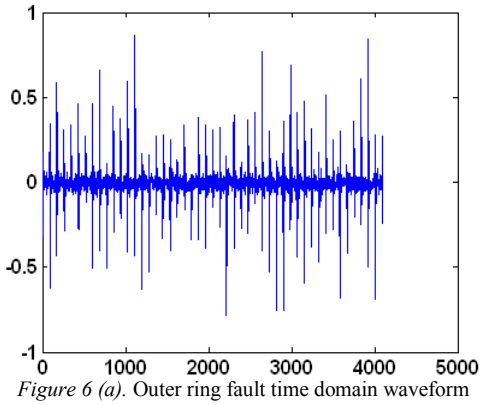


Figure 6 (a). Outer ring fault time domain waveform

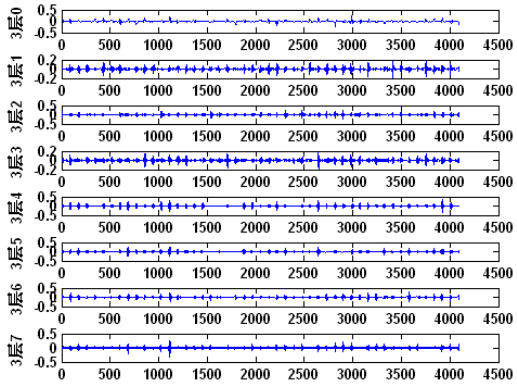


*Figure 6 (b).*Reconstruction waveform after outer ring fault three layers wavelet packet decomposition

5.1. Preparing Training Network

After three layers wavelet packet decomposed the collected data through MATLAB, we reconstructed the third layer decomposed data, evaluated the frequency band energy according to the former formula, normalized to the training samples are shown in table 2.

TABLE 2. Bearing operation sample data

Data number	Feature samples	Fault state	Fault vector
1	0.0690 0.0677 0.1777 0.0112 0.0568 0.0701 0.4876 0.0599	Outer ring fault	(0 0 1)
2	0.0142 0.0566 0.2915 0.0480 0.0194 0.0197 0.5110 0.0396	Outer ring fault	(0 0 1)
3	0.0483 0.0350 0.2448 0.0343 0.0485 0.0718 0.4637 0.0536	Outer ring fault	(0 0 1)
4	0.0531 0.0717 0.6778 0.0259 0.0241 0.0363 0.0456 0.0655	Rolling fault	(0 1 0)
5	0.1574 0.0585 0.4877 0.0940 0.0141 0.0563 0.0893 0.0427	Rolling fault	(0 1 0)
6	0.1931 0.0095 0.5605 0.0970 0.0254 0.0369 0.0321 0.0455	Rolling fault	(0 1 0)
7	0.0236 0.5035 0.8768 0.0205 0.0312 0.0003 0.5723 0.0125	Inner ring fault	(0 1 1)
8	0.0375 0.1283 0.6785 0.0342 0.0331 0.0127 0.6595 0.0326	Inner ring fault	(0 1 1)
9	0.0653 0.0293 0.5825 0.0667 0.1211 0.0678 0.6825 0.0122	Inner ring fault	(0 1 1)
10	0.9788 0.0078 0.0010 0.0087 0.0016 0.0006 0.0008 0.0007	No fault	(1 0 0)
11	0.9795 0.0078 0.0015 0.0088 0.0004 0.0004 0.0010 0.0006	No fault	(1 0 0)
12	0.9792 0.0079 0.0008 0.0091 0.0011 0.0005 0.0008 0.0006	No fault	(1 0 0)

5.2. Create Neural Network

The designed Elman network had eight input layer neurons, 3 output layer neurons and 13 hidden layer neurons. Hidden layer neuron transfer function using s-tangent function tansig, the output neuron transfer function using s-log function logsig. Using the previous data obtained training the network.

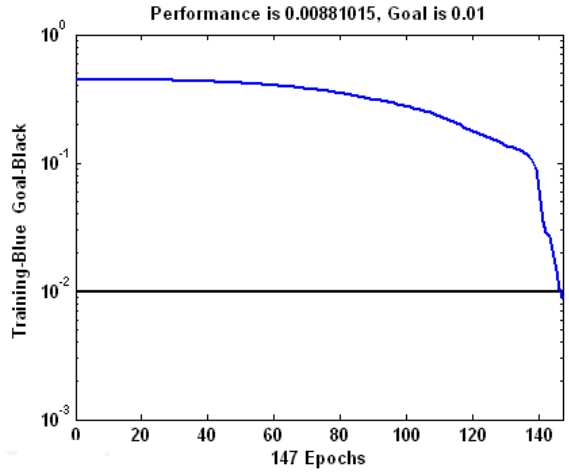


Figure 7. Elman network raining result

After 147 trainings, the network performance meets the application requirements, as shown in figure 7.

The designed BP network had eight input layer neurons, 3 output layer neurons and 14 hidden layer neurons. Hidden layer neuron transfer function using s-tangent function tansig, the output neuron transfer function using s-log function logsig. Using the previous data obtained training the network which is shown in figure 8.

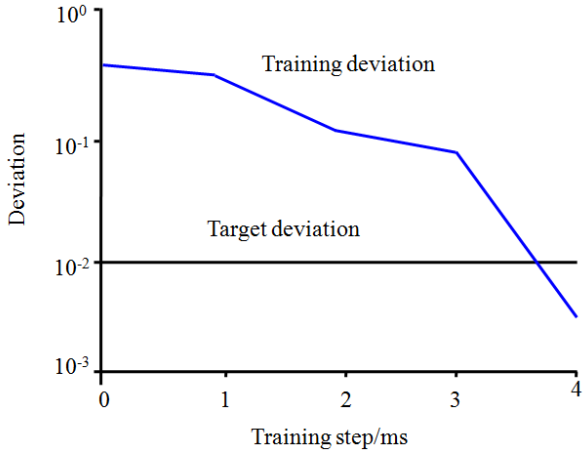


Figure 8. BP network training result

5.3. Test Neural Network

The test data was constructed with same approach to collected test data, as shown in table 3.

TABLE 3. Test data

Data number	Feature samples	Fault state
1	0.9722 0.0093 0.0042 0.0102 0.0004 0.0005 0.0024 0.0008	No fault
2	0.0841 0.0038 0.5236 0.0600 0.0669 0.0834 0.1039 0.0743	Rolling fault
3	0.0937 0.0744 0.1225 0.0927 0.4234 0.0498 0.0914 0.0521	Outer ring fault
4	0.0653 0.0293 0.5825 0.0667 0.1211 0.0678 0.6825 0.0122	Inner ring fault

Input the new test data to Elman network and got output as follows:

$$Y=\begin{bmatrix}0.9172 & 0.0151 & 0.0236\\ 0.0564 & 0.8231 & 0.0523\\ 0.0523 & 0.0723 & 0.8643\\ 0.0345 & 0.7829 & 0.8621\end{bmatrix}\rightarrow\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 1 & 1\end{bmatrix}$$

Input the collected data above to BP network and the output is shown as follows:

$$Y=\begin{bmatrix}0.9913 & 0.0145 & 0.0550\\ 0.0217 & 0.9952 & 0.0477\\ 0.0185 & 0.3278 & 0.9434\\ 0.0326 & 0.8327 & 0.9352\end{bmatrix}\rightarrow\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 1 & 1\end{bmatrix}$$

6. Discussion and Conclusion

From the result of output, we can conclude that the error between the actual output and the desired output is very small. Both Elman and BP neural network could recognize the fault state and fault type. We can completely achieve the intended purpose of identifying problems through this methodology.

Aiming at the characteristics of subway vehicle running system, this paper proposed to acquire vibration signals of subway vehicles running gear bearing through the sensors, combines wavelet analysis with Elman neural network to analyse the vibration velocity or acceleration signals for feature extraction, identifies fault types, realizes the purpose of safety assessment of running gear

bearing, and achieves remarkable results by actual tests. The method can be widely used in the safety assessment and fault diagnosis of rotating machinery such as bearing.

References

1. Li, X., Zhu, X.N., Cai, G.Q. (2010) Research on System Integration Alliance of Urban Rail Transit Safety Monitoring. *Journal of Convergence Information Technology*.10, 36-41
2. Nikolaou, N.G., Antoniadis, I.A. (2002) Rolling Element Bearing Fault Diagnosis Using Wavelet Packets. *NDT & E International*, 35(3), 197-205.
3. Li, X., Zhu, X.N., CAI, G.Q. (2010) A Novel Fault Diagnosis Expert System Knowledge Acquisition Method of Metro Vehicle Equipments. Proceeding of 2010 International Conference on Computer and Automation Engineering.
4. Guzmán, L.G., Gómez, A.S., Ardila, C.J., Jabba, D. (2013) Adaptation of the GRASP Algorithm to Solve a Multiobjective Problem Using the Pareto Concept. *International Journal of Artificial Intelligence*, 11(A13), 222-236.
5. Meng, X.L., Qin, Y. (2013) A New Fuzzy Linear Programming Model and Its Solving Method, *International Journal of Applied Mathematics and Statistics*, 49(19), 256-264
6. Guo, Q.J., Yu, H.B., Xu, A.D. (2005) Modified Moret Wavelet Neural Networks for Fault Detection. International Conference on Control and Automation, 1209-1214.
7. Lin, F.J., Shieh, H.J., Huang, P.K. (2006) Adaptive Wavelet Neural Network Control With Hysteresis Estimation for Piezo-Positioning Mechanism. *IEEE Transactions on Neural Networks*, 17 (2), 432-444.
8. Huang, N.E., Shen, S.P., et al. (2005) Hilbert-Huang Transform and Its Applications. Singapore: World Scientific.
9. Pang, P.L., Ding, G.B. (2008) Wavelet-based Diagnostic Model for Rotating Machinery Subject to Vibration Monitoring. Proceedings of the 27th Chinese Control Conference, 303–306.
10. Liu, H.M., Wang, S.P., Ouyang, P.C. (2009) Fault Diagnosis Based on Improved Elman Neural Network for a Hydraulic Servo System. Beijing Jiaotong University.
11. Ramos, G., J.López, J., Pueo, B. (2009) Cascaded warped-FIR and FIR filter structure for loudspeaker equalization with low computational cost requirements. *Digital Signal Processing*, 19, 393–409.
12. Yang, J.W., Cai, G.Q., Yao, D.C. (2010) Fault Diagnosis Method for the Rolling Bearing of Railway Vehicle Based on Wavelet Packet Transform and BP Neural Network. *China Railway Science*, 11, pp.68-73
13. Liu, L.P., Lin, F.T. (2008) Fault Diagnosis of Rolling Bearings Based on Wavelet Packet Energy Eigenvector and Neural Network. *Bearing*, 4, 659-660

Received on the 20th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 162-167
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

ENVIRONMENT COMPATIBILITY OF SERVICES INTRACTION AND ITS REACHABLE ANALYSIS

B. Chen[1,2,*], X. Huang[3]

[1]Service and Embedded Computing Key Lab of Ministry of Education, Tongji University, Shanghai, 201804, China
[2]College of Computer, Guangxi University of Science and Technology, Liuzhou,545006, China
[3]Libray of Guangxi University of Science and Technology, Liuzhou, 545006, China
[*] Corresponding author E-mail address: cb31@sina.com

The environment of services were not considered in most of analyses of interface interaction of composite services. In this paper, interface of interaction of services under specific environment are modelled as interface automata. The concept of environment compatibility of services is proposed. By giving the expression of criterion for compatibility checking and traversing the interaction model of composite services, reachable analysis for interaction environment compatibility checking can be carried. Some results of verifications are different from that of without environment and are accordance to actual situations.

Keywords: interface interaction, composite service, interface automaton, environment compatibility, and compatibility verification.

1. Introduction

A Service can combine with other atomic services or composite services through their interface interactions as an complex multifunctional software across organizations and platforms. However, combination mechanism also bring some risk and nondeterministic for composite services. Many scholars have modelled and analysed the behaviour of service interaction.

Bultan [1] regarded each peer participated in choreography as a finite automaton and studied the compatibility of interface interactions of two peers. Zhao [2] expressed the behaviour model of each participator in choreography with a special finite state machine named LOTOS. Cambronero [3] mapped the service interaction protocol WS-CDL to timed automata to express the behavioural of service. More scholars analysed interface interactions of composite web services with process algebra [4-7]. These studies are basically divided into two classes. One is the analysis of global interaction and answered whether one subservice in some interaction abided by the global agreement. Another were associated with local interactions and answered whether the interface interactions of two subservices is compatible?

This paper belongs to the latter. The difference is that the environment of interface interaction is introduced into the models and analysis. De Alfaro [8] discussed the interaction compatibility of two component under environment. His work focused on answering whether or not exists an environment, under which the interaction of two components is compatible. In this paper, we discuss the specific environment of composite web services and analyze their interaction compatibilities. It will been seen that if the interaction is compatible without considering the environment, the situation would be different.

2. Interface Automaton Model for Services and Interaction Environment

Interface automaton can be used to analyse the interface interaction of component system [8]. Emmi [9] noted that the system modelled by interface automaton has some advantage over the same system modelled by I/O automaton. Because some deadlock interaction in interface automaton system can be detected and that would not occur in I/O automaton system. We adopt interface automaton to model web service interaction for its characteristic is very similar to that of component system.

Definition 2.1 an interface automaton is a hex-tuple, $A=(S, I, Act^I, Act^O, Act^H, \delta)$, where

(1) S is the set of finite states.
(2) I is the set of initial states.
(3) Act^I is the set of input activities
(4) Act^O is the set of output activity symbols
(5) Act^H is the set of noninteractive but visible activity symbols

(6) $\delta \subseteq S \times (Act \cup \{\tau\}) \times S$ is a state transition relation, $Act = Act^I \cup Act^O \cup Act^H$

Specially, if $Act^I = Act^O = \varnothing$, the corresponding interface automaton A is called closed, otherwise, it is called open. If there exists a state transition, (s,a,s') $\in\delta$, then a is called enabled in states. An interface automaton can be used to model interface interaction of web service.

Definition2.2 let $A=(S_A, I_A, Act^I_A, Act^O_A, Act^H_A, \delta_A)$,$B=(S_B, I_B, Act^I_B, Act^O_B, Act^H_B, \delta_B)$ are two interface automata. They are composable if they are uncontradicted in signature, i.e. $Act^I_A \cap Act^I_B=\varnothing$, $Act^O_A \cap Act^O_B=\varnothing$, $Act^H_A \cap Act_B=\varnothing$, $Act^H_B \cap Act_A=\varnothing$.

Definition2.3 let A, B are two composable interface automata. Their composition is an interface automaton, $A \otimes B=(S, I, Act^I, Act^O, Act^H, \delta)$, where

(1) $S=S_A \times S_B, I=(I_A, I_B)$.

(2) $Shared(A,B)=Act_A \cap Act_B=(Act^I_A \cup Act^O_A \cup Act^H_A) \cap (ActI^I_B \cup Act^O_B \cup Act^H_B) = Act^I_A \cap Act^O_B \cup Act^O_A \cap Act^I_B$

(3) $Act^I = Act^I_A \cup Act^I_B \backslash Shared(A,B)$.

(4) $Act^O = Act^O_A \cup Act^O_B \backslash Shared(A,B)$.

(5) $Act^H = Act^H_A \cup Act^H_B \cup Shared(A,B)$.

(6) $\delta = \{((u, v),a,(u', v)) \mid u \in S_A \wedge u' \in S_A \wedge v \in S_B \wedge (u,a,u') \in \delta_A \wedge a \notin Shared(A,B)\} \cup$
$\{((u, v),a,(u, v')) \mid v \in S_B \wedge v' \in S_B \wedge u \in S_A \wedge (v,a,v') \in \delta_B \wedge a \notin Shared(A,B)\} \cup$
$\{((u, v),a,(u', v')) \mid u \in S_A \wedge u' \in S_A \wedge v \in S_B \wedge v' \in S_B \wedge (u,a,u') \in \delta_A \wedge (v,a,v') \in \delta_B \wedge$
$a \in (Shared(A,B) \cup \{\tau\})\}$.

Easily see that the composition above is associative and commutative.

Definition2.4 let $A_E=(S_E, I_E, Act^I_E, Act^O_E, Act^H_E, \delta_E)$, $A=(S_A, I_A, Act^I_A, Act^O_A, Act^H_A, \delta_A)$ are two interface automata. A_E is called the environment of A if A_E satisfies that

(1) A_E and A are composable in signature,i.e. $Act^I_E \cap Act^I_A = Act^O_E \cap Act^O_A = Act^H_E \cap Act_A = Act^H_A \cap Act_E = \Phi$

(2) A_E is not empty (means that its initial set is not empty.

(3) $A \otimes A_E$ is closed

The condition (3) means that interface automaton and its environment compose a closed system. Factually, BPEL4people standard supports role based people interaction. It is reasonable to consider the client activities as a subservice and the composite web service is closed.

3. The Example of the Interface of a Service

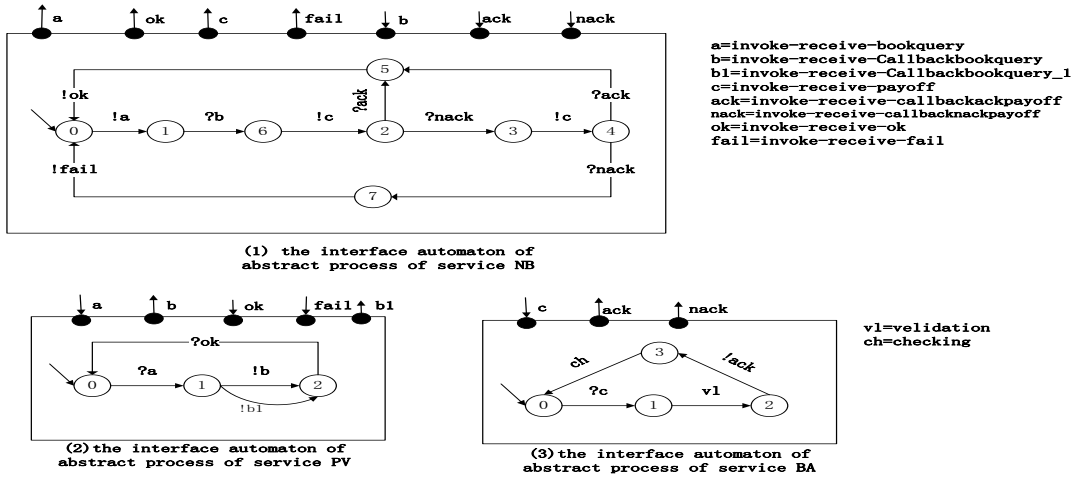


Figure 1. The interface automaton model of services in Online Book Service System

Online Book Service System: the system includes three services, Network Book Purchase, Service (NB for short), Book Provider Service (PV for short) and Online Bank (BA for short). For brevity, NB is considered as the combination of client and service interface and it represents client to interact with PV. Their interaction models can be described in Figure1.

4. The Compatibility of Interface Interaction

The compatibility definitions in existed works basically accorded with three situations in Bordeaux [10]. The compatibilities in this section are under specific environment and are distinguished with that without considering environment. The weak compatibility is just like the second definition in Bordeaux [10] under environment and the strong compatibility excludes one deadlock situation in the definition of Bordeaux [10].

Defition4.1 let A and B are two composable service interfaces. (u,v) is a reachable state of composite interface A⊗B. If there exists an activity symbol a∈shared(A,B) such that $a\in Act^{O}_{A}$ (u) (or $a\in Act^{O}_{B}(v)$) and $a\notin Act^{I}_{B}(v)$(or $a\notin Act^{I}_{A}(u)$), then (u,v) is called an irresponsive state of A⊗B.

The set of all irresponsive states is denoted as Irresponsive(A,B). Easily proof the following formulae.

$$\left\{(u,v)\in S_A\times S_B \mid \exists a\in shared(A,B).\begin{pmatrix} a\in Act^{O}_{A}(u)\wedge a\notin Act^{I}_{B}(v)\\ \vee\, a\in Act^{O}_{B}(v)\wedge a\notin Act^{I}_{A}(u)\end{pmatrix}\right\} \tag{1}$$

Irresponsive(A⊗B,C) =Irresponsive(A, B⊗C) (2)

Irresponsive(A,B⊗C)=

$$\left\{\begin{array}{l}(u,v,w)\in S_A\times S_B\times S_C \mid \\ \exists a\in Act^{O}_{B\otimes C}\wedge a\in Act^{I}_{A}.\begin{pmatrix} a\in Act^{O}_{B}(v)\wedge a\notin Act^{I}_{C}(w)\wedge a\notin Act^{I}_{A}(u)\\ \vee\\ a\in Act^{O}_{C}(w)\wedge a\notin Act^{I}_{B}(v)\wedge a\notin Act^{I}_{A}(u)\end{pmatrix}\\ \vee\\ \exists a\in Act^{I}_{B\otimes C}\wedge a\in Act^{O}_{A}.\left(a\in Act^{O}_{A}(u)\wedge a\notin Act^{I}_{B}(v)\wedge a\notin Act^{I}_{C}(w)\right)\\ \vee\\ \exists a\in shared(B,C)\wedge a\in Act^{O}_{A}.\begin{pmatrix} a\in Act^{O}_{A}(u)\wedge a\notin Act^{I}_{B}(v)\wedge a\notin Act^{I}_{C}(w)\\ \vee\\ a\in Act^{O}_{B}(v)\wedge a\notin Act^{I}_{C}(w)\wedge a\notin Act^{I}_{A}(u)\\ \vee\\ a\in Act^{O}_{C}(w)\wedge a\notin Act^{I}_{B}(v)\wedge a\notin Act^{I}_{A}(u)\end{pmatrix}\\ \vee\\ \exists a\in shared(B,C)\wedge a\in Act^{I}_{A}.\begin{pmatrix} a\in Act^{O}_{B}(v)\wedge a\notin Act^{I}_{C}(w)\wedge a\notin Act^{I}_{A}(u)\\ \vee\\ a\in Act^{O}_{C}(w)\wedge a\notin Act^{I}_{B}(v)\wedge a\notin Act^{I}_{A}(u)\end{pmatrix}\end{array}\right\} \tag{3}$$

Definition4.2 A and B are two composable service interfaces. (u,v) is a reachable state of composite interface A⊗B. If there exist two activity symbols a, b∈shared(A,B),a≠b such that $a\in Act^{I}_{A}(u)$, $b\in Act^{I}_{B}(v)$, $a\notin Act^{O}_{B}(v)$,$b\notin Act^{O}_{A}(u)$ satisfy that ∀c∈ (shared(A,B)\{a,b}), A and B cannot execute through co-operating c, then called the state (u,v) co-unrequested state of composite interface A⊗B.

Co-Unrequested(A,B)=

$$\left\{\begin{array}{l}(u,v)\in S_A\times S_B \mid \exists a,b\in shared(A,B)\wedge a\neq b\wedge\begin{pmatrix} a\in Act^{I}_{A}(u)\wedge a\notin Act^{O}_{B}(v)\\ \wedge\, b\in Act^{I}_{B}(v)\wedge b\notin Act^{O}_{A}(u)\end{pmatrix}\\ \wedge\,\forall c\in Act_A(u)\setminus\{a,b\}\wedge \bar{c}\notin Act_B(v)\end{array}\right\} \tag{4}$$

The symbol $c\in Act_A(u)$, $\hat{c}\notin Act_B(v)$ means that if c is the activity symbol in state u of A, then $\hat{c}$ is the corresponding reverse activity in state v of B (the reverse of an input activity is the corresponding output activity and vice versa.)

Co-Unrequested(A ⊗ B,C)=Co-Unrequested(A,B ⊗ C) (5)

Co-Unrequested(A,B⊗C)=

$$\left\{\begin{array}{l}(u,v,w)\in S_A\times S_B\times S_C \mid \\ \exists a,b \left(\begin{array}{l} a,b\in shared(A,B\otimes C)\wedge a\neq b. \left(\begin{array}{l} a\in Act_A^I(u)\wedge\neg Act_B^O(v)\wedge\neg Act_C^O(w) \\ \wedge \\ b\in(Act_B^I(v)\wedge\neg Act_C^O(w)\wedge\neg Act_A^O(u) \\ \vee\ Act_C^I(w)\wedge\neg Act_B^O(v)\wedge\neg Act_A^O(u)) \end{array}\right) \\ \vee \\ a,b\in shared(B,C)\wedge a\neq b. \left(\begin{array}{l} a\in Act_B^I(v)\wedge\neg Act_C^O(w)\wedge\neg Act_A^O(u) \\ \wedge \\ b\in Act_C^I(w)\wedge\neg Act_B^O(v)\wedge\neg Act_A^O(u) \end{array}\right) \end{array}\right) \\ \wedge \\ \forall c\in Act_A(u)\setminus\{a,b\}\wedge \bar{c}\notin Act_B(v)\wedge \bar{c}\notin Act_C(w) \end{array}\right\} \quad (6)$$

Definition 4.3 let $A_W=A_E\otimes A_1\otimes A_2$, an interface of composite service, where A_1 and A_2 are two service interfaces. A_E is an environment of them. If they satisfy following condition (1) and (2), then $A_1$and A_2 is called environment weak compatible. Written as $A\approx_E B$. If they satisfy all condition, then A_1 and A_2 under environment A_E is called environment strong compatible, written as $A\approxeq_E B$.

(1) the interface of composite service, A_W, is closed, namely $Act^I_W=Act^O_W=\varnothing$.

(2) there is no any reachable irresponsive state in composite interface interaction, namely, irrespon_sive(A_W)=∅.

(3) there is no any reachable co-unrequested state in composite service interface, i.e. Unrequested(A_W) = ∅.

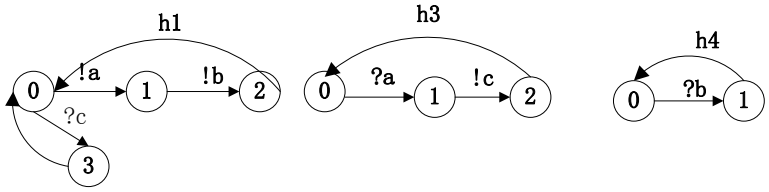


Figure2. Service A and B with co-unrequested state and environment C

Three interfaces are illustrated in Figure2. The composite of service A and B is not closed. The interface of service C is their environment. Without considering environment, A and B are incompatible because there is a co-unrequested state. After considering the environment C, the co-unrequested state is unreachable and the interface interaction of A and B is compatible under environment C.

Definition4.4 let A_E is the environment of service A. If Irresponsive(A_E,A)=∅, then the interface interaction of A and its environment A_E is called self environment weak compatible, written as $A\approx_E A_E$.

Definition4.5 let A_E is the environment of service A. If Irresponsive(A_E,A)=∅ and Co-Unrequested (A_E,A)=∅, then the interface interaction of A and its environment A_E is called self environment strong compatible, written as $\mathbf{A\approxeq_E A_E}$.

5. Compatible Checking for Interface Interaction

The compatible checking for interface interaction of services can be carried out by traversing the interface of composite service A_W and judging whether there is irresponsive state and co-unrequested state. The algorithm traverses very state of composite service interface automaton and judges whether there is irresponsive state and co-unrequested state according formulae (3) and (6). For space limited, only the procedure for judging if a state is or not an irresponsive state has been listed above, i.e. checking_irresponsive (s). The sub procedure checking_co-unrequested is similar. The number of the former sub-procedure loops is the total of activity labels in $A_E\otimes(A_1\otimes A_2)$. The number of the latter sub-

procedure loops depends its two loop variables. Suppose the N is activity labels in $A_E \otimes (A_1 \otimes A_2)$. The time complexity of two sub-procedures is O(9N) and $O(4N^2)$ respectively. Where 9 and 4 is most number of judges in one loop of two sub-procedures respectively. ErrorFound procedure traverse $A_E \otimes (A_1 \otimes A_2)$ in depth-first fashion. The time complexity if $O(n^2)$, where n is the number of nodes in graph.

Algorithm 5.1(environment strong compatible checking):

IsCompatible_Envi

Input: $A_i=(S_i,I_i,Acit^I_i,Act^O_i,Act^H_i,\delta_i)$,i=1,2 are two service interfaces,$A_E=(S_E,I_E, Acit^I_E, Act^O_E, Act^H_E,\delta_E)$ is the

environment. $A=A_E \otimes (A1 \otimes A2)$ is the composition of service interface.

Output:"yes" if A contains a irresponsive state or unrequested state, otherwise "no".

```
{  Set of states R:= ∅;                                  // the set of visited states in first outer DFS.
   Stack of states:U=ε;                                  / /the stack for first outer DFS.
   Boolean error_found:=false; irresponsive:=false;
   unrequested:=false; State I:=(I_E, I_1,I_2);
   State s←I;
   errorfound(s);
   if error_found  return ("yes")
  else return("no");
}

Procedure Boolean Errorfound (state s)
{ push(s,U);   R←{s};
  repeat {
      s'← top(U);
      if  Post(s') \ R  then {          //the successor of s' not been visited exists
           let s"□Post(s') \ R;
           push(s",U);                  // push the unvisited successor of s' and mark it reachable
           R←{s"};
      }
      else{
           pop(U);                      // DFS finished for s'
           checking_irresponsive(s');
           if irresponsive==false  checking_unrequested(s');
           error_found←irresponsive∨unrequested;
      }
  }until ((U ==ε)∨error_found)
  return(error_found);
}

procedure Boolean Checking_irresponsive( state s)
          Input:   shared(A_E,A_1⊗A_2); shared(A_1,A_2); Act; state s.
          Output:  the  irresponsive  value of state s.
Step1:  activity label a; boolean irresponsive:=false;
        Act←Act_1∪Act_2∪Act_E;                  //initialize variables.
Step 2: if (irresponsive==fasle∧Act≠∅)     Let a∈Act;

        if (a∈shared(A_E,A_1⊗A_2) ) {
            if (a∈Act^O_A1⊗A_2∧Act^I_AE){
                if a□Act^O_A1(v)∧a□Act^I_A2(w)∧a□Act^I_AE(u)     irresponsive←true;
                if a□Act^O_A2(w)∧a□Act^I_A1(v)∧a□Act^I_AE(u)     irresponsive←true;
             }
            if (a□Act^I_A1⊗A_2∧a□Act^O_AE )
                if a □Act^O_AE(u) ∧a□Act^I_A1(v)∧a□Act^I_A2(w)  irresponsive←true;
         }
         if (a□shared(A_1,A_2)∧a□Act^O_AE ){
            if a□Act^O_AE(u)∧a□Act^I_A1(v)∧a□Act^I_A2(w) irresponsive←true;
            if a□Act^O_A1(v)∧a□Act^I_A2(w)∧a□Act^I_AE(u)   irresponsive←true;
            if a□Act^O_A2(w)∧a□Act^I_A1(v)∧a□Act^I_AE(u)   irresponsive←true;
         }
```

if (a□shared(A_1,A_2)∧a□Act^I_{AE}) {
if a□Act^O_{A1}(v)∧a□Act^I_{A2}(w)∧a□Act^I_{AE}(u) irresponsive←true;
if a□Act^O_{A2}(w)∧a□Act^I_{A1}(v)∧a□Act^I_{AE}(u) irresponsive←true;
}

Step 3: if (irresponsive =true∧Act≠∅) Act←Act\{a}; go to step 2;

Step 4: if (irresponsive =false∧Act=∅) return (irresponsive).

Without considering the composite environment, service NB and PV in Online Book Service System in section 3 are incompatible for the state 7 is a reachable irresponsive state. By introducing the environment BA, the reachable state 7 in their composite interface interaction model is limited as a unreachable state. Thus their interface interaction is environment strong compatible.

6. Conclusion

The interface automaton is applied to model composite web service interface interaction in this paper. By introducing the environment factor, the environment strong and weak compatibilities of service interaction under specific environment are proposed. We think this consideration is accord with factual service situation. After given the criterion of judge for irresponsive state and co-unrequested state, we can check the environment compatibility through reachable analysis. Future work will focus on global interaction analysis.

Acknowledgment

This research is supported by Open Project Program of Service and Embedded Computing Key Lab of Ministry of Education in Tongji University (No.2011-2), Research Program of Education Department of Guangxi (No. 201203YB124), The Program of Doctoral Fund of Guangxi University of Science and Technology (11Z05).

References

1. Bultan, T., Su, J.W., Fu, X. (2006) Analysing conversations of Web services. *IEEE Internet Computing*, 10(1), 18-25.
2. Zhao, H., Wang, W., et al. (2012) Research on Formal Modelling and Verification of BPEL-based Web Service Composition. In Proceedings of 2012 IEEE/ACIS International Conference on Computer and Information Science,Shanghai,China, 30 May-1 June 2012, (631-636)
3. Cambronero, M.E., Díaz, G., Valero, V., Martínez, E. (2011) Validation and verification of Web services choreographies by using timed automata, *The Journal of Logic and Algebraic Programming*, 1(80), 25-49
4. Carbone, M., Honda, K., Yoshida, N., Milner, R., Brown, G., Ross-Talbot, S. (2006) A theoretical basis of communication centred concurrent programming. (http://www.w3.org/2002/ wschor/edcopies/ theory/note.pdf).
5. Yeung, W.L. (2006) Mapping WS-CDL and BPEL into CSP for behavioural specification and verification of Web services. The Proceedings of the European Conference on Web Services, 297-305.
6. Liu, J., Wang, J., He, K. (2013) Using Pi-calculus to Model Web Service Interaction, *Journal of Computational Information Systems*, 9(5), 1759-1767.
7. Busi, N., Gorrieri, R., Guidi, C., Lucchi, R., Zavattaro, G. (2005) Choreography and orchestration: A synergic approach for system design, *LNCS*, 3826, 228-240.
8. de Alfaro, L., Henzinger, T.A. (2001) Interface automata. *ACM SIGSOFT Software Engineering Notes*, 26(5), 109-120.
9. Emmi, M., Giannakopoulou, D., Päsäreanu, C.S. (2008) Assume-guarantee verification for interface automata. *LNCS*, 5014, 116-131.
10. Bordeaux, L., Salaün, G., et al. (2005) When are two Web services compatible? *LNCS*, 3324, 15-28.

Received on the 10th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 168-173
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

THE RETARGETING TECHNOLOGY OF 3D MODEL BASED ON TOPOLOGICAL STRUCTURE

W. De-Cheng [1,*], Ch. Yan [1]

[1]School of Information Engineering, Anhui Xinhua University,
Hefei, Anhui 230088, China
[*] Corresponding author E-mail address: ding_xing@163.com

Retargeting technology is the key technology of reuse of motion capture data. This paper analyses the specific application of retargeting for three different topology structure target models and proposes improved algorithm for the target models with significantly different topology structures so that data reusability is increased.

Keywords: retargeting; motion capture; topology structure

1. Introduction

With the rapid development of the motion capture technology, it is very easy to obtain 3D human motion data, and motion capture data has been successfully used in computer animation, movies and many other fields. However, since motion capture is acquisition in specific scenes and for specific actors, motion capture equipment is expensive and data storage format is diverse, all has made reuse of motion capture data inconvenient. Thus, how to analyse and reuse the existing motion capture data has become one of the focuses of the research of current motion capture technology.

2. Human Skeleton Model

As shown in Figure 1, Human capture data describes the human body joints and skeletons and their connections by hierarchical skeleton model [1]. A motion consists of skeleton topology information and motion information. Skeleton topology information includes joint topology connection relationship and the offset of joints in the parent joint point local coordinate system. Motion information includes global translation and rotational translation of the root joint in each frame as well as the rotational components of other joints around the parent joint. Accordingly, the motion data F_i of each frame can be expressed as:

$$F_i = \left(p_i^{(1)}, r_i^{(1)}, r_i^{(2)}, ..., r_i^{(n)}\right) \tag{1}$$

Wherein, N represents the number of joints, $p_i^{(1)}$ means that the a Root joints in position in the world coordinate system, $r_i^{(1)}$ means that the orientation of Root in the world coordinate system, and $r_i^{(n)}$ indicates the rotation of the n-th joint relative to its parent joint.

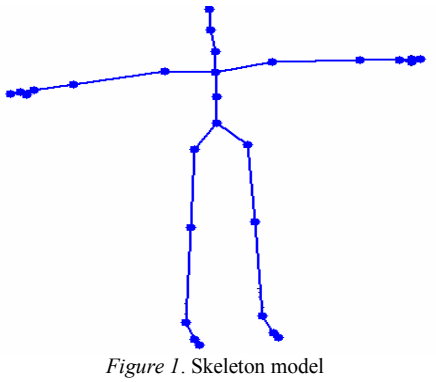


Figure 1. Skeleton model

3. Motion Retargeting

Motion retargeting is to modify the motion data captured from the original role, and map it to new role different from the original role to form motion of the characteristics of the new role. Motion retargeting technology mainly studies: 1) how to effectively apply collected original data to different role models; and 2) constraints for motion characteristics to be in line with the new role.

If the new role has the same topology structure and limb length as that of the original role, data mapping can be done directly. However, new role generally has different limb length and even different topology structure. Thus, direct mapping of the captured data to new role can easily lead to motion distortion and errors such as penetration of feet into the ground or off the ground. The greater the difference of limb length and proportion between the two role models is, the more obvious the phenomenon contrary to the original motion characteristics is. Thus, before application, data retargeting processing must be done in order to make the reuse of the data possible and even to form some actions and special motion that cannot be captured, thereby expanding the applicable scope of the captured data and further improving data reusability.

4. Retargeting According to Target Model

Currently, various motion capture systems record the trajectory of the human body and convert it into abstract motion data. Motion retargeting target model can be divided into three categories in accordance with the difference from the human body model: with the same topology structure, similar topology structure and different topology structure. For different type of target model, retargeting processing method is different.

4.1. With the Same Topology Structure

For the target model with the same topology structure as human body model, retargeting technology first solves the problem of the alignment of motion between human body and target model, and then maps the captured the human motion information expressed in translation and rotation vector to the target model. Generally, the limb length of target model is different from that of human body model, which will result in foot penetrating of the ground or off the ground and so on of the target model in case of motion. As shown in Figure 2, the left shows the motion sequences of the captured human skeleton model and the right shows the motion sequences by directly redirecting the motion information to the human body with short skeleton length.

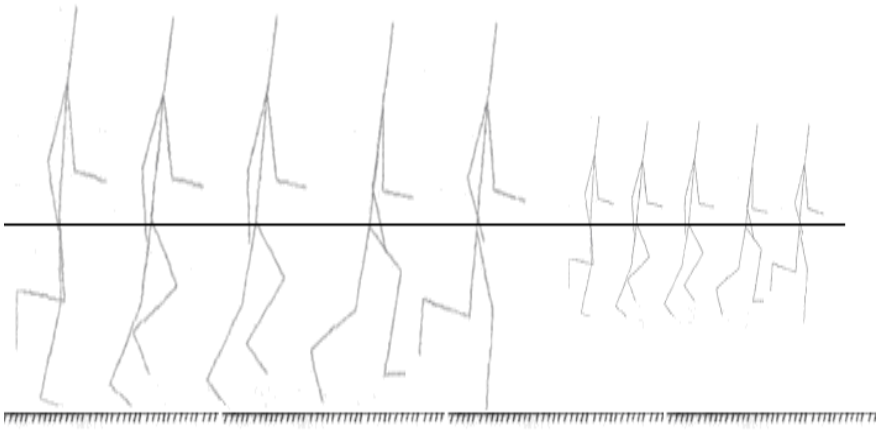


Figure 2. Feet off the ground

When redirecting among models with the same topology structure and proportional limb length, you can adjust the central position of the skeletons according to the proportion of the models and set the end effector in the feet and operate the entire joint chain by inverse kinematics so as to finally solve the entire motion information of retargeting target.

4.1.1. Adjustment of the Centre of the Skeletons

The centre of the skeletons of the human body model in initial frame can be adjusted in accordance with the following equation.

$$P_y^r = P_y^o \times \alpha \tag{2}$$

Where in, p_y^0 is the height of the centre of the skeletons of the human body model, and p_y^r is the height of the centre of the skeletons of the target model.

$\alpha \sum_{k=1}^{n} l_k^r / \sum_{k=1}^{n} l_k^0$ is the ratio of the total length of lower extremities of the human body model and the target model and is known as the pantograph ratio, wherein l_k^0 is the length the k-th segment of the lower extremity of the human body model, l_k^r is the length of the k-th segment of the lower extremity of target model, and n is the total number of segments of the lower extremity.

The latter centre position of the skeleton of the target model of each frame can be calculated by the position of the centre of the skeleton of the previous frame plus the displacement d^r of the centre of the skeleton of target model; d^0 can be obtained by loading motion data and calculation, d^r can be obtained by multiplying the displacement d^0 of the center of the skeleton of the human body model by pantograph ratio as the following formula:

$$d^r = d^\circ \times \alpha . \tag{3}$$

4.1.2. Setting of the Terminal Effector

The lower extremities of the skeleton of the human body model are zoomed according to pantograph ratio α to obtain a virtual skeleton, and set the position of this virtual skeleton terminal effector as mandatory constraints of that of the target model so as to make the terminal effector position of target model above the ground. The displacement of the terminal effector of the target model is able to cope with the displacement of the centre of the skeleton so as to achieve the requirement of the feet on the ground at motion.

4.1.3. Reverse Kinematics Solution

According to the new position of the terminal effector of the target model, the motion of the target model can be solved, and finally the centre of the skeleton of the target model can be adjust according to formula (2) and (3) to obtain the same height as terminal effector of human body model. The terminal effector of lower extremities is the feet. Thus, as long as the captured motion does not penetrate the ground or get off the ground, it can ensure that the target model will not suffer such errors.

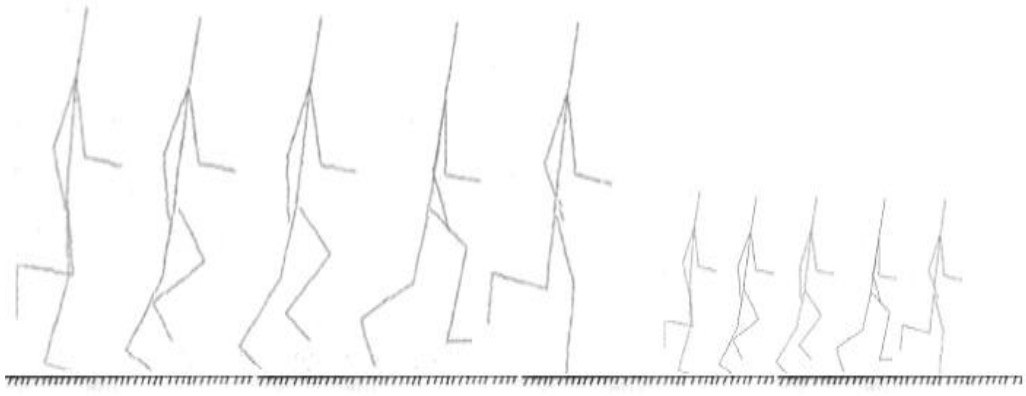


Figure 3. Processed results

4.2. With Similar Topology Structure

It is difficult to directly map the retargeting of the motion from the human body model to the target model with different topology structure because of inconsistency in the joint number, degree of freedom, length and coordinate system. Direct mapping without adjustment will cause unexpected results and the obtained motion can not maintain the original characteristics.

In order to redirect between the human body model with different topology structure and the target model, it is most likely to think of aligning their skeletons, then making their initial posture consistent and finally mapping the motion from the human body model to the target model.

4.2.1. Alignment of Skeleton

For alignment of the skeleton, the skeletons of human body model and target model need to be transformed. Usually, the human body model is simplified to be similar to the skeleton of the target model. Pullen [2] et al have pointed out that the freedom of motion of the human body has redundancy, simplifying the human body model can retain the freedom of motion as little as possible to reduce the complexity of solving. Popovic [3] et al obtain simplified human body model by human-computer interaction settings and deletion of joint freedom with little effect on the human body motion. Pollard [4] uses a simple mass-spring model as the similar human motion in order to achieve the motion retargeting.

In this paper, a simplified model is used in the experiment. Simplification of the model is completed by manual deletion of surplus degree of freedom of motion in advance, and the specific simplification operation can refer to the principle of simplification kinematics proposed by Pullen, including the deletion of degree of freedom, deletion of sub-nodes and symmetrical deletion and so on.

There are mainly three types of alignment between the skeletons: 1:1 alignment relationship, N:1 alignment relationship and no alignment (i.e. ignoring certain joints).

4.2.2. Initial Posture Alignment

Initial alignment mainly calculates the correction matrix between corresponding joints of two joint hierarchical models when the initial joint angle is zero in order to maintain consistent rotation direction between the joints at the time of mapping motion data. It recursively calculates the rotation angle between corresponding two skeletons from the root joint, and directly applies the calculation results to the target skeleton so as to enable the skeleton of the target model and the skeleton of human body model to be in the same rotation direction.

4.2.3. Motion Data Mapping

In spite of similar initial posture, there may be difference in local places between two different skeletons. For instance, direct mapping of the human hands and the claws of the bear may result in great difference between the retargeting results and the original motion. Take the i-th joint of human body model at the moment t as the example, if the Euler angle of this joint at this moment is set as $A = (\alpha, \beta, \gamma)$ and Euler angle rotation sequence is *ZYX* and rotation sequence of the Euler angle of the joint corresponding to the joint in the target model is also defined as *ZYX*, the corresponding Euler angles of the joints of the target model are calculated as follows:

1) Calculation of rotation matrix of the joint of human body model at the moment t:

$$R_i(t) = Rz(\gamma) \times Rr(\beta) \times Rx(\alpha) \tag{4}$$

2) Calculation of the rotation matrix of corresponding joints of target model

If the initial correction matrix of this joint of the target model is set as RM, the rotation matrix of corresponding joints mapped to the target model is calculated as follows:

$$R_i'(t) = R_M \times R_i(t) \tag{5}$$

3) Reverse of Euler angle

Reverse of Euler angle according to Equation 3, set:

$$R_i^{'}(t)=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{6}$$

If corresponding joints of the target model rotate from the initial position around their axis *Z*, *Y* and *X* successively by θ, φ and ψ and Euler angles coincide with $R_i^{'}(t)$, Euler angles can be calculated as follows:

$$\theta = \arctan(a_{12}/a_{11}) \tag{7}$$

$$\phi \arctan\left(\frac{-a_{31}}{a_{11}\cos\theta + a_{21}\sin\theta}\right) \tag{8}$$

$$\psi = \arctan\left(\frac{a_{13}\sin\theta - a_{23}\cos\theta}{a_{22}\cos\theta - a_{12}\sin\theta}\right) \tag{9}$$

The experimental results show that the motion in this method is some different from the real motion that the target model should have. It is not very realistic and needs to be further improved.

4.3. With Different Topology Structure

In Cartoons and 3D games, we often see some objects with human body motion characteristics, such as tables, chairs and fruits. Take table dance for example, the anthropomorphic effects need animators to draw each frame or key frame image by hand, with low efficiency. In fact, we can use the captured human motion data and map it to these objects to produce special effect that is difficult to see in the real world.

However, data obtained by motion capture is substantially motion data of human body model. The topology structures of the tables, chairs and fruits are greatly different from that of human body model and the motion mapping cannot be done in above method.

Idea 1: according to the characteristics of the target model, install the virtual limbs and set it as invisible to simulate the topology structure of human body, and then map the motion information of the human body model to the anthropomorphic target model, thereby generating cartoon model motion.

Idea 2: The human skeleton model is scaled and placed into the inside of the target model. For example, a human skeleton is placed in a banana so that the human skeleton is completely contained in the banana, and then the banana motion is driven by the skeleton to generate anthropomorphic effect. However, in case of great motion of human body, the problem is how to drive the model surface deformation in order to avoid the target model from the loss of the original shape due to deformation.

In order to achieve this retargeting, IIya Baran [5] has proposed a new idea: generating a skeleton model according to the users' choice, then use the skeleton to guide the segmentation of the three-dimensional model for conversion into similar or the same topology structure, and finally completing the animation directly in retargeting algorithm and real-time rendering algorithm.

In this paper, an improved model segmentation algorithm is used to guide the segmentation of the target model by this skeleton. Prior to execution of the algorithm, the user should first complete definition or importing of skeleton and drop the joint point to the appropriate location. The algorithm is as follows:

1) Calculate the distance D_i from each vertex to all the skeleton line segments, wherein it does not refer to the distance to the straight lines but refers to the distance to the line segments. When the pedal is outside of the line segment, the smaller value of the distance from P to the two end points should be regarded as the value of D_i.

2) Record the minimum value Dmin1 and the second smallest value Dmin2 for each vertex P and the serial number J1 and J2 of the corresponding skeleton (either end point of skeleton line segment represents one end of the parent), which means that the vertex P is affected by J1 motion matrix at the

probability of $\omega_1 = D_{\min} / (D_{\min 1} + D_{\min 2})$ and by J2 motion matrix at the probability of $\omega_2 = D_{\min 2} / (D_{\min 1} + D_{\min 2})$.

3) Take the adjustable parameters $\alpha \in (0,1)$, $\beta_1 > 0$, $\beta_2 > 0$ and then the formula for calculating the new position of the vertex P of is as follows:

$$P' = \left[(1-\alpha) * I + \alpha * \left(w_1 * e^{-D_1+\beta_1} * M_1 + w_2 * e^{-D_2+\beta_2} * M_2\right)\right] * P \quad (10)$$

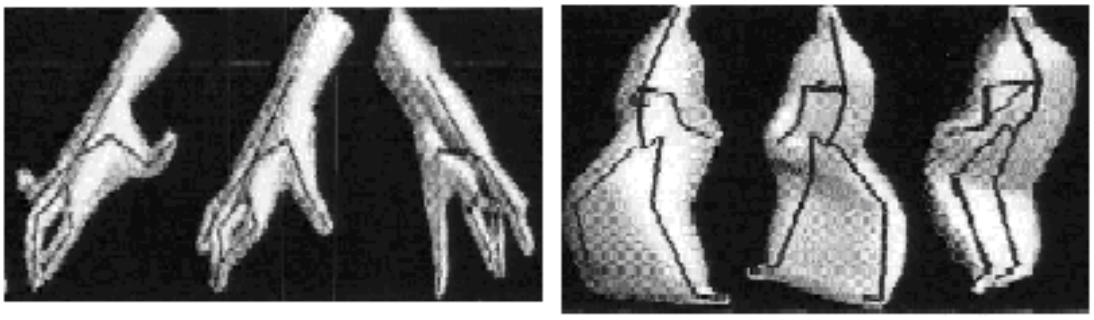


Figure 4. Anthropomorphic effect

5. Conclusion

This paper improves the motion mapping algorithm between the models with different topology structures to effectively improve the reusability of the data. The future work will focus on how to quickly generate realistic 3D model motion.

Acknowledgment

This work is supported by the Natural Science Foundation of the Higher Education Institutions of Anhui Province, China (Grant No. KJ2011B070).

References

1. Xu Jie, *Research of 3d cartoon character animation generation method combined with motion capture data*, Zhejiang University, 2003
2. Tourangeau, R. and Yan, T. Sensitive questions in surveys, *Psychological Bulletin*, Vol.133, No.5, 2007, pp. 859-883.
3. Pullen K, Bregler C. Motion capture assisted animation: Texturing and synthesis. *Computer Graphics Proceedings, Annual Conference Series, ACM SIGGRAPH, San Antonio, Texas*, 2002, pp. 501-508.
4. Popovic' Z., Witkin A. Physically based motion transformation. *Computer Graphics Proceedings, Annual Conference Series, ACMSIGGRAPH, Los Angeles, California*, 1999, pp. 11-20.
5. Pollard N. Simple machines for scaling human motion. *Proceedings of Eurographics Workshop on Animation and Simulation*, Milan, 1999, pp. 3-11.
6. Baran I., Popovic J. Automatic Rigging and insurance Animation of 3D characters. *San Diego: ACM SICCRAPH Conference*, 2007.

Received on the 10th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 174-184
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

A COGNITIVE TOPOLOGY STRUCTURE OBSERVATION APPROACH BASED ON MAN-MACHINE COUPLING STATES

Sh. Yang [1], W. Yu [1], Ch. Liu [2], R. Cheng [1], L.Guo [1], Sh. Meng [3]

[1]Department of Mathematics and Computer Science
Wuyi University, Wuyishan, Fujian, China, 354300
[2]School of Modern Information Service, Wuyi University
Wuyishan, Fujian, China, 354300
[3]Orient Potential Software Technology Co., Ltd, B•CSM
Wuyishan, Fujian China, 354300

The cognitive structure is the basis for learning and there is no doubt about its Objectivity. Due to its existence in the brain, detecting the cognitive structure can be very difficult. How to collect cognitive information and reconstruct the cognitive structure are the scientific issues worthy of studying. Based on the various cognitions such as the situated cognition, the distributed cognition, the extended cognition and so on, the man-machine cognitive coupling environment is built and the inner-brain cognition is induced to the outer-brain cognitive context by the Agent. In order to record the process of man-machine interaction process and obtain the cognitive sequence of symbols, a vector situation space is designed and the sensors for cognitive operations are deployed. What's more, the manifold reconstructed cognitions such as the topology and networks are built. From the perspective of cognitive observation, the man-machine interaction can be divided into "Brain cognitive body – cognitive coupling Situations - information Manifolds" (BSM) and becomes the observation platform for cognitive structures. The cognition observation approach based on BSM is called "coupling observation approach for cognitive structures". This method involves the duality between the cognition and the cognitive objects in the information space, the covariance of the cognitive patterns, the effectiveness of the cognitive manifold frame, the completeness of cognitive symbol sequences, the scientificalness of the restructured cognitive manifold spatial structure and etc. BMS can observe the cognition process more accurately and obtain quantities of real-time cognitive information. Consequently, it can provide support in the cognitive dynamic system construction, the cognitive structure imagery, the empirical education, etc.

Keywords: Man-machine Interaction, Agent, Cognitive Structure, Manifold, Category, Empirical Education

1. Introduction

The schema and cognitive structure are the basic concepts in pedagogic psychology within the cognitive science [1]. Piaget established the cognitive mathematical models by introducing mathematical concepts such as Category, Morphism, Group Lattice and so on [2]. However, similar to the research paradigms in psychology, pedagogy, etc., these studies end up with the same weaknesses such as inadequate precision on the data acquisition, the difficulties in the description of dynamic process, the loss in the operational feasibility etc. [3-4]. Therefore, it is a significant trend to introduce the empirical research paradigms in natural sciences based on observation and data.

Based on research paradigms of natural science, if the cognitive structure does exist objectively, then questions rises up like how to observe, quantify and visualize the cognitive structure and what kind of scientific question it is to describe the cognitive structure in a mathematical way. The empirical study of the cognitive structure involves theoretical design, scientific hypotheses, data collection, experimental verification etc. Among them, the tool selection and improvement and the cognitive structure observation are important. For example, after Galileo Galilei improved the telescopes in 1610, he published his observed results from the accurate observation about the satellites of Venus and Jupiter. He overthrew Aristotle's mechanics system, which was the foundation in mechanics studies by then and established the modern mechanics theory by carrying out theoretical analysis and experiments. Tools, methods and technologies for the observations of cognitive structures are the keys and basis. By using the Bayesian Network Approach [5], Structural Equation Modelling Approach and other mathematical methods, the cognitive structure can be further studied. What is more, the psychological information acquired through the man-machine interaction has gradually become a research focus [7-9].

The empirical approach is a new trend within education development and it proposes ideas like developing high qualified and efficient educational methods and the Precision Teaching (PT) [10]. It encourages teachers to work based on more data and tools like doctors.

In general, the brain is the carrier for cognitive structure, and therefore, the first challenge is how to observe the cognitive structure within the brain. Currently, the research methods for cognitive structures in psychology and pedagogy are mainly through questionnaire, psychological laboratory

equipment, observation, introspection etc.

In natural sciences such as physics, observation is closely related to tools. Nowadays, computer is undoubtedly the most widely used tool in the world. Furthermore, in the field of cognition and pedagogy, it is also used widely for data analysis or conducting the online survey to replace the traditional one. Is it possible to observe the cognitive structure through computers? As the bond between the brain cognitive space and the information space, the man-machine interaction is a complete new cognitive environment.

The research of man-machine interactions can be carried out on the level of philosophy, mathematics and technology. Specifically, the situated cognition, distributed cognition, embodied cognition, extended cognition and other cognitions in the Cognitive philosophy [11] provide a new prospective for the outer-measurement of the brain cognition, i.e. the inner-brain cognitions are coupled to the outer-brain environment for measurement. This indicates the development roadmap of Cognitive Philosophy from the disembodied cognition, the embodied cognition to the extended cognition.

Based on the outer-brain cognition view, the man-machine interaction learning system is studied and the cognitive coupling learning environment is established, furthermore, the inner cognitions are induced to the digital space using the Agent in order to collect the behavior information of the Agents, to obtain the cognitive topology and to try the observation methods for cognitive structures.

2. The Philosophy of Outer-Brain Observation for Cognitive Structures

The first problem for cognitive structure observation is the theoretical issue, i.e. can the inner-brain cognitive structures be coupled and induced to the outer-brain information space? Or where does the cognition take place? Is the Skull the brain's physiological boundary as well as the boundary of cognitive activities? The new cognitive view answers these questions from a complete new perspective, i.e. the cognition occurs in the coupling interaction between the brain and the environment, and through the Agent the cognition is distributed and extended into the physical environment. Nowadays, the active cognitive philosophy includes the following aspects.

1) Situated Cognition: The development of cognition, knowledge and intelligence are rooted in the interaction, mutual adjustment and adaptation of agent - environment. Cognition is an event existing in the overall context of the agent-environment interactions and correlations. The environment, more than just being a cognitive object in epistemology, also supports the existence of agent. Cognition is existence-oriented, activity-directed and it is an activity pointing to environment and can exist in the outer-brain environment.

2) Embodied Cognition: The Organisms and the environment enfold and unfold each other in the basic cycling, which presents the life itself. The cognition and knowledge are generated in the life system coupled with the environment, which is known as the second-generation cognitive scientific perspective, which is a development on the former one thinking cognition is equivalent to computing.

3) Distributed Cognition: Hutchins published his representative work- “Cognition in the wild” in 1995 and proposed that the cognitive activities occur and distribute in the cognitive environment composed of the other humans, the technical artefact, the external characteristics and the environment. Cognition is a distribution phenomenon and it is already applied in the analysis of man-machine system. Jim Hollan and Ed Hutchins established the Distributed Cognition [12-13].

4) Extended Cognition: The extended cognition and the extended mind is the cognitive philosophical hypothesis, which raised important shock and controversy in the world. Proposed by Clark in 1998 [14], this hypothesis claimed that the mind can extend outward and enter a series of cognitive objects such as the tools, the medium and the other humans. The cognitive agent can extend to the outer-brain tools. The extended cognition emphasizes that the cognition is a dynamic coupling system. Clark and Chalmers held the view that the human body and the external environment are associated interactively, and together they constitute a coupling system, in which all components are in the role through a positive causality. If these external components are removed, the behaviour capability of this system will be reduced, just like removing the constituent parts of the brain. For this reason, regardless of its existence is complete in the brain, the coupling process should be considered as a cognitive process.

Although the new cognitive perspective is still under development and even challenged, it is our focus to put outer-brain environment into the cognitive boundaries, especially special outer-brain environment like the computer, which has a coupling relationship with the inner-brain cognition. The man-machine system involves the duality relation of physics - information space and the coupling relation of psychology - information space [15].

Definition 1. Man-Machine Cognitive Coupling States: Throughout the man-machine interaction, the design of cognitive content, logic rules and procedures and orbits from the researchers also include

the understandings of the cognitive scenarios designed by the researchers and the brain's response to the scenarios after the understanding. The computer determines the next operation node according to the response process of the brain and achieves the cognitive modulation. If the activities of brain act in accordance with the intended orbit, which means the cognitive design of designers is matched with the individual cognition, then the related cognitive rule gets a validation. Otherwise, either cognitive rule needs to be modified or the cognitive modulation needs to be improved or etc. [16].

The sum of the object, content, rules, processes, orbit and so on, of the brain cognitive operations are called the cognitive information space. The introduction of cognitive information space is aimed to "physicalize" the brain activity information using the computer software and to replace the brain behaviour of the object using the Agent. The coupling information can be abstracted into the Agent behaviour in the cognitive information space, similar to the particle in the electromagnetic field.

From the outer-brain perspective, the man-machine interaction is the digital cognitive observation platform, which is constituted by Brain cognitive body - coupling Situation - information Manifolds (BSM) [17].

Definition 2. Brain Cognitive Body: the brain cognitive body takes the skull as the physiological boundaries, the physiological carrier and the material basis of current individual intelligence.

Definition 3. Coupling Situation: The coupling situation involves the man-machine situation interface, the interaction situations of the smart devices and the local objective environment. It regards the objects that can be perceived and cognized by the brain as boundaries such as the visual man-machine screen and the interface extension. The situations include the interactive situation, the cognitive data collection, the control panel and other components.

Definition 4. Information Manifold: The Agent vitro carrier includes the cognitive modulator which can modulates situation and the brain cognition into the coupling states, the cognitive data processing system and the cognitive visualization instrument. The main function of information manifolds is to image the learning process and the cognitive structure.

We hope to build the cognitive digital power system that is similar to the cloud chamber in cosmology, which is full of clean air and alcohol (or ether) saturated steam. When an invisible charged particles breaks in, it will become the core of the "cloud" condensation and form fog points to show the "footprint" of the particles movement.

Definition 5. Cognitive Coupling Digital Manifold: Based on the outer-brain perspective, the coupling cognitive man-machine interactive environment, the Agent technology the cognitions can be projected onto the information manifold, where the cognitive processes can be imaged. This manifold is called "cognitive coupling digital manifold" or "cognitive digital manifold" [17]. We hope that the "cognitive digital manifold" can become the "cloud chamber" of cognitive observation.

To summarize, the man-machine interaction system is analysed and the BSM cognitive observation platform is formed based on the cognitive coupling states from the outer-brain cognitive perspective. From the observing aspect, we still need to check whether the object observed from the man-machine interaction system and the one from the inner-brain cognitive space are unified and whether the process observed from the man-machine interaction system and the one from the inner-brain cognitive space are covariant. Therefore, it is necessary to study the mathematical structure of the cognitive coupling states and clarify the mathematical principle for BSM observation.

3. BSM Observation Mathematical Principles

Morphism, Category and Category theory can be considered as the most imagination approach generated when we human beings cognize the tools. It provides a new perspective to visualize the dynamic changes from concreteness to formalizations during the cognitive process. The famous psychologist Piaget started the research in the Category-based cognition from 1960s [2]. Now it is believed that there is Category Morphism between the brain cognition system and the computer system and their relations can be described using the Category theory.

Definition 6. The BSM Category System: During the man-machine interaction, assuming that the brain cognitive body is B category, the coupling situation is S category, the information manifold is M category and all of them are called BSM category or man-machine coupling category, then the following morphisms exist:

1) The Cognitive Morphism: It is the morphism existing between the inner-brain cognitive objects. It is based on the brain thoughts and connects the inner-brain cognitive space.

2) The Symbol Morphism: This morphism exists in the outer-brain information manifold. It is based on the symbol computation in the computer and is the morphism in the system of symbols, logic

and operations.

3) The Coupling Morphism: This morphism is constructed between the cognitive morphism and the symbol morphism by the coupling situation, so as to implement the mutual links between these two parts.

Theorem 1. The Information – Cognitive Space Cognitive Natural Isomorphism: The morphism f∈hom(X,Y) is called equivalent (f :X≈Y). If the bilateral inverse f -1hom(Y,X) exists, i.e. there is equivalent morphism between the information and the cognitive category, then the cognitive category objects are equivalent. This is called the Cognitive Natural Isomorphism Theorem, as shown in Fig. 1.

hom(B,S)×hom(S,M)→hom(B,M)

hom(M,S)×hom(S,B)→hom(M,B)

Theorem 2. The Information – Cognitive Space Cognitive Coupling Transformation: According to the Category Functor Principle, the cognitive functor keeps the coupling transformation and the covariance of the cognitive patterns between the inner-brain cognitive morphism and the information manifold morphism, as shown in Fig. 2.

The Cognitive Natural Isomorphism Principle and the Cognitive Coupling Transformation Principle can guarantee the unity of the cognition – information space cognitive objects and the covariance of the cognitive patterns, serving as the mathematical basis of BSM cognitive observation and imagery.

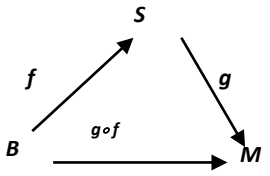


Figure 1. Cognitive Natural Isomorphism

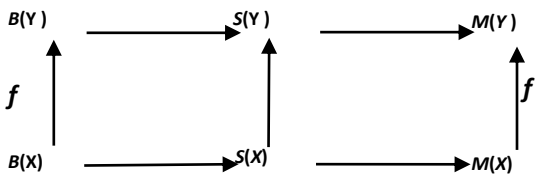


Figure 2. Cognitive Coupling Transformation

Based on the Category Theory, The Cognitive Natural Isomorphism Principle and the Cognitive Coupling Transformation Principle are concluded to provide the mathematical foundations for the cognitive coupling observation in the man-machine environment. The data acquired through the man-machine interaction process based on the Morphism Theory reflects the moving status of the objects in the inner-brain cognitive space.

4 .The BSM-Based Cognitive Structure Observation Approach

Now since the mathematical principle of the outer-brain cognitive structure observation is clear, then the contents and the approach of the observation cognitive structure should be the next step. Compared to the thinking mode of Newton and Einstein for building the Mechanical System, then a question can be raised: does the cognitive structure exist in the cognitive space or the cognitive manifold. Different form the common psychological or educational observation, the man-machine interaction observation can collect the data continuously, precisely and in large quantity, which is similar to the dynamic observation for the spatial object in Mechanics.

The basic elements in the cognitive structure are defined as the cognitive objects, which have the consistent meaning with the objects in the computer science. Therefore the truth holds to observe the cognitive objects attributes and the cognitive process through the cognitive objects in the man-machine interaction observation.

The observation and the reconstruction are closely related, i.e. the manifold hypothesis that the observation objects already exists is already included when designing the observation approach. The cognition is high-dimensional manifold, and from the information flow's perspective, the man-machine cognitive situation S is local measurement for the high-dimensional cognitive manifold, which is to say, the cognitive coupling situation during the man-machine interaction is the local mapping, reflection and expression of the high-dimensional cognitive manifold and the man-machine situation is the local measurement for the high-dimensional cognitive manifold, i.e. the cognitive manifold is partly coupled and projected to the man-machine situation space.

Definition 7. Cognitive Manifold: Manifold is the space locally with the Euclid space features. Here the brain cognition is the manifold and the Euclid space is the cognitive situation that is coupled to

the outer-brain and exists in the man-machine interaction. According to the manifold theory, the Euclid space is a measurement space and it can be observed. Similarly, the man-machine situation can be observed to locally measure the cognitive manifold, so as to study the total cognitive manifold.

Definition 8. Cognitive Frame: The local Euclid Space can be measured using the coordinates or the frame in the Manifold Theory. Similarly, the cognitive frame can be built to measure the cognitive situation. The frame structure reflects our understanding of the cognitive situations and the cognitive frame determines in the cognitive situations which information to collect and which to ignore. Based on the cognitive background, it is believed that the cognitive contents, cognitive operations and cognitive math should be collected, i.e. the frame of those contents, the COM frame should be built to collect and measure the cognitive situation.

Definition 9. The Cognitive Symbol Series: Assuming the man-machine interaction approach is used for students to dictate the English words, then 1) the computer pronounce [buk], 2) the students type in the word in the text box, 3) the computer check the correctness. Assume the student types a "b", and then followed by "u,k,o,o", then the common learning software would just record the results of the dictation and represent it with either 1 or 0. However, for the COM architecture, the operation process and the consumed time of the students should also be recorded and the following records will be generated: a) b-bu-buk-bok-book, the start time (2013-5-1 14:00) and the time span (10 seconds), rather than just b) b-u-k-o-o or c) 0 or 1 for the result. Here approach a) is called the cognitive process oriented data collection, i.e. the cognitive observation mode based on the COM architecture.

BSM-based cognitive observation is closely related with the concepts like the cognitive manifold and the cognitive frame, with the aim of building the cognitive dynamic system and put forward the cognitive research into the empirical and precise phase.

5. Analysis on the Cognitive Topology Structure

When collecting the cognitive process data based on the COM frame, how the cognitive symbol series should be handled and how the cognitive structure and process should be reconstructed? As now the cognition is abstracted into concepts such as object, process and relation, then these concepts should be presented and visualized. To reduce the complexity, the cognitive objects should be abstracted into vertex and the relation and process to edge, then the cognitive structure and the cognitive process can be built and described using the plane geometric shapes and topology, which is called "cognitive topology structure" and the "cognitive topology manifold".

Based on the elements such as vertex and edges, several models including the complex network are introduced to describe the cognition, so the complex cognitive network model can be built, the cognitive process can be simulated, and therefore the cognitive research can be visualized. From the perspective of manifold, the analysis of cognitive topology is to project the high-dimensional cognitive manifold into the low-dimensional cognitive manifold; for example, the cognitive network ψ is the bottom space for the high-dimensional cognitions.

1) The Basic Object In Cognitive Topology Structure: points and lines are the basic elements in network ψ, in the visualized cognitive networks, the point means the cognitive object P, with the basic attributes such as the name, the size, the colour. Line L means the relations between two P, with the attributes such as the name, the size, the direction and the colour. The points and lines are the basic topological elements in cognitive structure and they can make up a geometric shape or a visualized network object, forming a topological structure.

In the real cognitive analysis, the cognitive points and lines can either be built based on knowledge, psychology, pedagogy or experience, or the new cognitive objects can be abstracted through the analysis of the cognitive observation data.

2) The Basic Parameter in Cognitive Topology: The order parameters are important parameters for system evolution. For the cognition, the cognitive time, the cognitive counts etc. are all important order parameters.

3) The Cognitive Flow: The cognitive process can be described using the set {L, P} to form the time and process series within the cognitive process and present the process information of the cognitive activities. The cognitive flow can be abstracted into the cognitive mode and it is the cognitive orbit on the dynamic system level.

4) The Cognitive Logic Relations: The points can form the basic cognitive relation system, but another important factor for cognition is the cognitive logic coL, and it requires not only the lines but also the mathematical logic operators to express, for example, the logic relation between P1，P2 can be expressed using $\rightarrow$.

5) The Analysis On the Cognitive Object Group: Group is an important concept in both physics and math. The analysis on the cognitive group aims at studying the relations between the cognitive objects, such as the duality relation, the symmetry relation, etc.

6) The Cognitive Topology Space: The observed cognitive objects gather into the cognitive set, therefore, the Topology Theory is required to do the analysis on it. According to the Topology Theory, adding the related additional structure into the cognitive set and classifying the cognitive objects in a new way can form the cognitive topology space $\prod$. The cognitive topology is important for cognitive research, as not only the topology space is easy to build, but also the cognitive topology is the foundation for connecting the cognitive measurement and the Dynamics.

To sum up, the basic concepts in the geometrical and topological description of cognition is defined and it serves as the foundation for the analysis and reconstruction of the cognitive observation data.

6. The Observation Workflow for Cognitive Topology Structures

For the specific cognitive process, the man-machine interaction software, the cognitive symbol series are designed and the man-machine situation interface is vectored to implement the cognitive process oriented information observation. The observation workflow is shown as follows:

1) Vectored the man-machine interaction situation interface. The cognitive situations within the man-machine interaction are the aim of the cognitive structure observation and the cognitive process is done in this situation interface. According to the requirement, the man-machine interaction area is deployed to collect the man-machine operation information, the interface elements such as the buttons; the text box, the circular zone, etc. can all record the users' operations and reflect the cognitive process information.

2) The Design of Cognitive Symbol Series Based on COM frame: According to the cognitive contents, the exact cognitive symbol series structure is designed, including elements such as cognitive objects, attributes, the order parameters etc. The simplest cognitive symbols are the cognitive situation information collected from the keyboard, the mouse, the time etc. from the man-machine interaction.

3) The Design of Multi-dimensional Multi-Sized Cognitive Objects Set: Building the set is the basis for further mathematical studies and an important step for processing the cognitive data. Based on the individuals, the group, the sex, the location, etc. the multi-dimensional cognitive symbol series set can be built, and based on the start time, time span, the age, the grade, the area, etc. the multi-sized cognitive set can be built.

4) Visualized Analysis for the Cognitive Topology: The visualized cognitive objects such as points, lines, networks and graphs, etc. can be built from the perspective of geometry and topology, and the cognitive process can be projected to the low-dimensional manifold for further process.

5) The Design of Cognitive Dynamics System: The cognitive data observed is discrete, so how to recover the data into continuous or even differentiable mathematical structures? Based on the continuous parameters like time, age, grade, etc. the discrete data can become continuous.

The visualization of cognitive structure is to build more "object" cognitive mathematical model and to externalize and physicalize the inner-brain cognitive structure into the measurable, visual objects in the information space.

7. Applications for Cognitive Structure Observations

Take the 24-points cognitive computing tasks for students as an example to conduct the above analysis. Within the 24 points game, the student is given four numbers from 1 to 9 and then combines them with four math operations such as plus, minus, times or divides so that the result is 24 (or no solution).

Using the 24-points cognitive computing man-machine interaction software system, where each calculation operation is a record, nearly 80,000 cognitive data records are collected from more than 2000 students ranging from the 3-Grade pupil to the college student and then the cognitive structure images are built based on the above approach.

1) The Interface for Cognitive Symbols Series Collection: The computer gives four numbers in random and then the user clicks on the corresponding operators to form an arithmetic expression. The cognitive symbols series man-machine interface for the 24-points calculation is shown in Fig. 3.

Figure 3. The man-machine interface for the 24-points calculation

|(|(2|(2+|(2+4|(2+4+|(2+4+2|(2+4+2)|(2+4+2)*|(2+4+2)*3|_true_spent:8s
|(|(2|(2+|(2+4|(2+4+|(2+4+2|(2+4+2)|(2+4+2)*|(2+4+2)*3|_true_spent:8s
|(|(2|(2+|(2+4|(2+4+|(2+4+2|(2+4+2)|(2+4+2)*|(2+4+2)*3|_true_spent:8s
|3|3*|3*(|3*(4|3*(4+|3*(4+2|3*(4+2+|3*(4+2+2|3*(4+2+2)|_true_spent:11s
|3|3*|3*(|3*(4|3*(4+|3*(4+2|3*(4+2+|3*(4+2+2|3*(4+2+2)|_true_spent:11s
|3||(|(4|(4+|(4+2|(4+2*|(4+2*2|(4+2*2)|(4+2*2)*|(4+2*2)*3|_true_spent:9s
|3|3*|3*(|3*(2|3*(2*|3*(2*2|3*(2*2+|3*(2*2+4|3*(2*2+4)|_true_spent:65s
|(|(4|(4+|(4+2|(4+2+|(4+2+2|(4+2+2)|(4+2+2)*|(4+2+2)*3|_true_spent:8s

Figure 4. The cognitive symbol series collected for the 24-points game

2) Cognitive Symbol Series: The related data is collected when the learners click on the related cognitive sensors and the records are generated, shown in Fig. 4 (the user ID and time point are not listed here).

3) Cognitive Set: How to handle large quantities of cognitive symbol series? Classifying and Setting up sets are the basis. For example, when the records are classified by the average time for each calculation, then the records with the toughest solution can be searched. In Fig. 5, the calculation question, the test count and the average time are all listed, for example, the 3-3-5-8 is tested by 81 times and the average time is 15.1 seconds.

3-3-5-8	81	15.1
2-3-7-8	83	15.1
2-4-4-5	41	15.0
7-8-8-9	15	14.9
6-6-6-9	14	14.9
1-4-6-6	50	14.7
3-4-9-9	50	14.7
2-4-4-9	34	14.6
2-3-6-6	58	14.4
2-2-8-8	27	14.4
2-6-6-9	25	14.3
1-6-7-8	174	14.3

Figure 5. The average time for the 24-points game

4) The Analysis of Cognitive Topology: The cognitive topology space is built based on the cognitive sets. For example, using the algebra approach, (4+2+2)*3 can be mapped into (A+B+C)*D, therefore cognitive records like (3+3+2)*3 and (1+3+2)*4 can all be mapped into the image (A+B+C)*D, and all the images can compose the new level cognitive topological space, which is also called quotient topological space.

After the algebra analysis of 62401 records, 232 images are acquired, which is much less compared to the number generated by the pure permutation and combinations, which means that the cognitive operation is different from the pure mathematical way. Because of these cores are fundamental, so we call them the gene of 24-points cognitive computation, as is shown in Fig. 6.

((A*B)-C)/D	((A+B-C))*D	((A-B)-C)*D
((A+B)-C)*D	((A-(B-C)))*D	((A-B)/C)*D
((A+B)/C)*D	((A-B))*(C+D)	((A-B+C))*D
((A+B+C))*D	((A-B))*C*D	((A/B)+C)*D

Figure 6. Part of the patterns

5) Cognitive Topological Imagery: Four edges of a rectangle Q1, Q2, Q3 and Q4 are built to represent the four numbers in 24-points game, with each cognitive calculation of the 24-points game

mapped to a point on the rectangle. Then the Q's coordinates can form the diagram shown in Fig. 7. Furthermore, different colours can be given to the points to represent the corresponding time level of the cognitive calculation, for example, red means the calculation with most time, while green means the one with least time. Then a cognitive topology diagram can be formed, shown as in Fig.8.

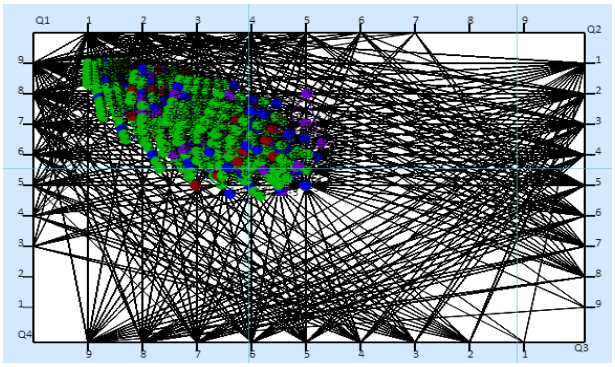


Figure 7. The imagery of the cognitive topology

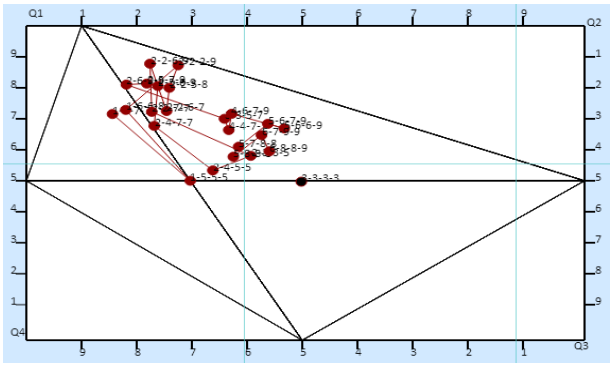


Figure 8. The network diagram of cognitive topology

To sum up, the 24-points cognitive calculation is analysed in detail on its process level, topology level and dynamics level based on the man-machine interactions. By using the visualization technology, the related abstract data are presented to support the educational application.

6) Cognitive Topology Dynamics Simulation: The dynamics simulation of 24-points is conducted for a further understanding of its cognitive process.

1. The Mathematical Possibility: Assume the computer gives four numbers as 2,6,6,6, then after the exhaustive search, 46 possible algorithms and expressions are acquired, part of which are listed below.

1:A*B+C+D-2*6+6+6: 18.0(47.36%)
2:A*B-C*D-6*6-2*6: 5.0(13.15%)
3:(A*B)+C+D-(2*6)+6+6: 4.0(10.52%)
4:A+B+C*D-6+6+2*6: 4.0(10.52%)
5:(A+B)+C*D-(6+6)+2*6: 2.0(5.26%)
6:A*(B/C)+D-6*(6/2)+6: 2.0(5.26%)
7:A/B*C+D-6/2*6+6: 1.0(2.63%)
8:(A/B)*C+D-(6/2)*6+6: 1.0(2.63%)
9:(A*B)-C*D-(6*6)-2*6: 1.0(2.63%)
10:((A*B)+C)+D-((2*6)+6)+6: 0.0(0%)
11:(A*B)+(C+D)-(2*6)+(6+6): 0.0(0%)
12:(A*B+C)+D-(2*6+6)+6: 0.0(0%)
13:A*B+(C+D)-2*6+(6+6): 0.0(0%)

Figure 9. Part of the algorithms for the 24-points game with 2,6,6,6

2. Psychological Implementation: In the practical cognition, only nine solutions out of the 46 possibilities are observed, while the others do not appear in the psychological world. According to the collected data, the application percentage of each algorithm can be calculated and 2*6+6+6 is the most popular one.

3. Order Analysis: Is there any difference between 6*6-2*6 and (2*6)+6+6? The former one can be the result from 36-12=24, while the latter is the result from 12+12=24. Here 6*6-2*6 and (2*6)+6+6 are called first order, 36-12 and 12+12 are called second order. Through the order analysis, the user preference can be acquired.

In Fig. 10, on the left side the possibilities of using 2,6,6,6 to calculate 24 are listed, together with their percentage in the psychological implementation. In the middle, the first two columns present the states of the two first order calculations, the third column presents the algorithm for the second order calculation and the fourth column lists the time for calculation.

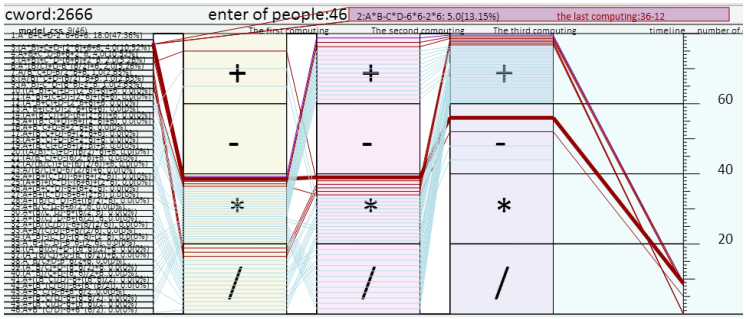


Figure 10. The order analysis of the 24-points calculation

Besides the research on the 24-points game, the English dictation for Chinese students is also used for the cognitive observation. 1) The computer pronounce [BUK] and the student write down the result, 2) If the student does not get the accurate answer after a while, then a tip is given in Chinese from the textbook, 3) If the student still does not get the right answer, then the test for this word is done and a new test starts for further dictation. The "+" means the situation when the Chinese tip appears, and part of the cognitive series is: book_b-b-bl-b+bo-boo-book;book_b-bo-box-bo-b-b-bo-boo-book;book_b-bo-box-bo-b-bu-bux-bu-b-bo-boo-boox-boo-bo-b+;book_b-bo-bok-bo-boc-bo-b-b-bo-bog-bo-b-d-do-dog-do-d-b-bo-boo+book;book_b-bo-boc-bocj-boc-bock-bck-buck-bck-back-bck-bock-bck-beck-bec+be-b-b-be-bec-beck-bck-bick-bck-beck-beack-beck-bck-bc-boc-boce-boc-bo-b or sick_s-si-sit-si-s-s-so-sou-soup-sou-so-s-s-si-sil-silk-sil-si-s+s-si-sil-sill-sillk-sill-sil-si-sil-si-s-s-si-sit-si-sil-silt-sil-si-sic-sick. Sometimes the student get the first letter wrong for some worlds, for example, in the case of book_d-do, it is a confusion between B and D. By analysing the 18,900 records, the following results can be acquired, as shown in Fig. 11. For example, It is quite common to mix up A and O, then E and H. For O it is easy to mix up with A. (The red means the two letters are easy to mix up mutually)

C	E1	E2	E3	C	E1	E2	E3	C	E1	E2	E3	C	E1	E2	E3	C	E1	E2	E3
A	O	E	H	F	T	S	H	K	N	C	P	P	H	B	F	U	A	O	Y
B	D	P	M	G	J	D	B	L	R	W	B	Q	C	K	A	V	W	B	M
C	K	S	T	H	P	T	A	M	N	B	I	R	L	W	B	W	L	R	B
D	B	G	T	I	A	E	B	N	M	T	L	S	C	T	A	Y	E	U	H
E	A	I	H	J	G	H	D	O	A	B	W	T	B	F	D	Z	J	T	E

Figure 11. The list of easy-to-mix-up letters for Chinese students

A-Z mix-up analysis is similar to the group analysis in Mathematics, and it reflects the relations among A-Z.

8. Conclusions

From the new cognitive perspective, the man-machine system is an information-cognition coupling system. As the inner-brain cognitions are distributed into the outer-brain environment, it is

possible to observe the inner-brain cognition from outside. From the perspective of Category Theory, there is cognitive morphism for the information-cognitive space and the cognitive natural transformation maintains the cognitive natural isomorphism.

The cognitive symbol series collected based on the man-machine coupling states contain the inner-brain cognitive structure information: Through the design of objects such as cognitive points, lines, connections, logics, the cognitive symbol series are analysed and the cognitive topological network diagram is built, so as to realize the visualization of the cognitive structure based on the complex cognitive networks.

Through the cognitive structure observation based on the man-machine coupling states, large quantities of cognitive data can be acquired and its dynamics model can be built using the mathematical approach, so as to provide new ideas for the research on the cognitive structures.

References

1. Zhang, G., Yang, J. (2010) The Study on the Generation and Development of the Cognitive Structure. [J] *The Journal of Jilin Education College.* 26(2), 100-101. (in Chinese)
2. Piaget, J., et al. (2005) The Morhpism and the Category: Comparsion and Transformation [M]. Shanghai, Huadong Normal University Press, September, 11-37 (in Chinese)
3. Cacioppo, J.T., von Hippel, W., Ernst., J.M. (1997) Mapping Cognitive Structures and Processes Through Verbal Content:The Thought-Listing Technique. *Journal of Consulting and Clinical Psychology*, 65(6), 928-940.
4. Yoram, B.-T. (2010) When the Need for Cognitive Structure does not Cause Heuristic Thinking: The Moderating Effect of the Perceived Ability to Achieve Cognitive Structure. *Psychology*, 1, 96-105. DOI:10.4236/psych.2010.12013.
5. Li, L., Wang, S., Du, R. (2011) The Dynamic Bayesian Network Classifier for Cognitive Structure Evaluation. *The Journal of Central South University (Natural Science Version).* September, 42(1), 615-619. (in Chinese)
6. Dimitrov, D.M., Raykov, T. (2003) Validation of Cognitive Structures: A Structural Equation Modeling Approach. Multivariate Behavioural Research, 38(1), 1-23, Lawrence Erlbaum Associates, Inc.
7. Froese, T., Suzuki, K., Ogai, Y., Ikegami, T. (2012) Using Human–Computer Interfaces to Investigate‘Mind-As-It-Could-Be’ from the First-Person Perspective. *Cogn Comput.*, 365–382. DOI :10.1007/s12559-012-9153-4.
8. Kim, S., Kwon, S., Kweon, I.S. A Perceptual Visual Feature Extraction Method Achieved by Imitating V1 and V4 of the Human Visual System. *Cogn Comput*. DOI:10.1007/s12559-012-9194-8.
9. Kievit-Kylar, B., Jones M.N. (2012) Visualizing multiple word similarity measures. *Behav Res,* 44, 656–674. DOI: 10.3758/s13428-012-0236-7.
10. Moran, D.J., Malott, R.W., Interpreted by XIAO Yan and SHAO Ran. Empirical Educations. China Light Industry Press. September 2006:1-11. (in Chinese)
    Yu Xiaohan (2010) The Systematic Study on the Cognitions-From the Perspective of Distributed Cognition [D]. The Faculty of Humanities, Zhejiang Unviersity, 2. (in Chinese) http://hci.ucsd.edu[EB/OL].
11. Hollan, J., Hutchins, E., Kirsh D. (2000) Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research [J], *ACM Transactions on Computer-Human Interaction,* 7(2), 174–196.
12. Clark A. (2008) Supersizing The Mind. by Oxford University Press, Inc, 3-43.
13. Xu, G., Tao, L., Zhang, D. et al. (2006) The Duality Between the Physical Space and the Information Space. [J]. *Science Communications,* 51(5), 610-616. (in Chinese)
14. Meng, S., Cheng, R. (2011) Cognitive Coupling States Based on Tree Cognitive Fields, 2011 International Conference on Computer Communication and Management Proc. of CSIT vol.5 Singapore, 2011 (pp. 593-597) IACSIT Press.
15. Meng S., Liu, C. et al. (2013) The Cognitive Digital Manifold Based on the man-machine interactions. [J] *Computers and Information Technology*, 1(1-8), 49. (in Chinese)
16. Liu, C., Meng, S., (2012) The Cognition-oriented 1:1 Digital Classroom Information Series [J]. *The Journal of Changchun Normal College (the Natural Science version)*. 31(12), 121-124. (in Chinese)
17. Meng, S. (2012) Visualization Complex Cognitive Networks [C].2012 3rd International Conference on e-Education, e-Business, e-Management and e-Learning. IPEDR, 27, 2012, (pp.1-5) IACSIT Press, Singapore.

Meng, S., et al. (2012) Complex word network model based on cognitive coupling states [C], Information Society (i-Society), 2012 International Conference on Date: 25-28 June (pp. 411- 414) Copyright © 2012 by Infonomics Society, IEEE conference publications: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6285006

18. Bullmore, E.T., Sporns, O. (2009) Complex Brain Networks: Graph- theoretical Analysis of Structural and Functional Systems Nature Reviews Neuroscience, 10(3), 186- 198.

Received on the 10th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 185-190
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

INTERACTIVE CLOTHING SEGMENTATION BASED ON *JSEG* AND *GRABCUT*

W. Li[a,b], J. Wang[a*], G. Lu[a]

[a]State Key Laboratory of CAG&CG, Zhejiang University
Hangzhou, China
[b]School of computer, Guangxi University of Science and Technology
Liuzhou, China
Phone: (+86)13588154448. E-mail: dwjcom@cmee.zju.edu.cn, gxlwl@zju.edu.cn

Clothing segmentation is a challenging issue of research, which is gaining more and more attention. This paper presents an interactive clothing segmentation method based on JSEG and GrabCut. Firstly, we pre-process the image from pixels to regions using JSEG algorithm for the purposes of lessening computer loads and reducing the noise of image, and then build a graph using regions as nodes, and define the seeds of clothing or non-clothing areas interactively. Finally, solve the energy function of the graph using min-cut (max-flow) approach, and separate the clothing from the image. In some case of inaccurate boundary, we adopt repeated segmentation technology to repair the results. Some given samples show that the approach is effective and can achieve promising results.

Keywords: interactive segmentation; JSEG; Grabcut; clothing segmentation

1. Introduction

Clothing is one of the key information of human appearance and segmentation is one of the most challenging tasks in computer vision and image processing. Clothing segmentation widely used in many fields, such as human detection and recognition [1-2], virtual try-on, e-commerce, fashion-aided design and so on. In recent years, more and more researchers pay close attention on it.

One of the most popular approaches to clothing segmentation is based on graph cuts [1-7]. They use a graph cut segmentation technique involves a Markov Random Field (MRF) framework [8, 12], separated the clothes, the human body and other regional respectively. Although these approaches can tackle a diverse range of clothing robustly, they still suffer in accuracy, gaining very rough segmentation results. In addition, they often require a lot of study process and computational loads, especially in the face of some complex clothing styles and patterns. The others are based on model or template [9-11]. These segmentation methods are restricted to the numbers of the model species, unable to adapt to the complex clothing style and changeable human posture.

Since clothing image having rich colours, patterns and styles, numerous complex structure, and a variety of different lighting and background conditions as well as humans and other objects blocking characteristics, the traditional image segmentation algorithm cannot meet the needs of clothing segmentation.

In this work, we present an improved grabcut [12] method for interactive clothing segmentation. Firstly, the image is pre-processed from pixels to regions by JSEG, and secondly a graph is built using regions as nodes. Then specifies the seeds of clothing and non-clothing regions interactively. Finally, min-cut/max-flow algorithm was implemented to solve the graph, and the clothing regions are segmented.

2. Methodology

2.1. Pre-Process the Image from Pixels into Regions by JSEG

JSEG [13] is one of the most robust methods to identify homogeneous and non-homogeneous regions of a colour image. The algorithm takes the image's colour information and spatial information into account. Firstly, we quantize colours in image into several representative classes that can distinguish regions in the image. Then combine smaller image areas by the region growing method.

Let Z be the set of N pixels contained in the image and m be the mean of all the N points, $m = 0.5$. Suppose that Z has been classified into C classes, $Z_i, i = 1,2,\cdots\cdots,C$. And denote the spatial

mean of pixels in Z_i as m_i .

$$m_i = \frac{1}{N_i}\sum_{z\in Z} z \tag{1}$$

The main process of JSEG is as follows:

Step1: convert the image colour space to LUV colour space;

Step2: smoothing and denoising the image retain the edge information of image and remove the influence of particles noise;

Step3: calculate the ' J ' value of the template centre according to the colour distribution in the template, and get the map of " J " value; the measure J is defined as follows,

$$J = (S_B - S_W)/S_W \tag{2}$$

Where, S_B is called between class scatter matrix, it measures the distance between different classes. And S_W , a measure with-in class scatter matrix is defined as follows.

$$S_B = \sum_{z\in Z} \|z - m\|^2 \tag{3}$$

$$S_w = \sum_{i=1}^{C}\sum_{z\in Z} \|z - m_i\|^2 \tag{4}$$

Step4: calculate the regional growth of seed points in class map; grow from the non-seed region to seed region according to the distance of the adjacent regions;

Step5: merge the colour histogram of the minimum Euclidean distance between two adjacent areas until the area is larger than the smallest gathering region-merging threshold of image segmentation.

Because the JSEG algorithm may produce over-segmentation, we combined threshold limits by adjusting the number of regions allowed to exist. The results is shown in Fig 1.

Figure 1. Results of JSEG algorithm

2.2. Construct a Graph Using Regions as Nodes

Let $Z = (z_1, \cdots, z_n, \cdots, z_N)^T$ be the set of image regions by JSEG, where z_n is the mean of colour scale of the region *n*. Define the undirected graph $G = \langle V, E \rangle$, where V is the set of image regions, E is the set of the connections between regions. Therefore, the image segmentation converts to the node segmentation of graph G. For each region a_n defines in an array $a = \{a_1, \cdots, a_n, \cdots, a_N\}$, $a_n \in \{0,1\}$, 0 indicates that the region belongs to the background, 1 indicates that the region belongs to foreground (target), the image segmentation problem is the problem of solving an array of a_n. By minimizing the

energy function $E(a,\theta,z)$, we can obtain the optimal solution a, the energy function including regional boundary characteristic feature items $U(a,\theta,z)$ and items $V(a,z)$.

$$E(a,\theta,z)=U(a,\theta,z)+V(a,z) \tag{5}$$

Where $U(a,\theta,z)=\sum_{n}-\log(p(z_n;a_n))$ describes the split characteristics of region a_n, and the probability of foreground is $-\log(p(z_n;1))$ or the background is $-\log(p(z_n;0))$, and the models of foreground and background are decided by the predefined seed points; $V(a,z)=\gamma\sum_{(m,n)\in C}B_{\{m,n\}}[a_n\neq a_m]$ is the boundary characteristics of region a_n, the coefficient $B_{(m,n)}=dis(m,n)^{-1}\exp(-\beta\|z_m-z_n\|^2)$ is the penalty coefficient of the region m and n, if the region m and n are similar, the coefficient will be relatively large, while the greater the difference of the region m and n, the coefficient will be smaller; γ is the weight coefficient of them.

2.3. Define the Seeds of Clothing and Non-clothing Regions Interactively

According to the results of image pre-processing segmentation, we define the seed points interactively in the picture using dashes, and indicate which areas are clothing, which areas non-apparel. In this paper, we identify the red interaction curves in clothing area, the blue curve in background, as shown in Fig 2.

Figure 2. Definition of clothing and non-clothing regions

If the colours of garment are quite complex, interactive determining seed point must across the region contains more colours, it can get more accurate segmentation results.

2.4 Solve the Graph with Min-Cut (Max-Flow) Algorithm

GrabCut [12] is an interactive image segmentation algorithm based on colour Gaussian Mixture Model (GMM) and iterative energy minimization. The energy minimization is optimized by min-cut (max-flow) algorithm [14]. The main process of the whole segmentation algorithm is as follows:

1) Inputs the image and pre-process the pixels to regions using JSEG.
2) Creates an initial trimap by defining some seed regions interactively. Pixels inside the regions with blue brush are marked as known background; Pixels inside the regions with red brush are marked as known foreground. Pixels outside of regions are marked as unknown.
3) Creates an initial image segmentation, where all marked foreground regions are tentatively placed in the foreground class, and all marked background regions and unknown regions are placed in the background class.
4) Constructs Gaussian Mixture Models (GMMs) for initial foreground and background classes using the K-means clustering algorithm.

5) Assigns each region in the foreground class to the most likely Gaussian component in the foreground GMM. Similarly, each region in the background is assigned to the most likely background Gaussian component.
6) Builds a graph and using min-cut (max-flow) algorithm to find a new tentative foreground and background classification of regions.
7) Updates the parameters of GMMs from the regions sets created in the previous set.
8) Repeats Steps 5-7 until the classification converges.
9) Obtains the result of clothing segmentation.

3. Results and Discussions

In order to verify the correctness of our method, we use a computer with Pentium 4 2.4GHz, 1G memory and Matlab2007b software; choose the images of size 480 × 640, several common kinds of clothes styles. The results are shown in Figure 3.From top to bottom, the image number is named as Image-1, Image-2, Image-3.

Figure 3. Comparison with GrabCut [11] and proposed method

From the results, the proposed method can accurately distinguish background and objectives, and extract a better picture of the clothing.

To further verify the effectiveness of our algorithm, an objective criteria is taken to evaluate the merits of the algorithm. In this paper, we use the absolute error rate as the main indicators. Let n_0 is the number of pixels obtained ideally from the target, by visual observation using the manual segmentation method. n_i is the number of pixels obtained by the method adopted various target dividing, N is the total number of pixels in the image, so the *err* is defined as the absolute error rate.

$$err = \left| \frac{(n_i - n_0)}{N} \right| * 100\% \tag{6}$$

From the formula (3), we know that the absolute error is littler, the segmentation result is better. The absolute error of Fig 4 is computed in table 1.

TABLE 1. The absolute error of segmentation results

Image No	GrabCut [11]	Our method
	err	err
Image-1	1.3%	0.5%
Image-2	6.2%	0.2%
Image-3	0.3%	0.3%

When the clothing colour and background colour are very similar in the boundary or contrast is not obvious, and the result has some defects, as shown in Fig 4. To handle this problem, we provide an interactive segmentation method that allows users to select the area, which is not satisfied, and the system will divided again in the selected area, rather than in the whole image.

Figure 4. Results of repeated segmentations

4. Conclusions

In this paper, we present an interactive clothing segmentation approach that is effective and easy to use. Relative to the existing methods, our method do not need prior learning and training, requiring only a small amount of user interaction. In order to tackle the incorrect object boundary, we apply repeated segmentation to repair the separate results. However, we found that the method still has some shortcomings, especially in the case of background colour is similar with clothing colour, and the segmentation effect will be affected to a certain extent. In the next step, we will consider of texture, shape and semantic information, such as object features to enhance reliability of the segmentation results.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.51275460, 61103106), the project of public technology research in industry of Zhejiang province (No.2013C31046), the Doctoral Program of Higher Education of China (No. 20100101110025).

References

1. Gallagher, A.C., Chen, T. (2008) Clothing Cosegmentation for Recognizing People[C]. In Proceedings of CVPR 2008, 1–8.
2. Yang, M., Yu, K. (2011) Real-time clothing recognition in surveillance videos. In Proceedings of ICIP 2011, 2937-2940.
3. Hasan, B., Hogg, D. (2010) Segmentation using Deformable Spatial Priors with Application to Clothing. In: *BMVC*. 1-11.
4. Cushen, G.A., Nixon, M.S. (2012) Real-Time Semantic Clothing Segmentation. In ISVC, 272-281
5. Hu, Z., Yan, H., Lin, X. (2008) Clothing Segmentation Using Foreground and Background Estimation Based on the Constrained Delaunay Triangulation[J]. Pattern Recogn, 41(5), 1581–1592.
6. Yao, S.S., Khan, I.R., Farbiz, F. (2011) Clothing segmentation and recoloring using background subtraction and back projection method[C]. In Proceedings of ICIP, 3137-3140.
7. Huang, L., Xia, T., Zhang, Y., Lin, S. (2012) Pose-Invariant Clothing Segmentation Based on Saliency Detection and Graph Cuts[J]. *Journal of Computer-Aided Design&Computer Graphics*, 24(5), 620-627.
8. Boykov, Y., Veksler, O., Zabih, R. (2001) Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11), 1222-1239.

9. Wang, N., Ai, H. (2011) Who Blocks Who: Simultaneous Clothing Segmentation for Grouping Images[C]. In: *ICCV*, 1535-1542.
10. George, A.C., Mark, S.N. (2011) Markerless Real-Time Garment Retexturing from Monocular 3D Reconstruction. in: Proceedings of the IEEE International Conference on Signal and Image Processing Applications (ICSIPA2011), 88-93.
11. Chen, H., Xu, Z.J., Liu, Z.Q., Zhu, S.C. (2006) Composite Templates for Cloth Modeling and Sketching[C]. In CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 943–950.
12. Rother, C., Kolmogorov, V., Blake A. (2004) GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts[J]. ACM Trans on Graphics, 23(3), 309-314.
13. Deng, Y., Manjunath, B.S. (2001) Unsupervised segmentation of colour-texture regions in images and video [J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 23(8), 800-810.
14. Boykov, Y., Veksler, O. (2006) *Graph cuts in vision and graphics:theories and applications*. Handbook of Mathematical Models in Computer Vision. New York: Springer, 79-96.

Received on the 10th of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 191-195
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

HAND GESTURE TRACKING AND RECOGNITION TECHNOLOGY BASED ON COMPUTER VISION

X. Wang, X. Dong

School of Information Engineering, Huanghe Science and Technology College
Zhengzhou, 450006, China
Phone: (+86)13303820802. E-mail:38856735@qq.com

With the development of mobile terminal, multimodal human-computer interaction technology has become a hot spot in research field. Gesture recognition based on computer vision is a key technology of multimodal human-computer interaction, which has important theoretical research value and application prospect. The gesture image pre-processing, feature extraction, gesture tracking and recognition are studied, using the CamShift (Continuously Adaptive Mean-SHIFT) algorithm for real-time hand gesture tracking to achieve good tracking effect in monochrome background. In Microsoft Visual C++6.0 programming environment, it can achieve the recognition of 10 commonly used gestures of camera input, and then set up a simple gesture interactive system in order to verify the feasibility of the algorithm. The experimental results show that this method has good robustness in real-time tracking of target gesture, and good result of the recognition of gestures.

Keywords: Computer Vision; human-computer interaction; gesture tracking; gesture recognition

1. Introduction

With the rapid expansion of computer influence on modern society, multimodal natural human-computer interaction technology becomes more and more important in real life. The research aim on multimodal human-computer interaction technology is to create a harmony human-computer interaction environment, so that the users can use the computers conveniently. One of the most important link is to make the computer can accurately perceive natural language, gestures, facial expressions, and different human expression, achieving the people-oriented man-machine interactive technology [1]. As one of the ways of expression, hand gesture has been widely used in our daily life. Vivid intuitive gestures contain rich information, which are the important medium of communication between people. The gesture-based interaction has become an important content of human-computer interaction.

For the realization of multimode human-computer interaction, aiming at the problem of detection and tracking of alphabet gesture, a gesture recognition algorithm based on the maximum likelihood criterion of Hausdorff distance is presented in literature [2], it can be used to well identify the alphabet gesture and has good effect on the deformation (rotation and scaling) gestures. According to the basic gesture features used in the tracking process, a static hand gesture recognition method based on the PGH (pairwise geometric histogram) is proposed in literature [3]. However, with the increasing of gesture recognition, system error rate will also increase. A representation method of history gestures is adopted in literature [4], establishing a gesture appearance model with skin history image, and matching prototype template for gesture recognition. The system background is assumed static objects closed to human skin, without considering the effects of light. In literature [5], Li and Jarvis realize multimodal human-computer interaction based on gestures and the body. Using the binocular camera, three-dimensional position information and the gaze direction are obtained. The stereo matching ambiguity and its complexity leads to the unreliable 3D information [6]. At the same time, the calculation is complex and difficult to meet the requirements of real-time human-computer interaction.

This research focuses on the hand gesture tracking and recognition based on computer vision. Firstly, the algorithm of the hand gesture tracking and recognition of each stage is analysed, the hand gesture recognition algorithm based on Hu moment feature and radial basis kernel function SVM (support vector machine) is presented. In the development environment of Microsoft Visual C++6.0, it has achieved the recognition of 10 common gestures from camera input to verify the feasibility of proposed algorithm, and has established a simple gesture interactive system based on this.

2. Hand Gesture Tracking and Recognition System Based on Computer Vision

A general hand gesture tracking and recognition system based on computer vision is shown in

Figure 1. First, use the video camera to capture video data flow; to see if there are gestures in the stream interaction model, and segment the hand gestures from video signals. Then the image pre-processing and feature extraction are used to obtain ideal gesture segmentation effect. Choose the moment features as feature vector of gesture image; adopt the CamShift algorithm for real-time hand gesture tracking. The Hu moment feature and "one-to-many" radial basis kernel function SVM algorithm is applied to gesture recognition. The 10 digital sign simulation experiments are done, and then set up a simple gesture interactive system.

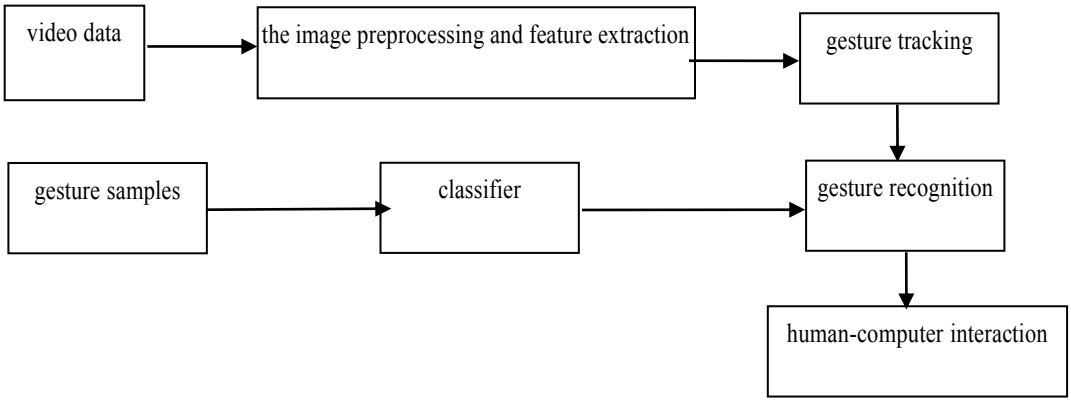


Figure 1. Hand gesture tracking and recognition system based on computer vision flow chart

2.1. The Gesture Image Pre-processing and Feature Extraction

It is very important to establish a good sample database for statistical recognition. The gesture images obtained by the camera are segment into monochrome gesture images to set up a sample library. Specifically, the first step is to do sample collection. Limit the shooting background for monochrome background in the acquisition of gesture samples, but allow the light intensity changes. Set the gesture image normalized to the 200×240 pixel size. The samples selected are preserved in corresponding folder according to certain rules. In this paper, the identification of 10 digital 0-9 gesture is made, as shown in Figure 2 and Figure 3. Figure 2 is the RGB image acquisition from the camera, figure 3 is the corresponding image segmentation, which is the gesture sample to be saved.

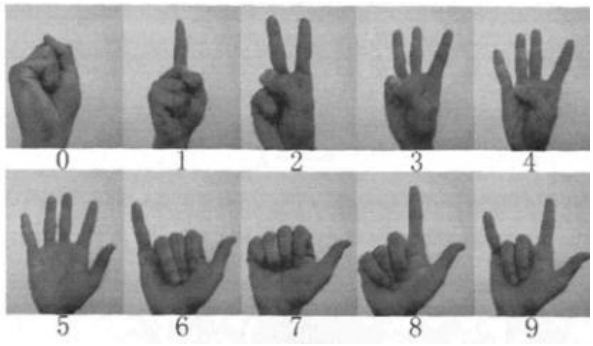


*Figure 2.*Digital gesture RGB image

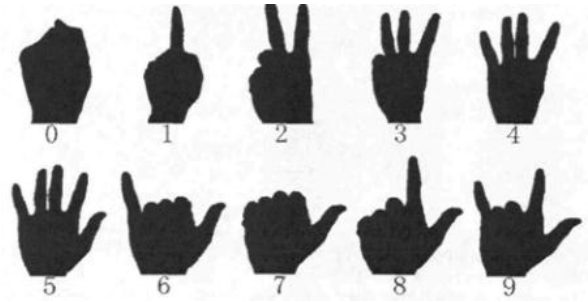


Figure 3. Schematic diagram of digital gesture segmentation

The gesture image transform from RGB space to the YCbCr space, using the characteristics of the luminance and chrominance separation of YCbCr colour space. To a certain extent, this method avoids the interference of light, and then use the Otsu method to make image processing on the image of Cb and Cr channels. The experimental results show that, Cb and Cr colour channel has better segmentation effect compared with the Y channel. The specific process of Otsu method is as follows [7]:

1) The calculation of normalized histogram $h(i)$;

2) Calculate the mean grey-scale value $u_T = \sum_{i=0}^{255} i * h(i)$;

3) Calculate the histogram of zero order cumulant $w(k)$ and cumulative moment $u(k)$:

$$w(k) = \sum_{i=0}^{k} h(i) \quad k = 0,1,...255;$$

$$u(k) = \sum_{i=0}^{k} i * h(i) \quad k = 0,1,...255;$$

4) The calculation of separation index: $\sigma_B(k) = \frac{[\mu_T \omega(k) - \mu(k)]^2}{\omega(k)[1-\omega(k)]} \quad k = 0,1,...255;$

5) The k corresponding to $\max(\sigma_B(k))$ is the optimal threshold value T;

6) Get binary image according to the optimal threshold value:

$$g(x,y) = \begin{cases} 255 & f(x,y) > T \\ 0 & f(x,y) \le T \end{cases}.$$

In order to better the region extraction to obtain the gesture area accurately, grey projection method is used in the paper. Specific means is to use the grey projection method in the vertical direction and the horizontal direction respectively, so as to obtain the gesture coordinate range in X direction and Y direction. As shown in Figure 4, (a) RGB image is captured from real-time video in the hand, (b) is corresponding to the processed image of (a), (c) is hand region determined by grey projection.

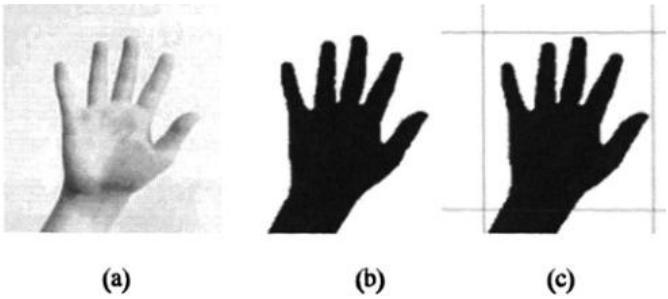


Figure 4. Hand region using grey projection method

2.2. Hand Tracking

Camshift is the Mean Shift algorithm of continuous adaption [8]. It is the extension and improvement of Mean Shift algorithm, which can handle dynamic changes of target tracking. When the video sequence changes frame by frame, the Camshift automatically adjusts the search window size and position the tracked target. With the current known localization results, it can predict the target position in the next frame. This iteration of each frame image leads to dynamic tracking. The experimental results show that, CamShift algorithm does not have high requirements on system resources. It has better real-time interactive scene and robustness in the actual performance.

2.3. Real-Time Gesture Recognition

Gesture library is created firstly in gesture recognition. Image pre-processing and feature extraction are done to obtain ideal gesture segmentation effect; and moment characteristic components are used as the feature vector of gesture image. Then the algorithm of support vector machine is used for

the classification and recognition of hand gesture image, meanwhile, several common multi-class classification algorithms of support vector machine, multiple classification algorithm based on the posterior probability, and several traditional multiple classification methods are applied to gesture recognition, and the results are compared.

In order to search for the best model of support vector machine for hand gesture recognition, we have done many experiments, comparing the use of a variety of multi-class classification method and the classification accuracy of different kernel function. The experimental results are shown in Table 1. In the table, The "C" is the penalty parameter; "b", "d", "r" is the kernel function parameters. When testing the support vector machine classifier, the Hu moment feature vector of test samples are used as inputs, according to well trained model file, svmpredict application is used to do the classification and output and test sample's category.

TABLE 1. The classification accuracy of training set and testing set

Kernel function	Classified	The kernel function parameters	The training sample recognition rate	The test sample recognition rate
Polynomial kernel function	"one to one"	C=24,r=2-2,b=1,d=3	95.4%	95%
Polynomial kernel function	"one to many"	C=23,r=2-3,b=1,d=2	95.9%	95.2%
The radial basis kernel function	"one to one"	C=210,r=2-1	96.1%	95.6%
The radial basis kernel function	"one to many"	C=29,r=2-2	97.1%	96.5%
Minimum distance algorithm			84.2%	80.8%
Neural network algorithm			92.4%	87.8%

Conclusions are obtained from the above experimental results:

The SVM classifier outperforms the minimum distance method and neural network method, which shows that SVM is relatively stable and efficient classifier, and has good generalization ability and reflects advantage in SVM in the small-sample pattern recognition.

Using a variety of multi-class classification methods has little difference in classification accuracy, which is more than 95%. In contrast, the classification accuracy of radial basis kernel function is slightly higher than the polynomial kernel function. The identification of the training set and testing set rates are above 96% and 95.5%. The stability and the various radial kernel multi-class classification method is slightly better than other kernel function. "One-to-many" multi-class classification accuracy is generally higher than the "one to one". Especially the "one-to-many" identification of radial basis function classifier is the highest in all SVM classification models.

Because the parameters of radial basis kernel function of support vector machine is the only "r", which is the least in four kinds of kernel functions, such SVM parameters to be determined in the training process are only "C" and "r", which is easier than polynomial kernel function.

According to the above experimental data and conclusion, in the hand gesture recognition algorithms, we use the support vector machine classifier based on "one-to-many" radial basis kernel function to do gesture classification.

3. Application of Hand Gesture Tracking and Recognition in Human-Computer Interaction

In order to validate the gesture tracking and recognition algorithm, a simple Word digital input program is written based on Visual C++6.0 platform. The gesture recognition is used in human-computer interaction, realizing the operation of the Word document from camera input gesture.

The use of predefined digital gestures 0-9 realize the implementation of Word document operation. We defined in the program: before the word document open, digital sign 1 means click the left mouse button, digital sign 2 means Click of the right mouse button, digital sign 5 means double-click of the left mouse button, so as to realize the simple control of the mouse gestures; after Word document open, 1, 2, 5 Digital gesture mouse click or double-click event is no longer valid, the digital gesture 1-9 means the inputs of number 1-9 in Word documents, namely automatically write the number corresponding to the gesture in word documents through gesture recognition results. However, the digital gesture 0 is always defined as the gesture of closing Word document, so number 0 cannot be inputted in this system.

This system captures gestures and recognition through the USB camera, thus realize the simple operation of the Word. The functions of the gesture interactive system are shown as follows.

Open the Word document functions: firstly, use the relative position information module obtained from tracking module to simulate the mouse mobile event, move the mouse to the desktop shortcut Word,

then use the recognition result of digital gesture 5 to simulate mouse double-click event and open the Word document.

Digital input function: Since we stipulate that the mouse click or double-click events of digital gesture 1,2,5 will be invalid when Word opens, at this time, the system puts the recognition result of digital gesture inputted by camera (i.e., the number corresponding to the digital gesture input) into Word document, so as to realize the digital input function.

Save the document and close the document: after completing digital input, use gesture 0 to close the Word document. The "whether to save the changes for document 1" dialog box would pop up when you close Word, at this time, the input digital function will be stopped and the mouse click or double-click events of digital gesture 1, 2, 5 will be activated through programming.

4. Conclusion

The human-computer interaction based on human gesture recognition is an important research subject in current intelligent human-computer interaction. The gesture recognition based on vision is to use the camera to capture the gesture information, which do not need to buy expensive equipment, and it is more natural, convenient in operation, and people-oriented. Therefore, the study on this problem has important theoretical value and application prospect. In the Microsoft Visual C++6.0 development environment, using OpcnCV open source library, gesture tracking and recognition algorithms of different stages based on computer vision are studied and validated. A simple Word digital recording system is developed, combining the gesture recognition in human-computer interaction. This research still needs to be improved. In the hand gesture tracking and real-time recognition, we also need to increase the kind of gestures, and further improve the recognition speed and accuracy.

Acknowledgements

This work was financially supported by Zhengzhou Key Laboratory of embedded System Application Technology, China (Grant No. 2011C520017) and the Natural Science Research Project of Huanghe Science and Technology College, China (Grant No.KYZR201203).

References

1. Sadka, A.H, Crookes, D. (2010) Multimodal biometric human recognition for perceptual human-computer interaction [J]. *IEEE Transactions on Systems, Man, and Cybernetics - Part C：Applications and reviews*, 676-681.
2. Hang, Z., Qiu-Qi, R. (2007) Bare-hand alphabets gesture recognition based on VCM and ROI segmentation. *Journal on Communications*, 94-101. (In Chinese).
3. Yu-Jin, L., Yong, C., Hui-Yue, W., Feng-Jun, Z., Guo-Zhong, D. (2009) Approach to tracking deformable hand gesture under disturbances from skin colour. *Computer Engineering and Applications*, 164-167. (in Chinese).
4. Guofan, H., Xiaoping, C. (2009) History-Based Dynamic Hand Gesture Recognition. *Journal of Southwest University (Natural Science Edition)*, 106-110. (In Chinese).
5. Li, Z., Jarvis, R. (2009) A multi-modal gesture recognition system in a human-robot interaction scenario. Proceedings of IEEE International Workshop on Robotic and Sensors Environments. Lecco, Italy: *IEEE Instrumentation and Measurement Society*, 41-46.
6. Kaiser, A., Schenck, W., Möller, R. (2012) Stereo matching and depth perception by visual prediction. Proceedings of SAB Workshop on Artificial Mental Imagery. Odense, Denmark: *IEEE*, 7-10.
7. Katkere, A., Hunter, E. (1994) Telepresence using Hand Gestures Technical report VCL-94-104. Visual Computing Laboratory, University of California, SanDiego.
8. Zhen-Jun, Y., Wang-Cheng, Q., Chun-Lin, L. (2004) A Self-Adaptive Approach of Multi-Object Image Segmentation. *Journal of Image and Graphics*, 674-678. (In Chinese).

Received on the 10th of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 196-199
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

RESEARCH OF INTELLIGENT RETRIEVAL BASED ON THE SEMANTIC WEB

N. Zhang

College of Information Science and Technology, Donghua University,
Shanghai 201620, China;
Department of Computer Science and Engineering, Henan University of Urban Construction,
Pingdingshan Henan 467036, China;
Phone:15937565010, E-mail: znwang1010@126.com

The disadvantage of the traditional information retrieval system is that it disregards the semantic information of the keywords itself, so the efficiency of the research is low, the author proposes a new intelligent information retrieval system based on the Semantic Web and computes the effective Semantic comparability, furthermore, it introduces an ideology of the Object Oriented Design into the domanial Ontology design and implements the semantic retrieval on the electronic business website. The practice proved that this semantic retrieval could research the correlation with the keywords and make the information retrieval entirely and precisely.

Keywords: semantic retrieval, the Semantic Web, Ontology

1. Introduction

The traditional information retrieval system based on the keywords doesn't usually meet the user's demands because it doesn't express the retrieval requirement completely[1], for example, firstly it identificate the website with isolated keywords which can't express the logical meaning of the documents. Because it usually research the irrelevant information，its efficiency is very low；secondly, the retrieval based on keywords which has not the semantic support does not search the relevant information, such as synonymous and similitude conception. The Recall Ratio is low; finally, people has to browse hundreds and thousands of results one by one which takes much time and strenuosity, this is so called "Information Overloading" [2, 3]. Aimed at these questions, the article processes the mechanism of semantic information retrieval based on Ontology on the Semantic Web.

2. The Semantic Ontology

The Semantic Web adopts Ontology as the carrier of information contents, the Ontology is regarded as an explicit formulation conceptually which defines a group of typical conceptions, these conceptions and the relation of which describe the domain world[4, 5, 6, 7, 8]. This article carries out the work in the domain Ontology of electronic business, it expands the semantic Ontology, for example, first, when the users not only obtain the information included in the keyword itself but also get hold of the same meaning as the keywords when he inputs the keywords, such as "DB" has the same meaning as the database, no matter which the user inputs, it will show the information of the both. With respect to the retrieval based on keywords, the semantic retrieval moves up to the knowledge and semantic level. This condition implements the synonymous retrieval; second, the semantic retrieval can retain the gradation research, in the electronic business website, DB is belong to the computer in book domain. In the system of Ontology, it is belong to the up and down relation. If the user want to lookup the information about DB, he can also input the word "computer". This condition is described as the retrieval based on the class and subclass.

Above all, it introduces the semantic Ontology which implements the extended retrieval. The semantic Ontology can improve the recall ratio and the precision ratio. In the domain of Ontology, the description of the information resources is defined with the concepts and property of Ontology and defines as RDF instance document. The semantic retrieval system can locate the information resources according to the conceptions and the property of the conception, which the users input. For example, this is the description about the book:

<book rdf:ID="the Website of the Datebase">
<book:property rdf:resource="#keywords">
<book:property rdf:resource="#ISBN">

<book:fullDesc rdf:datatype=”http://www. w3. org/2001/XMLSchema#string”>This book describe the fundational theory, basic technology and elements</book:fullDesc>

<book:relatedbook rdf:resource=”#DB”>

<book:relatedbook rdf:resource=”#Computer”>

<book:bookname rdf:datatype=”http://www. w3. org/2001/XMLSchema#string”>the conspectus of the DB</book:bookname>

The implementation of semantic retrieval constitutes the foundation of the semantic Ontology, which can impress the concepts and the relation of the concepts accurately. On a certain extent, Ontology has the ability of understanding. On the point of view on philosophy, “Relation exists forever.” The extent, which Ontology describes, determines the range of the application and the goals obtained, from this point of view, Ontology is infinite great and it can describe the objective reality.

3. Created Ontology by Object-Oriented Method

Nowadays, there is not an efficient and uniform method to create the Ontology in the Semantic Web, on the foundation of rational created methods, it is introduced the Object Oriented technique into the field Ontology created. This article process a new method to create the Ontology-Object-Oriented methods. On the one hand, it maps the classes and its hierarchy in the Object-oriented system into Classes and its hierarchy in the Ontology, maps the relation of the classes into the relation in the Ontology, and maps the poverty of the classes into Ontology. Furthermore, the author applies the method into the system of semantic retrieval. The concrete approach is below:

(1) Define the classes and the its hierarchy, The concept Hierarchy which is constructed and the hierarchy of concepts which is abstracted from the knowledge of the field directly map into the concept hierarchy of the Ontology, and the real semantic conception relation in the hierarchy tree is not considered. Fig 1 defined the class and its hierarchy in the literature.

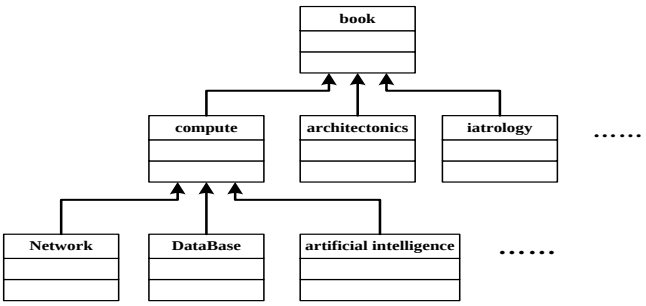


Figure 1. Classes and its hierarchy

(2) Construct the Relation Hierarchy among the concepts. By means of intercalating the domain the relation, the relation of concepts is restricted. From the angle of interactional relation, it embodies the practical semantic of HC. The figure shows the relation of the classes and the poverty of the classes. Fig 2 show the relation of the classes，Fig 3 show the poverty between the classes.

Figure 2. Classes the it’s relation

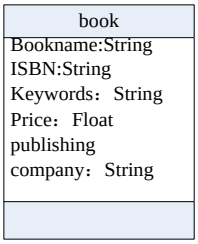


Figure 3. Property of classes

(3) Define the synonymy and near synonymy relation, it mainly defined by means of Protégé, such as:

<owl:Class rdf:ID=”Network”>

<owl:equivalentClass rdf:resource=”&computer Network”/>

</owl:Class>

The Orient-objected method has characteristic and significance, such as below:

(1) It is easily learned and used, reducing the participation of the field experts. So far, the orient-objected method and thought has already developed maturely and has applied in many fields. Therefore, this method is easily accepted by software developers who embark upon the Object Oriented Design.

(2) Using this method can decrease the amount of work of the system, when developing the semantic retrieval system based on the Ontology, the prior-period demand analyse and database design has intercrossed with the process of method implements. Therefore, this method can decrease the amount of work of the system.

(3) This method make the robotization of the Ontology created impossible. A large member of concepts of the Object Oriented Design have related with the definitions of the Ontology. Therefore, API which is applied in the Object-Oriented modelling means possibility resolve the robotization of the Ontology.

4. The Process of the Semantic Retrieval

Firstly, you must create the Ontology in some fields; Ontology exists as the RDF txts, which is called Ontology knowledge database. Secondly, you must semantically label the information resources by the concepts defined in the Ontology and the relation of the concepts, and form into instance documents. Because the semantic retrieval system aims at the unity documents which base on XML and describe the information resources by RDF. At the same time, we combine the metadata in the meta-database to form into the instance document. These metadata, the concepts and its poverty is unity. On the other hand, we label the resources automatically by Protégé, and this process is understand by computer. Thirdly, this system adopts the B/S model, the users input the requests in the Internet Explorer, Application program receives the requests, subsequently it query and match into the Ontology meta-database, which is the extended Ontology database. Ontology have reasoning ability, so it can implement the semantic retrieval. The process of the semantic retrieval is showed at Fig 4.

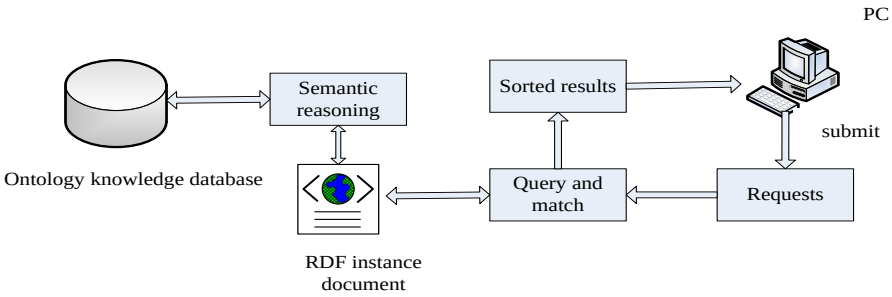


Figure 4. The process of the semantic retrieval

5. The Interface of the Semantic Retrieval

The semantic retrieval system is developed in the WEB, which implements two kinds of query methods, such as:

(1) conditioning query, Users input the query information that they want to search, after that, the system will return the results .

(2) navigational query. At some time, Users cannot make sure the retrieval conditional expression, Therefore, the system provide the flexible navigational query, namely, Users have not necessity to input the query expression, they can query according to the classification of the books.

The interface of the semantic retrieval is showed in Fig5.

Figure 5. The interface of the retrieval system

6. Conclusion

The paper advances a retrieval method based on the Ontology, it introduces the Ontology into the system, and adopts the Object-Oriented method to create Ontology and label the information resources automatically. Ultimately implements the semantic retrieval system, because the Ontology can extend the semasiology by the semantic reasoning， the system improve the Recall Ratio and Precision Ratio, compared with the traditional retrieval system based on the keywords.

References

1. Rodriguez, M., Egehnoefr, M. (2003) Determining Semantic Similarity among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2), 442-456.
2. Navigli, R., Velardi, P. (2004) A knowledge-based approach to ontology learning and Semantic annotation. *Proceedings of the16th International Conference on Advanced Information Systems Engineering*, *CAISE.* Riga, Latvia, 1-15.
3. Navigli, R., Velardi, P. (2006) Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics,* 30(2), 151-179.
4. Grigoris, A., Frank, V. Web Ontology Language: OWL. *International Handbooks on Information Systems*, 1(4), (2004), 67-92.
5. Eero, H., Kim, V., Jouni, T., Seppälä, K. (2008) Building a National Semantic Web Ontology and Ontology Service Infrastructure-The Finn-ONTO Approach. *The Semantic Web: Research and Applications Lecture Notes in computer Science,* 5021(2), 95-109.
6. Peter, M. (2005) Ontologies Are Us: A unified Model of Social Networks and Semantics. The Semantic Web-ISWC 2005 Lecture Notes in Computer Science. 3729(2), 522-536.
7. Anupriya, A., Markus, K., Thanh, T., Denny, V. (2007) Proceeding of the 16th international conference on world wide web, 1(1), 825-834.
8. Scott, F., Terry, L. (2003) A linguistic ontology for the semantic web. *GLOT*. 7(3), 1-4.

Received on the 10th of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 200-203
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

FOUR REMOTE MONITORING POWER SYSTEM DESIGN

Jingxia Liu, Guimei Cui

Inner Mongolia University of Science and Technology Information Engineering Institute
Baotou, Inner Mongolia, China
E-mail: 958428915@qq.com

"Four remote monitoring" means power system telemetry, remote communication, remote control and remote setting, this design is based on laboratory equipment of power system lab to simulate the actual power system to realize the four remote SCADA system through the force control 5.0 configuration software. It is made use of the configuration software to make monitoring interface so as to realize the four remote monitoring of small grid. Realized the four remote functions of analogue power system consisting of "THLDK - 2 monitoring experiment platform" and "THLZD - 2 integrated automation experimental plat-form", included the real time measure in all kinds of data, monitor and the record with all kinds of information, such as remote communication remote control with circuit breaker state, telemetry and remote setting of generator operating parameters, alarm records, history reports, etc.

Keywords: configuration, Four control, Remote monitoring

1. Introduction

"Four remote monitoring" is an integral part of modern electric power system. It is consisted of master station system installed in the dispatch centre of, remote terminal unit and remote access installed in a power plant or substation, etc. its main function is to collected operating parameters and information of power system in real-time on-site, monitor and control continuously.

Early power system dispatching, it was mainly rely on telephone contact be-tween dispatching centre and every plant station, the real-time performance for scheduling was bad.

After using the telemechanique of power system, the telecontrol equipment of the end of the plant stand transmit the telemetry and the remote communication information to the dispatch centre in real time, these information can be displayed visually in screen and simulation screen of the control centre, it can make the operator see the real-time running parameters and the operation mode of the system at any time, Realized the effective monitoring for running state of the system. When needed, the dispatcher can operate in the dispatch centre, complete to transmit the remote control or remote adjustable to the device of plant station. Due to generate the information in telecontrol equipment, the speed for transmission and processing is very fast, adapted to the real-time demand for power system to scheduling work, it made the dispatching management of electric power system enter into the automatic stage.

The telecontrol system of dispatching automation system consisted of the remote master station, remote terminal RTU and channel.

Remote terminal (RTU) cooperate with master station can realize the "four control" function, namely:

1) telemetry: collect and transmit the real-time parameters of power system operation.

2) remote communication: collect and transmit the action information of relay protection, status information of circuit breaker, etc. in the power system.

3) remote control: from the dispatch centre issued a command to change the condition of operating equipment

4) remote setting: from the dispatch centre issued a command to realize the distance adjusting operation parameters of power plant or transformer substation

The laboratory of Power system is professional teaching laboratory for electrical engineering and its automation, the equipment of the laboratory can simulate a small network for power system, from power generation, power transformation and transmission to distribution and utilization. This provides a more comprehensive teaching experiment platform; it can helps students understand the structure of power system in-depth through the experiments of power system.

The laboratory has been opened to the electrical professional majors. the open experiment: "four remote" experiment of power system is necessary to do for students each year to understand the monitoring knowledge of power system, but the existing monitoring software of the laboratory is not perfect, in order to make the teachers improve the teaching quality and professional level for specialised

experiment, students can realize the theoretical analysis and experimental analysis, now is the interface design with more targeted, more professional about "four control" monitoring software in power system. from the real-time monitoring to the experiments can be showed in the superior machine demo operation.

2. The Composition of the Power System Network

In this design, the network structure of power system consisted of four "THLZD - type 2 power system integrated automation experimental platform" is equal to the power plant in actual power system, one "THLDK - type 2 power system monitoring experiment platform" is equal to the dispatching centre in actual power system. Of the laboratory for electric power system in Inner Mongolia university of science and technology of 1, article 6 power transmission lines of different length and 4 different power load, three winding transformer and the infinity of simulation system of multiple autotransformer, etc., the main electrical wiring forms as shown in figure 2.1:

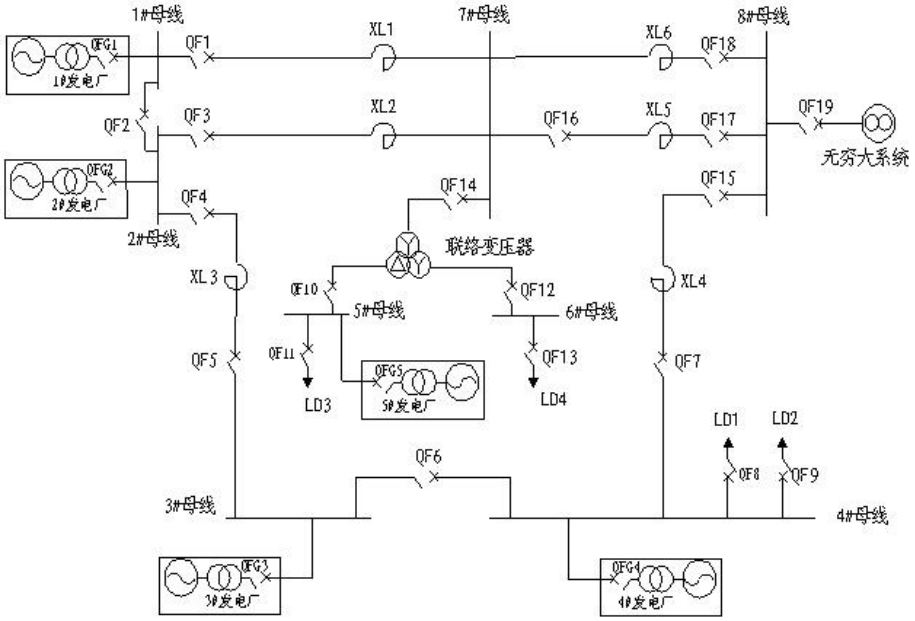


Figure 1. The main electrical wiring forms

3. The Design Method of Configuration

Interface and graphic design is finished in the system development environment in 5.0 force control configuration software, the system development environment is the working environment that a design engineer must rely on for carry out the control scheme, with the support of configuration software for the application of the system-generated work. Through establishing a series of user data files, to produce the final graphic target application system, used for system running environment runtime. System development environment is composed of several configuration program, such as graphical interface configuration program, real-time database configuration, etc. Force control main graphical interface with Windows 5.0 system document structure, provides the convenient and flexible graphing tool, can draw nice graphics.

4. The HMI Interface Design of Four Remote Monitoring Of Power System

The design process for HMI (human machine interface) of Four control power system simulation can be divided into: the drawing and layout for interface, the definition and set of I/O devices, engineering point and variable definition, data connection, animation, preparation of action script (programming), machine commissioning and so on.

5. Telemetry, Remote Communication, Remote Control, Remote Setting Interface

Force control 5.0 configuration software is divided into development system and running system. In this section, I will introduce mainly the operation of the power system for four remote monitoring software, namely the function and interface con-figuration. The monitoring object is the electric field that simulated by "THLDK - 2 monitoring experiment platform" and "THLZD - 2 integrated automation experimental platform". Its need to implement, the real-time measurement for a variety of data, monitoring. The data including: the power flows, voltage, active power, reactive power, generator RPM, frequency, etc. In addition, this software also need to make remote control to the circuit breaker by in the power site simulated by "THLDK - 2" experiment platform and "THLZD - 2 integrated automation experimental platform". Operators can make remote control to each power plant and power line only through the software, to achieve the "four control" function of power system. At the same time, depending on the alarm window, history report, remote communication, and other configuration module, we can also make the operating records, data record and query history to the electric field simulated by lab "THLDK - 2 monitoring experiment platform" and "THLZD - 2 integrated automation experimental platform".

6. The Remote Control Function of Power System

Remote control: remote control or protect the open-shut, start-stop state, etc. of the electrical equipment and electrical mechanization.

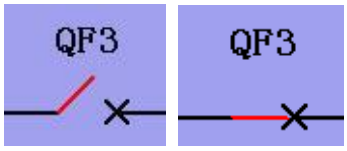
Remote setting: setting for distance and adjust the working parameters, standards of the equipment it controlled.

Remote control not only relates to the safety analysis of the power system, but also relates to the economic dispatch of power system, it has the profound significance.

Power system security analysis is to analyse if the power system in running state can continue to maintain the normal operation in the event of an accident. First function of the security analysis is to determine if the current operation state is safe in an accident. The so-called incident is the assumed incident that based on the experience of operating personnel. The results of these accidents, either make one or several electric power line disconnect, or make the transformer, generator, load disconnect, or the combination of the above happens. To determine whether it is security for the system when an accident is to calculate through the computer whether load or exceed the permitted transmission cable pass the limit in various assumptions above accident occurs, if there is a node overvoltage, if the system will lose stability and so on. If it will not appear the situation, the current running state of the system is safe, otherwise it is not safe. The second function of the security analysis is to confirm the control measures for keep the safe operation of the system. If the results of safety analysis found that the current running state of the system is unsafe during a particular accident, the safety analysis should be put forward to the calibration, adjustment and the control measures for keeping the safe operation of the system .

Economic dispatch is one of the important task of power system dispatching management, its purpose is: under the condition of safe operation, make the equitable distribution to the generator of power system active power load, and energy saving to the largest extent, make the lowest cost of power supply. The dispatching automation of power system, it should make computer give full play to the strong point in data processing, to make the work for economic dispatching more and more detailed, and saving more energy. The economic operation of power system is not only to adjust individual plants, but also realized by unity dispatching for power system. The economic dispatch of generator set not only need to optimization allocation to use the scientific method, but also needs to have a monitoring and evaluation to ensure that the optimized scheme and it needs to supply accurate performance parameters for optimization analysis. Therefore, realize the economic dispatch of generating set, need to work on the following aspects: (1) the unit output for optimal allocation to make the total power generation efficiency is the highest or the lowest cost; Generators, and (2) real-time monitoring unit output make the actual output is within the envelope line of program output; (3) count up and check unqualified power for each operation team, helps to raise the operation level with economic responsibility system; (4) establishing the coal consumption rate and the electricity rate curve for optimization analysis.

The remote control of Power system is implemented in master scheduling system interface, in the interface; you can see the points close state for circuit breaker in actual operation, convenient for analysis for the user, graph QF_3 circuit breaker switching state and brake state with standard electrical symbols.



The details for HMI (human-machine interface) drawing connect process simulated by “Four remote monitoring” system, including the overviews of the design and function of each interface, the establishment and the configuration of real-time data-base, flash and animation compilation, here no longer described in detail.

7. Conclusions

The design for four remote monitoring of power system is developed based on the Beijing sunway 5.0 configuration software, and realized the "four control" function of power system simulated by power system lab. In addition to the traditional "four control" function, in this design, also added the function of the circuit breaker operation send text messages and facilitate the dispatch centre of real-time monitoring function of remote viewing based on Internet network.

Due to the design objects is "THLZD - type 2 power system integrated automation experimental platform" and "THLZD - type 2 power system integrated automation experimental platform" of electric power system laboratory in our school , these two kinds of experimental equipment of communication adopts standard MODBUS protocol (RTU transmission mode), so the communication protocol in the design is MODBUS communication protocol, adopting RS485 serial interface for information transmission.

This monitoring design supports the secondary development; the user can be modified according to real need.

References

1. Beijing sunway configuration software company. Force control 5.0 configuration software specification (HMI interface section). Beijing: Sunway science and technology, (2005) （in Chinese).
2. Shandong create technology co., LTD. EDA9033 three-phase electric parameters data collecting module - product manuals. Shandong: Make science and technology (2006) (in Chinese).
3. Beijing sunway configuration software company. Force control 5.0 configuration software specification (function). Beijing: Sunway science and technology (2005)（in Chinese).
4. Beijing sunway configuration software company. Force control 5.0 configuration software specification (control strategy). Beijing: Sunway science and technology (2005)（in Chinese).
5. A Beijing sunway configuration software company. Force control 5.0 configuration software specification (database/communications section). Sunway science and technology Beijing: Sunway science and technology (2005) (in Chinese).
6. Beijing sunway configuration software company. The standard MODBUS setup instructions. Beijing: Sunway science and technology (2007) (in Chinese).
7. Zhen, Y., Researched for some problems of remote monitoring for electric power system [D].
8. Yong, W. (2010) Based on the digital substation technology of IEC61850-9-2 and GOOSE network transmission [J], electric power system protection and control, (12), 131-132.（in Chinese).

Received on the 20th of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 204-209
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

RESEARCH OF AN NEW AC-C/S-BASED ELECTRONIC DOCUMENT SECURITY SYSTEM

X. Xu

School of Mechanical&Electrical Engineering
Changchun institute of Technology, Changchun, China
Phone: (86)431-85711304. E-mail: xuxuedong1000@126.com

An AC-C/S (Anti-Crack Client/Server) architecture model is presented in this paper to overcome the software crack risk at the existent DRM clients with anti-crack improved by the shell. The AC-C/S comprises the framework and key components transferred to be executed at the remote server, both interacting through the data stream technology. The realization of the electronic document system and its application is described to show that it has improved the security and flexibility at DRM clients.

Keywords: Electronic Document, Digital Copy Rights Management, Anti-Crack, Data Stream

1. Introduction

All documents have become electronic with the rapid development of computer and network. Moreover, their copyright and security technique also develops rapidly, among which is the typical technique of Data right Management (DRM). This technique is effective in preventing unauthorized use or illegal acquisition of electronic files, for example [1, 2], such as Adobe Content Server, InfoGuard and Unnoo. These products comprises client, licensing server and content server etc. First, the file is encrypted by packaging tools with the use authority and the key stored in the license. The user must acquire the license from the corresponding licensing server for their downloaded documents from the content server [3,4]. Such DRM file security products have the intrinsic demerit that the authentication and decryption are executed on the user's computer. Once the software were cracked, there would be no protection for the files.

Therefore, ensuring the security of DRM system is important when the software is relatively fragile at client. Based on the popular network and improved bandwidth, the present paper has proposed a new approach in protecting DRM software at the client. The program was disassembled into a frame component and a key component. The software could be normally executed through the real time communication between network interfaces for the key parts transferred to remote security server.

Logically, cracker must establish the architecture transferred to crack the software. The cracking is even more difficult than programming a new software with more interaction in frequency and data between the separated software. For lower crack risk at the client, data stream transportation technique is required for the necessary processing of large amount of data.

Based on such a design philosophy, we have defined a new system architecture on the basis of C/S, named "Anti-Crack Client/Server, AC-C/S". In addition, based on it we constructed electronic files security management system to safeguard the system and files (documents) through remote processing of the key part and decryption module.

2. AC-C/S Anti-Cracked Architecture and its General Design

2.1. AC-C/S Architecture Model

Researchers have recently noticed the system failure risk [5] brought about by the fragility at clients in DRM system. Although there are many ways of protection such as code encryption and shell technique, malicious cracker have the chances to analyse the plan for software protection and cracks clients through ways such as decompiling, inserting skip command to skip the protection mechanism [6].

We argue that the key for the system security is that the vital components run on the remote server to get rid of the cracker's control. There is however a technical difficulty to integrate the separated parts and have high efficiency. First, there must be a communication protocol between operating systems, different interfaces and firewalls and we have chosen the "RPC Remote Procedure Call Protocol" as an ideal one. Second, there must be a data structure-processing model for real time interaction of great

amount of data. Based on our previously studied technique for data stream processing, we have designed a Data Streaming Operation Tree (suitable for execution and processing in RAM) and query-operation tree dynamic scheduling algorithm by incorporating the compression coding and Sketch. This design solved the problems of the space limitation and query speed in data processing. Therefore, we chose the data stream technique as the interaction model for communication between the modules.

Based on the study above, we modified the C/S Architecture by disassembling the client into a front part (for user's interface, data stream integration and bottom level monitoring) and core part (for the core job), with the former stored at the clients and the latter at the server. The RAM vacancy at the clients left after the transfer is filled with stubs and data stream. We defined this derived architecture for preventing cracked as the AC-C/S (shown in figure 1).

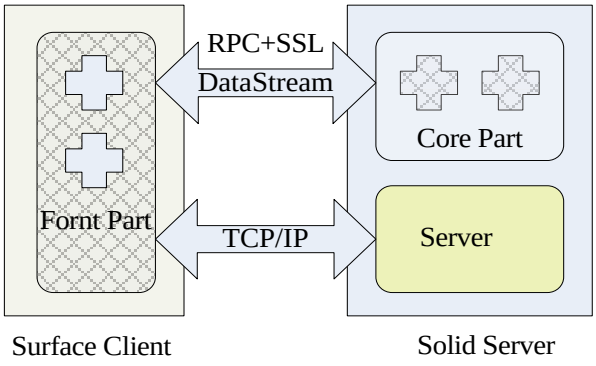


Figure 1. SC/SS Architecture Model

AC-C/S Architecture could be represented as four tuples:

$$SS = \{SC, SS, CP, DM\} \quad (1)$$

in which :

$$SC = \{\text{Front part}, \text{Data stream interface}\} \quad (2)$$

SC is defined as surface client, which comprises model, stubs filing the vacancy and data stream interface at conventional clients.

$$SS = \{\text{Core part}, \text{Server}\} \quad (3)$$

SS is defined as Solid server, which comprises core part and disassembled part from clients at the ordinary server.

$$CP = \{\text{RPC}, \text{TCP/IP}\} \quad (4)$$

CP represents Communication protocol. Multi-threaded RPC was adopted between the disassembled parts from clients. The call interface is SOAP protocol, adopting two-way SSL for communication. TCP/IP was adopted for ordinary interaction between clients and the server.

$$DM = \{\text{Data stream pattern}\} \quad (5)$$

DM represents the data model between two disassembled parts, realized by process-based real time data stream model.

2.2. General Design of AC-C/S-based Document Security System

2.2.1. System Requirement

The system should provide the requirement for document security management long life cycle from its creation to destruction, such as the setting up of the access policy and reliable encryption to prevent unauthorized access to and modification of the copy and print, including:

1) In the network environment, administrator manages security levels and authorities, and copy the files and can read or modify them under the control of the system. All the files would be in cipher text format once away from the system.

2) DRM encryption depends on operating system, not with the format or programs. The operating system for the encryption is Windows XP (vista) at the clients and Windows2003 server at the server.

3) The system has high reliability, and strong power against crack. It monitors and records the clients' operation.

2.2.2. The General Architecture

The System is based on AC-C/S, comprising surface clients and solid server.

Surface clients execute interaction surface, authority identification data stream interface, sends to the security server information collected about the clients, and monitors the uses of the electronic files.

Authentication, granting access, user's behaviour audit, file decryption and client's key modules are executed at the Solid server.

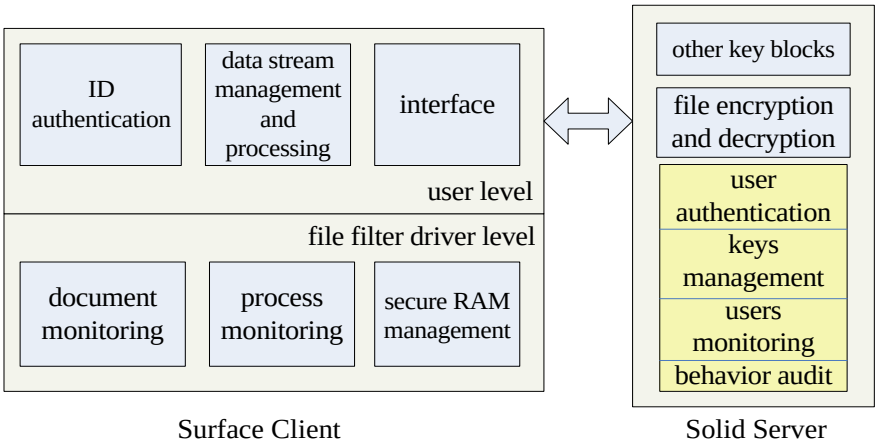


Figure 2. Block diagram of the security systems

The encryption module and some important lower-layer functions of the documents at the DRM file clients are packaged and transported to the server. As there is no decryption algorithm and keys at clients, the files decrypted at the server are sent to RAM and opened at clients. The RAM used for the files are locked and accept no access.

The reading or writing of the files are controlled at the bottom level of the system by file filter driver at the client to protect all kinds of files such as the files in format of pdf or doc [7]. At the same time, the clients are monitored for their operation or block any unauthorized operation, such as copying, reading and modification of the files. Clients sent the detailed description to the security server for query and alert against malicious behaviour.

3. Key Techniques for the Realization of the System

3.1. Realization of the Transfer of Key Module at Client

Remote encryption and decryption process of DRM files is shown at Figure 3.

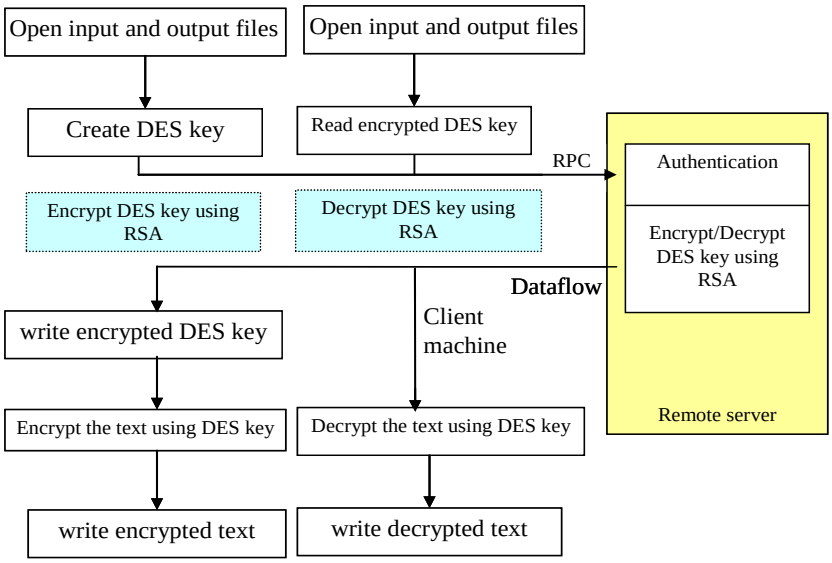


Figure 3. Remote encryption and decryption process

Encryption module is executed at remote server. After they passes the authorization, the surface clients start to sequence the files specifically in the form of stream, transfer them through RPC after compressing them into byte stream [8]. Upon receiving the data stream, the server generates query-operation tree in accordance with special connection management policies, and executes encryption algorithm after dissembling the data blocks. After decryption, the data is transferred to the RAM-area secured in SSL encrypted stream through the established security channels at clients. Surface clients counter-sequence the data in the locked security area, add load dropping operators to realize efficient load dropping by using frequent items algorithms in the data stream. This ensures improved file response rate at the clients for the netflow limit (for this technique, please refer to our other papers). In this model, the client could not acquire the authority to process the files before passing the authentication.

As the key to the encryption and decryption is on the remote server, the client could not process the files at clients independently of the remote server. Thus, the cracked DRM is logically blocked at the client.

At the same time, we transported many of the bottom level modules at the client to the server and runs in the above mechanism. Logically, cracker must rebuild the key parts at the client to crack the system at clients. This ultra-difficulty greatly lowers the possibility of cracked risk.

3.2. The Encryption and Decryption System

To lower the risk of misconduct and leak out by insiders, the system adapts a management strategy of "Power Sharing" [9]. We have designed a multi-level encryption system with keys in many categories for data encryption at customer's end through SM4 encryption card. The keys comprises start key, management key, restoring key, local key and base key (as shown in Table 1).

TABLE 1. System Key Categories

No	Categories	Type	Uses
1	*Start key*	*Asymmetrical*	*System starts key, protecting management key at two levels.*
2	*Management key*	*Asymmetrical*	*System management key, restoring and decrypting keys.*
3	*Restoring key*	*Symmetrical*	*Encryption key for local key, protected by key encryption.*
4	*Local key*	*Symmetrical*	*Encryption key for base key, protected by restoring key.*
5	*Base key*	*Symmetrical*	*Data encryption key, protected by local key*

Encryption Processing Flowchart is shown in Figure 4.In the encryption and decryption system, data is encrypted in the storage processors and sent to the storage equipment. The base key generates a

K1 through encryption by local key, which generates K2 through decryption restoring local key, which generates K3 through decryption by restoring key, which generates T through primary start key. Secondary start key encrypts T into TT. The K1, K2, K3, T and TT together are stored in the parameters databank of system.

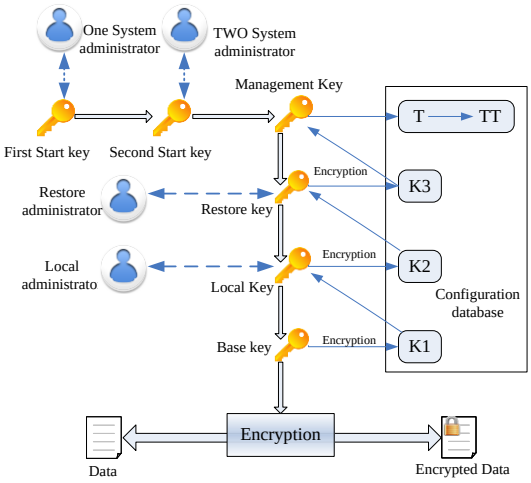


Figure 4. Encryption Grade System in SecureSAN

In this encryption system, the emphasis of protection is that the password stored in the key rather than in the data, in the base key is encrypted at multiple layers. This disperses the management power and lowers the risk of leak out and simple the steps of data encryption but with better functions.

4. Evaluation and Application of the System

4.1. Application Flowchart of the System

First, the administrator packages the DRM files, and stores access control privileges and keys at the server. The cipher text is sent on-line and off-line to clients.

Client logs into the security server through the client, and establishes security channel after ID is authenticated by user's key and certificate at client. Surface client sends the cipher text to the decryption module at the remote server and gets the plain text in the data stream at the client. The plain text is stored in secure RAM area with exclusive lock protection to prevent the call from external programs while the client is in the state of being monitored at bottom level. The client can have the call operation within the privilege while the software at the client monitors the operation and sends operation records to the software at security server.

The files are closed after the operation with the plain text at RAM automatically deleted and cipher text still at the client computer. In the case of creating a new file or re-editing a file, the remote server would arrange a new key and encrypt the new file and sends the file back to the client as a new edition of the file encrypted and stores it as a duplicate at the server before the saving starts with the packaged files [10].

The communication is realized through the two-way SSL security, which encrypts the channel to safeguard the security of the cipher text and plain text. Separated from the support from the server, the client could not independently decrypt the file. Thus, there is no possibility of cracking the cipher text.

4.2. Evaluation of the System

This technique was tried in a designing institute in Jilin Province. The run environment was the 100M LAN, with 2 Dell 5550 servers and 10 client computers. One server disposed security server system installed with a Dean Encryption System. The other server was used as a files server. The 10 client computers execute requests for opening, re-editing and saving of the files. This trial application showed that the designed aims were realized with the average time delay of 7.7 seconds, which was proved to the time for ID authentication and file decryption by monitoring the netflow at the remote

server. It is thus proposed that a better hardware is required for better and more satisfactory performances.

5. Conclusion and Future Work

In conventional DRM file security products, the decryption of the files and the ID authentication are both realized at the clients. Once cracked, DRM files are no longer protected. This paper proposed a new AC-C/S, which transfers the key parts to the remote server. The trial application of security system based on such an architecture was proved to solve the above two problems with ensured performance.

Our future work lies in the following three aspects: First, to probe into the realization of data stream parallel processing, and to support multi-encryption card load balancing and encryption technique for large-scale files. Second, by planning electronic documents management strategy, to increase the response speed by introducing cache prefetching strategy to read and process corresponding documents. Third, to explore watermarking technique of the files at clients so that to provide evidence, through taking pictures of the screen image etc., to trace anyone who leaks the documents.

Acknowledgment

The authors would like to thank Professor Yao Nianmin from Harbin Engineering University for kindly sharing his dataset with us. The work was supported by “Fund for National Development and Reform Commission high-tech”.

References

1. Biddle, P., England, P., Peinado, M., Willman, B. (2003) The darknet and the future of content distribution. In: Feigenbaum J.ed. Digital Rights Management, Lecture Notes in Computer Science 2696, Berlin:Springer-Verlag, 155-176
2. Lin, E.T., Eskieioglu, A.M., Lagendijk, R.L., et al. (2005) Advances in Digital Video Content Protection [J]. Proceedings of the IEEE, 93(1), 171 - 183.
3. Kuhlmann, D., Gehring, R.A. (2003) Trusted platforms, DRM, and beyond. In: Becker E. et al. eds. Digital Rights Management: Technological,Economic,Legal and Political Aspects, Lecture Notes in Computer Science 2770, Berlin: Springer-Verlag, 178-205
4. Bettini, C., Jajodia, S., Wang, X.S., Wijesekera, D. (2002) Provisions and obligations in policy management and security applications. In: Bernstein P.A., Loannidis Y.E., Ramakrishnan R., eds. Proc. of the 28th Int’l Conf. on Very Large Data Bases. Hong Kong: Morgan Kaufmann Publishers, 502-513.
5. Xu, X., Su, J. (2012) Research and implementation of iSCSI-based SAN static data encryption system. Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on . 29-31 Dec., 257 -260
6. Barth, A., Datta, A., Mitchell, J.C. (2006) Nissenbaum H. Privacy and contextual integrity: Framework and application. In: Proc. of the 27th IEEE Symp. on Security and Privacy. New York: IEEE Press, 184-198.
7. Chong, C.N., Ren, B., Doumen, J., Etalle, S., Hartel, P.H., Corin, R. (2004) License protection with a tamper-resistant token. In: Chae HL, Moti Y, eds. Proc. of the 5th Int’l Workshop Information Security Applications. LNCS 3325, Heidelberg: Springer-Verlag, 223-227.
8. Stallings, W. (2000) Network Security Essentials: Application and Standards. New Jersey: Prentice-Hall, 22-23
9. Garnett, N. (2001) Digital rights management, copyright, and napster. *ACM SIGecom Exchanges*, 2(2), 1-53
10. Davis, R. (2001) The digital dilemma. *Communications of the ACM,* 44(2), 77-83.

Received on the 20th of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 210-216
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

ANALYSIS AND SIMULATION OF NETWORKED CONTROL SYSTEMS DELAY CHARACTERISTICS BASED ON TRUETIME

Y. Wang[a], L. He[b]

[a]The Physics and Electronics Science College
Guizhou Normal University, Guiyang, China, 550025
Phone: (+086)-13007888851. E-mail: wyigz@126.com
[b]The Science College, Guizhou University
Guiyang, China, 550025
E-mail: 630897810@qq.com

The network-induced delay in networked control system is the most important reason of system performance degradation or even instability. In this paper, based on the analysis of structure of networked control system, the causes and integral part of delay was studied. Adopted the network control system simulation tool Truetime to build the simulation model of network control system delay. Discussed the effects of network delay on the performance of the control system in three cases. Then draw a conclusion that reduce the proportion of forward delay in total delay of the network can effectively improve the performance of networked control system.

Keywords: network delay, control performance, truetime, stability

1. Introduction

Recent years, with the increasing complexity of control object and continue expand of distribution area, the complex wiring, difficult to maintain, poor scalability, and other issues that brought by traditional point-to-point control system structure, has become increasingly prominent. It is difficult to meet the rapid development of the control system performance requirements. Thus the networked control systems (NCS) that exchange information between controller and actuator, sensor through a shared communication network has come into being. It is a complex fully distributed control system that combined communication network and control system as one unit. In addition, it reflects the development trend of networked, integration, distribution and intelligent. Compared with the point-to-point control, NCS has advantages of shared information resources, system wiring, reliability and flexibility, easy to extend and maintain, etc. So that it has rapidly develop and application, and catch more attention and applications in industrial control.

CAN bus control system is a kind of NCS that widely used. It has inevitable bring about some uncertain factors as the introduction of network. Moreover, mainly problem is performance degradation or even instability that caused by network-induced delay in NCS. In order to ensure the stability of control system, it is necessary to do an in-depth theoretical study to delay issue in NCS. In paper [1], in order to research stability issue, the discrete model of NCS was built. In addition, in paper [2], using the random delay modelling method to analysis the delay issues in NCS, and verify the effectiveness of algorithm by simulation results. While in article [3], it proposed a delay compensation strategy of segmentation dynamic matric to compensate the delay in the network transmission. Paper [4] put forward an adaptive PID control method based on BP neural network compensation. That can achieve online adaptive compensation network delay. Based on the analysis of the delay issues in the NCS, the paper [5] discussed and summarized the findings of delay problem. At last, this paper mainly studies the effects of forward and feedback delay in NCS to the performance of control system. Then draw a conclusion that there are some references value to the design of NCS.

2. Relationship of Delay and Sampling Period in NCS

NCS is a real-time closed-loop feedback control system that transmit data thought network. The typical structure is shown in figure 1.

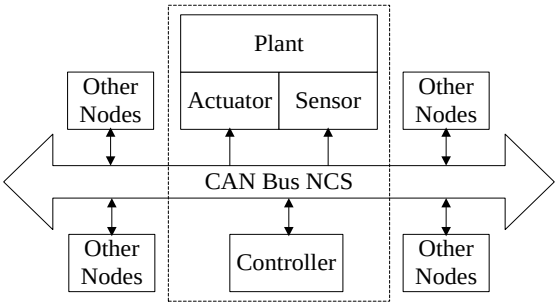


Figure 1. System Structure of NCS

As the intervention of network, delay of data transmission is inevitable [6]. From the perspective of control, the network delay can lagged the phase of system, deteriorated the system performance, narrowed the range of stability, or even caused system instability. In addition, from the scheduling perspective, due to the network delay, the information will not arrive on time, or even discard. The delay in network could do an important impact on the performance of control system. So during the analysis and design of vehicle NCS, it is the factor that cannot be ignored. As the presence of delay also increases the complexity of NCS research, the analysis and modelling of existing control theory cannot be directly applied to NCS [7,8].

Based on size, network delay can be divided into long delay and short delay. As the delay $\tau \in [0, \tau_{max}]$ and τ_{max} is less than sampling period h, that is $\tau_{max} < h$, then claimed it as the short delay. Else, the τ_{max} is greater than h, and then it is the long delay. As the long delay can convert into short delay by selecting sampling period and adopting appropriate scheduling algorithm and some other technical methods. Thus, this article only study short delay. The timing diagram is shown in figure 2.

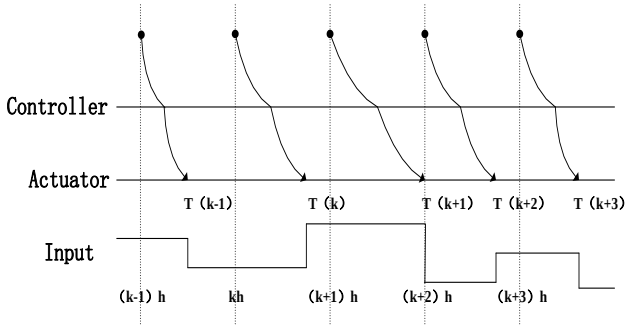


Figure 2. Timing Diagram of NCS

As the NCS that delay τ is less then sampling period h, the relationship between objects and controller can be expressed as the equation of state:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) \\ u(t) &= -Kx(t) \end{aligned} \tag{1}$$

In the equation, $x \in R_n$, $y \in R_p$, $u \in R_m$, represent the state variables, output and input variables of system. In addition, $A \in R_{n*n}$, $B \in R_{n*m}$, $C \in R_{p*n}$, $K \in R_{m*n}$, are constant matrix, represent the state matrix of system. Through figure 2, we can conclude that there are two controller outputs transmit to the actuator at most in one sampling period.

$$u(t) = \begin{cases} u(k-1) & t \in (kh, kh+\tau_k] \\ u(k) & t \in (kh+\tau_k, (k+1)h] \end{cases}$$

So, the state of NCS at $kh+\tau_k$ and $(k+1)h$ are:

$$\begin{aligned} x(kh+\tau_k) &= e^{A\tau_k}x(kh)+\int_0^{\tau_k} e^{As}dsBu((k-1)h) \\ x((k+1)h) &= e^{A((k+1)h-kh-\tau_k)}x(kh+\tau_k)+\int_0^{(k+1)h-kh-\tau_k} e^{As}dsBu(kh) \end{aligned} \tag{2}$$

Then the system discrete-time state equation is:

$$\begin{cases} x((k+1)h) = \Phi\cdot x(kh)+\Gamma_0(\tau_k)u(kh)+\Gamma_1(\tau_k)u((k-1)h) \\ y(k) = Cx(kh) \end{cases} \tag{3}$$

In the equation,

$$\Phi = e^{Ah}, \Gamma_0(\tau_k) = \int_0^{h-\tau_k} e^{As}dsB, \Gamma_1(\tau_k) = \int_{h-\tau_k}^{h} e^{As}dsB$$

Define the augmented state vector,

$$Z(kh) = \begin{bmatrix} x(kh) \\ u((k-1)h) \end{bmatrix} \tag{4}$$

So the close-loop system can be represented as:

$$Z((k+1)h) = \tilde{\Phi}\cdot Z(kh) \tag{5}$$

$$\tilde{\Phi} = \begin{bmatrix} \Phi-\Gamma_0(\tau_k)k & \Gamma_1(\tau_k) \\ -k & 0 \end{bmatrix} \tag{6}$$

As the spectral radius in equation (6) is less then 1, then the equation (5) is stability [9]. Thus, we can draw the relationship between τ_k and h:

$$\max\left\{\frac{h}{2}-\frac{1}{k},0\right\} < \tau < \min\left\{\frac{1}{k},h\right\} \tag{7}$$

The formula (7) is the constraint relationship between delay and sampling period. From the formula, the smaller the sampling period, the larger the stability allowable range of delay; the greater the sampling period, the smaller the maximum delay of allowed maintain system stability. It also shows that, in case of a stable system, the sampling period is small means that the large amount of data transmitted over the network will increase the network transmission delay.

3. Delay Model of NCS

Delay is one of the most common problems among all the issues of NCS. As the delay occurs, the control input of the system will not be up to date, and in severe cases the input timing disorder, and cause system degradation even instability. The entire control loop is passed in accordance with the the sensor→discrete-time controller→actuator→continuous-time controlled object, as shown in Figure 3.

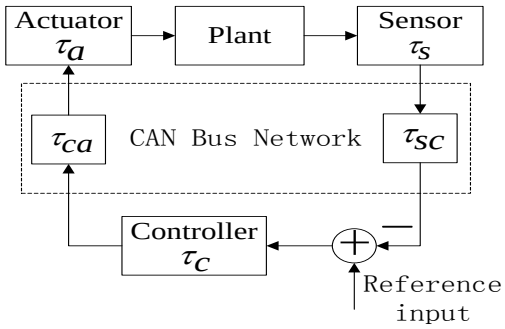


Figure 3. Delay Model of NCS

The whole process, that from sensor sample to actuator act in the plant, constitute a signal flow pass of the control system. In an ideal situation, it is often considered that the signal stream transfer process is instantaneous [10]. However, the forward delay and feedback delay in the control loop is inevitable for the introduce of network. That is: the sampling delay of sensor act sample to output; Calculate delay of controller do the calculate time to control amount signal; Driver delay of actuator led the control amount applied to controlled object; Then transmission delay during the information transmit in the network. Therefore, delay can be divided as implementation process delay and network transmission delay, and implementation process delay produced in the running process of sensors, controllers and actuators. Including the delay τ_s of sensors sampled data and complete the initialization of controller; after receive the packet, the time τ_c that controller calculate the amount of control signal and converts it into a message identifies required; and the computation time τ_a that actuators receives messages and translate them into the plant signals. Network transmission delay often consist by the feedback delay τ_{sc} that from sensor to controller and the forward delay τ_{ca} that from controller to actuator. Thus, the total delay of the control system can be expressed as:

$$\tau = \tau_{sc} + \tau_{ca} + \tau_s + \tau_c + \tau_a .$$

While delay of execution process mainly depends on the overall performance of each node, so it could reduce the execution impact of delay to a very small extent by select a high speed processor and optimization algorithm. Therefore, during the analysis and design of a NCS, the total delay of the NCS can be expressed as the sum of forward delay τ_{ca} and feedback delay τ_{sc}.

$$\tau = \tau_{sc} + \tau_{ca} .$$

4. The Impact of Delay in NCS

In order to study the impact of forward delay τ_{ca} and feedback delay τ_{sc} to the control performance of NCS, build a closed-loop control system by sensor, actuator, controller, interference, DC motor. As shown in figure 4. Using sensor to sample the speed of DC motor and process the data. Then sent information to the controller via network, and calculates the control amount, at last sent results to the actuator. Interference node is used for produce periodic and non-periodic signal, and adjust the load on the network traffic.

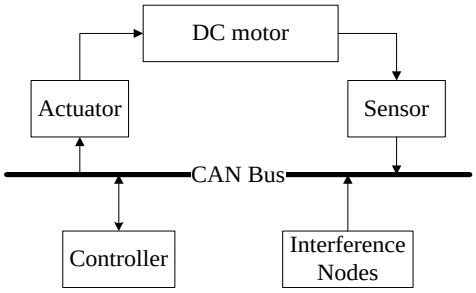


Figure 4. Network Coles-Loop Control System

Truetime is a toolbox that suitable for network simulation, which is developed by LUND University. It is a joint simulation tool based on Matlab/Simulink. It often used to study effect of network transmission delay time uncertainty to the control performance, and can be used on simulation of multi-scheduling algorithm in the NCS. The Truetime mainly includes six functional modules: Truetime Kernel, Truetime Network, ttGetMsg, ttSendMsg, Truetime Wireless Network and Truetime Battery. A control system simulation model can be build by connect the modules in Truetime and Simulink.

Truetime Kernel often work as nodes in the NCS, used to analog sensor, controller, actuator and other control elements, include A/D and D/A converter, outer channel, network, etc. The kernel module execute operations in accordance with the user-defined tasks, often use Matlab m-file or C^{++} language to write execute code. Truetime network is used to simulate a variety of networks. Optional paraments includes: amounts of network nodes, number of network, data transmission rate, MAC protocol, etc. It provides several MAC protocol: CSMA/CD, CSMA/CA, TDMA, FDMA, ROUND ROBIN. It is driven by an event mode, while the news come in/out from the network, it begins execute. Meanwhile, it defines many network scheduling policy (such as FP, RM, DM, EDF). Before use the module, it must be initialized. The main steps of build a NCS by Truetime are shown as follow.

1) through Truetime and Matlab/Simulink module to establish real-time network control system model;
2) design of nodes: adopt Truetime Kernel to build each node (such as sensor, controller, etc). Initialized all nodes, and set appropriate parameters;
3) design of network: Set parameters of network module, includes: type of network, packet loss probability,etc;
4) connected Truetime modules and Simulink modules, then execute it.

Figure 5 is the NCS build by Truetime kernel in the situation of Simulink.

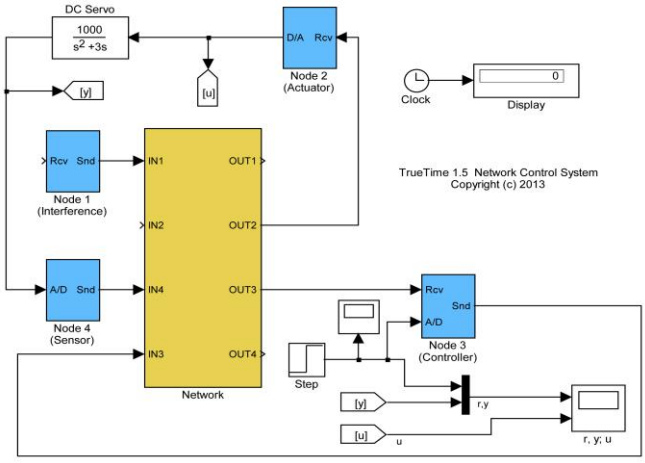


Figure 5. Simulation Model of NCS

Select the stepper motor as control plant, the mathematical model is:

$$G(s) = \frac{1000}{s(s+3)}$$

Select four Truetime Kernel modules respectively as sensor, controller, actuator and interfering node. The network module adopt CAN protocol (CSMA/AMP) to achieve communication.

At first, we analysis the effect of total delay in CAN bus NCS to performance of control system. The input is a step signal, and the system step response is shown in figure 6 while the delay is $\tau = 0.0005$, $\tau = 0.02$, $\tau = 0.06$ and $\tau = 0.3$.

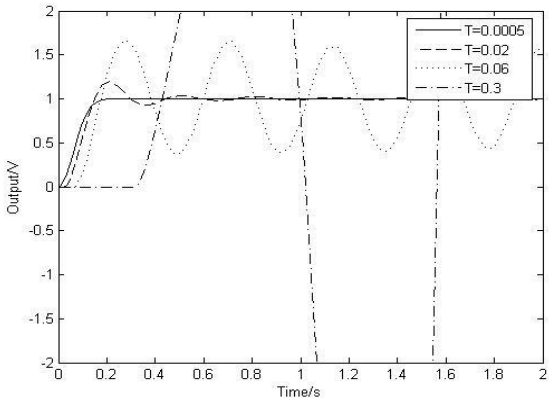


Figure 6. System Response in Different Delay

Simulation results show that the total delay of NCS smaller, the rise time of control systems and overshoot smaller, then the better the stability of system. With total network delay increases, the stability control system work deteriorated even instability.

Then study the affect of forward delay and feedback delay to performance of the control system. While the input signal is a square wave, set the $\tau = 0$, $\tau_{ca} >> \tau_{sc}$, and $\tau_{ca} << \tau_{sc}$. As shown in Figure 6.

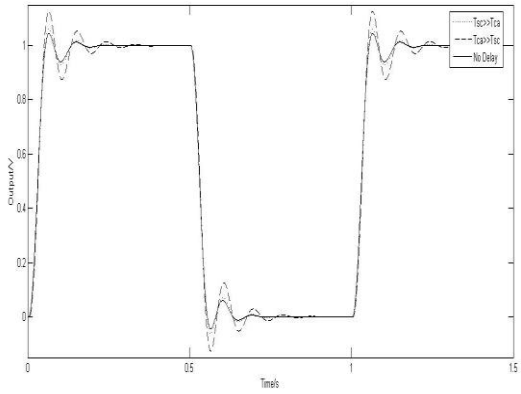


Figure 6. Simulation Results of Delay in NCS

Through the simulation results:

1) While the delay of NCS is zero, that is: $\tau = 0$. The rise time of system step response and overshoot are relatively small, this is consistent with the results of theoretical analysis.
2) When the NCS forward delay τ_{ca} is less than feedback delay τ_{sc}, the performance of the system is better than the situation that forward delay τ_{ca} is greater than feedback delay τ_{sc}, and the greater the forward delay, the system overshoot greater.

5. Conclusions

Based on the characteristics of NCS, analysis the effect of network delay to the performance of the

control system in the situation of sensors, controllers, actuators, and controlled object-shared network. In addition, established discrete mathematical model of the vehicle NCS, given the relationship of delay and sampling period under a stable conditions in NCS. Simulation results show that forward delay τ_{ca} has a greater effect than feedback delay τ_{sc} in the vehicle NCS. With the increase of τ_{ca} , the overshoot of system increases and the rise time becomes longer. So minimize the forward delay, during the design of vehicle NCS, could improve the performance of the NCS. The conclusion has some reference value for the design of NCS.

Acknowledgment

This work was financially supported by the National Natural Science Foundation of China (Grant No. 61262007), Guizhou Science and Technology Department School Cooperation Project (Qian Bureau No.[2013] 7003) and Guiyang Science and Technology Department Platform for innovation plan (2012303).

References

1. Zhao, H.Z. (2012) Stability Analysis of Networked Control Systems with Time Delay [J]. *Science Technology and Engineering*, 12(2), 342-344. (In Chinese).
2. Chen, M.W., Jing, C.Y., Zhu, J.L., et al. (2012) Design of Stochastic Interval Delay Controllers of Networked Control Systems [J]. *Journal of Guangdong University of Technology*, 29(1), 59-63. (In Chinese).
3. Zhu, K.Y. (2010) Compensating strategy based on dynamic matrix control for networked control system [J]. *Microcomputer & Its Applications,* 12, 75-77. (In Chinese).
4. Huang, J., Liu, J. (2009) Self-adapting compensation study to the network- induced delay of networked control system [J]. *Journal of Jiangsu University of Science and Technology (Natural Science Edition)*, 23(4), 335-337. (In Chinese).
5. Guo, G., Jia, E.N. (2009) Delays in networked control systems: Analysis and preview [J]. *Control and Decision*, 24(1), 1-6. (In Chinese).
6. Guo, X., Wu, B. (2010) The Impact of Time-delay on Networked Control System and Stability Region Analysis of PID Controllers [C]. *2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE)*, 163-166.
7. Hai, Z., Qing-Hai, H., Yin, L. (2012) Analysis and Simulation of Network Control System Delay Based on Truetime [J]. *Industrial Control Computer*, 25(2), 29-35. (In Chinese).
8. Xu, L., Lei, H., Ru J. (2011) Research on Predictive Control of Stochastic Network-induced Delay for Network Control Systems [C]. *2011 Chinese Control and Decision Conference (CCDC)*, 929-932.
9. Jing, N., Wang L., An, B.-Y. (2009) Analysis and policy of network latency in CAN control system [J]. *Computer Engineering and Design*, 30(20), 4599 -4602. (In Chinese).
10. Rui, D., Xian-Ming, T., Yu, J.-S. (2010) Analysis of Delay Compensation in NCS based on TrueTime Tool Box [J]. *Industrial Control Computer*, 23(2), 24-26. (In Chinese).

Received on the 20th of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 217-223
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

A UNIFIED MODELING AND PARALLEL CO-SIMULATION METHOD OF CYBER PHYSICAL SYSTEM

C. Song[1], Ch. Du[1], G. Li[2]

[1]Department of Computer
Northwestern Polytechnical University, Xi'an, China
Phone: (+029)-88431545. E-mail:songcy@mail.nwpu.edu.cn
[2]Department of Computer
North China Electric Power University, Baoding, China
Phone: (+0312)-7525255. E-mail: ququ_er2003@126.com

CPS system features with close interaction among cyber subsystem, physical device and the dynamic environment. The model semantic gap between the three kinds of subparts hinders the verification in the early design stage. In this paper, a unified modelling and parallel co-simulation method is proposed, in which, the discrete cyber subsystem, the continuous physical device and the working environments are modelled in an object model where objects with deterministic state chart are connected with asynchronous event channels and synchronous data channels. A parallel scheduler based on temporal event is designed to guarantee the interaction temporal relationship between the heterogeneous subparts. Finally, the method is realized based on Rhapsody and a moving control system is modelled and simulated to show the feasibility and effectiveness. The executable unified model provides a test bed not only for the cyber subsystem, but also for the physical part to show the properness of physical arguments.

Keywords: CPS, unified modelling, parallel co-simulation, temporal event

1. Introduction

Cyber physical system (CPS) comprises networked physical devices, working through deep interaction among the technologies in computing, communication and control (3Cs for short) based on automatic perception of the environment and the system itself [1]. For example, in the moving control CPS area, the 3Cs technologies are used to control moving device like train, aeroplane and so on. Usually, the running environment of the moving device is uncontrollable through the device, and whether the cyber subsystem is designed properly depends on the features of the physical subsystem and the environment. Thus, verifying the cyber subsystem is difficult during the early design stage. It is important to make a unified formalism model and attain the simulation results of the close-loop system, to verify not only the correctness of the computing logic but also the timeliness of the computing process.

Since both the discrete cyber behaviour and the continuous physical process consume time simultaneously and continuously, the heterogeneity in the model of the time in cyber domain and physical domain is the essential problem to overcome when establishing a unified CPS model, which shows the complete dynamic behaviour with temporal property [2]. Edward A. Lee has led his team in developing the Ptolemy [3] and PTIDEs tools [4]. Ptolemy provides a unified modelling and simulating framework for CPS based on actors and kinds of composable model of computing (MoC), but is lack of temporal semantic to model and simulate the whole temporal behaviour of a CPS system. PTIDES models the time-consuming property of the computing process, and design of embedded system with PTIDES results in deterministic processing of events, but does not focus on the interaction between cyber and physical processes. Hybrid simulation framework, which is built through the common communication interfaces like those of HLA/RTI[5], connect and reuse existing discrete simulator like Ptolemy II and continuous simulator like Matlab/Simulink[6], still doesn't emphasize the time consumed by the cyber process yet. Rhapsody[7], which is an embedded system development environment based on UML, provides a Simulink profile to support the integration of C/C++ code generated by Simulink/RTW into the object model in Rhapsody; It supports concurrent running of the cyber functions and physical processes, but does not guarantee the temporal relationship of the cyber physical behaviours. The OSMC organization in Europe develops a ModelicaML profile and integrates the state charts into Modelica physical model to support the timing during the behaviour simulation [8]. The hybrid theories such as hybrid automata [9] and timed hybrid Petri net [10] focus on the dynamic behaviour but without the structure of the system.

The method in this paper tries to provide a unified view of heterogeneous subparts in UML 2.0

objects and state charts, and implement the simulation integrating the integral algorithm of physical and environment features. In section 2, the moving control CPS is taken as an example to show the architecture of a closed model of a CPS system and gives the basic classes for the modelling. Section 3 illustrates the modelling method for the co-simulation model, and provides a unified and executable model in Rhapsody. Section 4 describes the parallel time forward mechanism and the scheduling algorithm. The features of the method are analysed in section 5, and section 6 concludes the paper.

2. Architecture of CPS System

Take the moving control CPS for example, the computing and control logics are embedded in the physical device. The computing centre communicates with external cyber logics through wired/wireless network. Usually, to keep the efficiency of the perception and control, sensors and controllers are connected by certain bus inside the vehicle with high speed. More than one bus may be adopted when needed. A multiple task-computing centre is placed as the brain of the smart physical node, and each sensor node and controller node are controlled and driven by the brain. Besides the cyber subparts, the dynamic features of the environments and the physical device should be modelled too. A closed architecture model in a unified semantic provides a unified view and helps simulate and watch the effect of each subpart of the CPS system in one single platform.

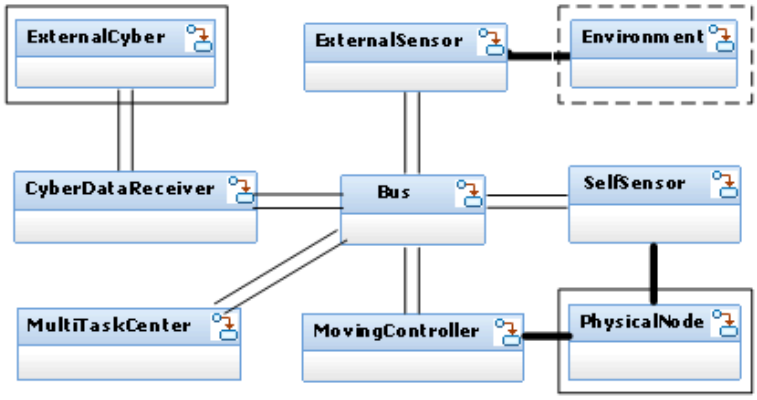


Figure 1. Architecture of a closed moving control CPS model

As shown in figure 1, nine basic classes are extracted for a closed loop model of CPS system. The internal behaviour of each class is modemed as a deterministic state chart and the interaction among classes are modelled as channels. The state chart are the same as that in UML 2.0, and can be translated into certain programming code. The semantics and implementation for the channels are illustrated in section 3. The six classes without any wire frame are the cyber classes for an embedded system node, and the main features of them are as follows:

- *MultiTaskCentre*, represents embedded nodes with concurrent OS.
- *Bus*, represents devices that transfer information among cyber nodes.
- *ExternalSensor*, means sensors that perceive physical environment.
- *SelfSensor*, means sensors that perceive the physical device itself.
- *CyberDataReceiver*, nodes that exchange message with external cyber nodes.
- *MovingController*, are devices that receive control decisions from a *MultiTaskCentre* object and actuate the actual physical device.

Besides the cyber classes, three other classes named PhysicalNode, Environment and ExternalCyber are designed to represent the physical subsystem, the external physical environment and the external cyber system respectively. An object based model is adopted for the simulation, and when modelling a given CPS system, each object of the system should belong to one and only one class from the nine specified above.

3. Unified Object Model for Simulation

The co-simulation view for a specified CPS system is an object model structure with state chart binding with the objects. The objects are connected with channels, which should be realized with

executable model elements. Both the discrete and the continuous behaviour are modelled in state chart. The semantics and implementation are illustrated in detail.

3.1. Semantics of Channels

The connection rules among the specified classes are given in figure 1. Channels are divided into EventChannel and DataChannel to model the communications in discrete domain and continuous domain respectively by the while bars and black bars.

- *EventChannel*, means asynchronous message channels, which simulate communication among the internal and external cyber objects. In an *EventChannel*, when an event in one end is generated, it will be sent to the other without stopping itself. Event is the trigger for a state transition in a state chart. Generally speaking, an event represents the condition that continues the behaviour of the object from the current state. Thus the condition can be defined and simulated by a unique event.
- *DataChannel* means synchronous data channels, which represent data flow among environments, physical device and sensors. A *DataChannel* connects two variables of the same type in two objects, and when one data changes, the other variable updates to the same value of the changed one. *DataChannel* are used between two objects, one of which has a variable always with the same value as a variable of the other. For example, the fact that the velocity sensed by a velocity sensor is actually from the vehicle object, can be modelled as figure 2. *DataChannels* are maintained by a *DataMonitor* module that is responsible for monitoring the variables and updating related values automatically.

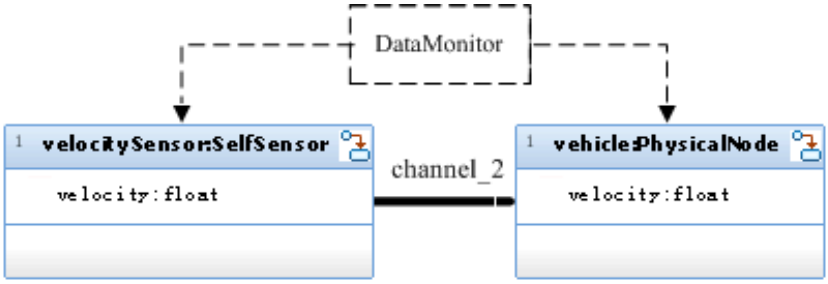


Figure 2. Objects connected by *DataChannel*

3.2. State Chart for Continuous Behaviours

In a unified object model of a moving control CPS with its surroundings, state charts are adopted to model the behaviour of all the objects. Usually, state chart is a suitable model for cyber objects, not for the continuous ones. To model and simulate the behaviours simultaneously, state chart is defined for continuous behaviours.

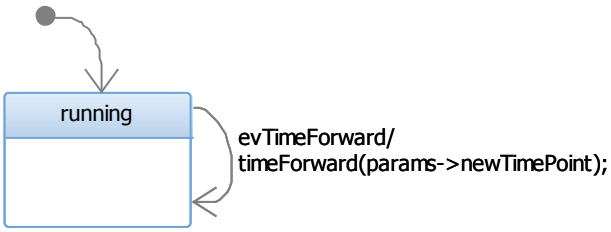


Figure 3. State chart for continuous object behaviour

Take the PhysicalNode object vehicle in figure 2 for example, the state chart for a continuous object is defined and implemented as shown in figure 3, which is created in Rhapsody. Here the evTimeForward is an event with an argument newTimePoint indicating the new time point where the continuous physical behaviour should pause. An interface timeForward is implemented to update the feature variables of the vehicle to the current time point. The interface is implemented integrating the integral algorithm of the continuous behaviour, which is usually modelled by differential equation as in Matlab/Simulink. In our implementation, a synchronize dataflow model for the vehicle moving

control is established in Simulink and the integral code is generated by RTW as described in reference [6].

3.3. Unified View of the Co-simulation Model

A co-simulation model of a vehicle CPS system is established in Rhapsody according to rules specified in section 2 in order to execute the model directly reusing the object execution functionality without developing an execution platform from scratch. The vehicle system moves in an outside environment with changing road surface, and makes electricity from the sunlight. All the nodes inside the vehicle are connected through a CAN bus and controlled by the computing centre, the GNC object, which navigates the vehicle and controlled the running of other cyber nodes.

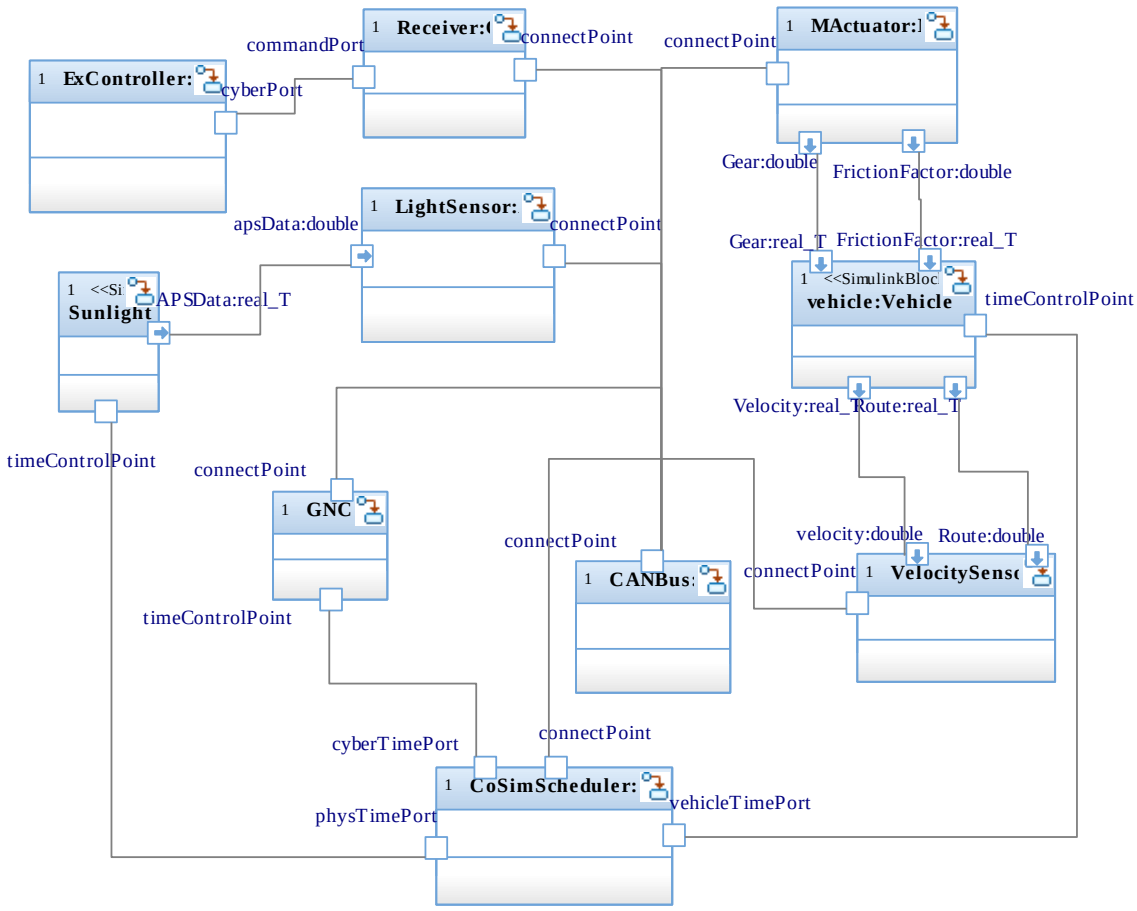


Figure 4. Unified Co-simulation object model of a vehicle CPS system.

The closed simulation model is shown in figure 4. The vehicle object represents the physical device, the Sunlight object represents environment entity, and the rest belong to the cyber domain. Channels are modelled with ports and connections, and the state chart of each continuous object is set according to a specification document. State charts for continuous objects including Sunlight and vehicle are the same as shown in figure 3. What's different between them is the implementation of timeForward interface because of the difference in their original physical law. For the vehicle object, two data ports named Gear and FrictionFactor are modelled to reflect the dynamic gear control and the effect of the road surface for the moving of the vehicle.

The class of each object inherits from one of the nine basic classes except the CoSimScheduler, which is defined to maintain the interaction temporal relationship of the cyber objects. The CoSimScheduler is not an component of a CPS system, but a kernel component for the co-simulation. The time forward mechanism is described in next section.

4. Temporal Event Driven Parallel Scheduler

A cyber physical objects parallel scheduler is designed and implemented based on temporal event to execute the simulation model.

4.1. Temporal Event

A temporal event is an event with an argument indicating a time point value in this paper, not the one described in [11], where the time value represents the occurring time of the event. For example, the event evTimeForward in figure 3 is a temporal event driving the continuous integral behaviour, which simulates the dynamic features of PhysicalNode objects and Environment objects. Besides, temporal event is used in MultiTaskCenter object to drive the sense-control behaviour periodically. Here the period is assumed sufficient for all the cyber behaviour and all other cyber objects are common event driven and controlled by the MultiTaskCentre object.

In the co-simulation, temporal events can be generated only by the CoSimScheduler object. During the simulation, both the periodic discrete cyber behaviour and the continuous integral behaviour are driven by discrete events along the time axes. The CoSimScheduler object implements a parallel time forward mechanism to reflect the actual interaction points among the cyber, the environment and the physical devices.

4.2. Parallel Time Forward

The fact, that the 3Cs technologies are based on indeterminate time and discrete event, while the continuous domain behaviours strictly with time, actually means a parallel relationship between the cyber domain and the continuous domain.

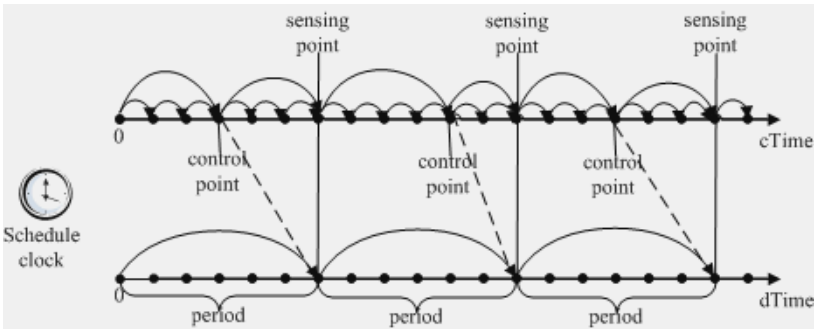


Figure 5. Parallel steps of discrete and continuous domain.

As shown in figure 5, the time of the continuous domain and the discrete domain are represented as cTime and dTime respectively. During the simulation, Both the cTime and dTime are forwarded periodically. The period in figure 5 is the greatest common divisor of the concurrent sensing-control tasks in the MultiTaskCentre object.

The dTime is updated period by period directly, while each period of cTime is cut at a predicted control point when the control effect actually reaches the PhysicalNode object. The control point may locate in different places in different periods due to the indeterministic time consuming of cyber process, but always belong to the designed time-consuming area. In each period, the control point is given by the scheduler as a random value in the area. The cTime forwardness is similar with Simulink, and is moving on by the integration micro steps. For each micro step, the integration method is called to update the current physical values.

A co-simulation function for each period is defined as doPeriod(), and the algorithm is described as follows:

```
doPeriod() {
  Update values of sensors;
  Call cyber behaviour and store the control command for physics;
      Compute the predicted control point;
  Forward the cTime to the control point;
  Send the control command to the physical object;
```

Forward the cTime to the next sensing point;
If environment is changed, update the physical arguments;
}

The co-simulation scheduler works by calling the doPeriod() function in loop until a specified time point or other conditions. For example, the simulation has successfully reported whether the navigation algorithm can avoid vehicle crash with obstacles as shown in figure 6. Adjust the argument in the GNC object in figure 4, rerun the simulation, we get the simulated data for all the micro steps for the vehicle route, and make the route curve. When an obstacle is 50 meters away from the vehicle as the initial state, the left and the right diagrams shows the result when an obstacle checking algorithm takes 5 meters and 6 meters as obstacle argument respectively. Both the two times finished within 1000 ms in logic, and intersect point in the right diagram reports a bad argument for the obstacle checking algorithm.

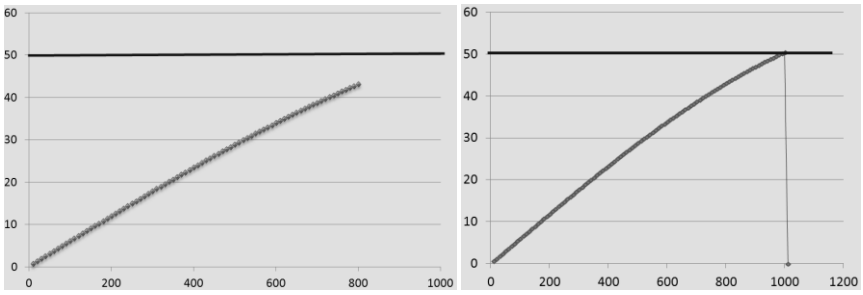


Figure 6. Co-simulation result for an obstacle checking situation

5. Discussions

The modelling and co-simulation method is based on objects with state chart, event channel and data channels. It reuses the seamless integration functionality of object structure and dynamic behaviour provided by Rhapsody instead of defines a new modelling language and develops parsers and code generators all from scratch. The features of the method are:

- Integrated simulation of discrete behaviour and continuous physical change in a unified event-driven and periodic way. Since the Simulink profile in Rhapsody does not guarantee the integral behaviour as specified in a given CPS system, new state chart is defined for the simulation.
- Accurate semantic of interaction between the cyber, physical and external environments. The parallel co-simulation scheduler actually reflects the interaction of heterogeneous subparts in an executable model, which is more accurate than a specification document in natural language.

Theoretically speaking, the structure of the unified view can be represented by some component-based modelling language such as AADL. However, the event driven behaviour of a component is less intuitionistic than state chart in UML. The integral function is from Simulink/RTW in our implementation, but unlimited to RTW. The code can be modified easily when necessary. Once the simulation model is established, it helps watch the effect of different physical or environment argument; it also provides a test bed for the cyber subsystem, and the behaviour of the cyber objects can be refined iteratively promising that the channels outside unmodified.

6. Conclusion

A unified modelling and co-simulation method for CPS system is illustrated in this paper. The cyber subsystem are modelled together with the dynamic environment features and physical process in a unique object model where objects with state charts are connected by event channels or data channels to form a closed and executable simulation model. A parallel co-simulation scheduler is designed and implemented to model the interaction between the discrete and the continuous parts and drive the objects in right temporal sequence. Integral functions are integrated to complete the behaviour model of continuous parts. The method is illustrated with a moving control CPS system, but it applies to other CPS systems too. The co-simulation model provides a test bed for the cyber subsystem during the design stage before the real integration of the heterogeneous subsystems.

7. Acknowledgment

This paper is supported by the National High Technology Research and Development Program of China (No.2011AA010102) and the Fundamental Research Funds for the Central Universities (No.11MG50).

References

1. He, J. (2010) Cyber-physical systems, *Communications of the CCF*, 6(1), 25-29.
2. Lee, E.A. (2010) CPS foundations, In: Proc. of Design Automation Conference, ACM, 737-742.
3. Eker, J., Janneck, J.W., Lee, E.A., et al. (2003) Taming heterogeneity the Ptolemy Approach. In: Proc. Of the IEEE, 91(2).
4. Zhao, Y., Liu, J., Lee, E.A. (2007) Ptides: A programming model for time-synchronized distributed real-time systems, In: Proc. of RTAS 07, Bellevue, WA, United States, 259-268.
5. Sung, C., Kim, T.G. (2011) Framework for Simulation of Hybrid Systems: Interoperation of Discrete Event and Continuous Simulators Using HLA/RTI. In: Proc. of 2011 Workshop on Principles of Advanced and Distributed Simulation, Nice, France, 1-8.
6. Shi, R. (2011) *Simulation and realization principle of Simulink/Stateflow*. Master's thesis, Cheng Du: University of Electronic Science and technology.
7. IBM Corporation. *Rhapsody User Guide7.6*, (2013)
8. Henriksson, D., Elmqvist, H. (2011) Cyber-physical systems modelling and simulation with modelica. In: *Proc. of 8th Modelica Conference*, 502-509.
9. Henzinger, T.A. (1996) The theory of hybrid automata. In: Proc. of the Eleventh Annual IEEE Symposium on Logic in Computer Science (LICS), 278-292.
10. Gudino-Mendoza, B., Lopez-Mellado, E., Alla, H. (2012) Modelling and simulation of water distribution systems using timed hybrid Petri nets, *Simulation*, 88(3), 329-347.
11. Tan, Y., Vuran, M.C., Goddard, S., Yu, Y., Song M., Ren, S. (2010) A Concept Lattice-based Event Model for Cyber-Physical Systems. In: *ICCPS'10*, 50-60.

Received on the 20th of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 224-228
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

GPS LOCATION ACCURACY IMPROVEMENT BY WLAN

Fan Zhang[1,2], Xincai Wu[1,2], Shang Zhang[2]

[1]Faculty of Information Engineering, China University of Geosciences(Wuhan),
Wuhan city, Hubei province, 430074, China
Phone: (+86)-13886092211. E-mail: zhangfan0725@sohu.com
[2]Zondy Cyber Group Co., Ltd
A10#,Optical Valley Software Park,#1, Wuhan Hubei, China

Global Positioning System (GPS) has been widely used all around the world. However, in many cases, the use of the location accuracy cannot meet the users' requirements. On the other hand, many places are covered with WLAN signals. Therefore, comprehensive utilization of the satellite signal and WIFI signals of WLAN to realize the optimization of GPS location become possible. This paper is to optimal location accuracy based on the satellite signals and WLAN signals.

Keywords: GPS; WLAN; location; error analysis; accuracy improvement.

1. Introduction

Nowadays, global satellite positioning system has been widely used in various fields. With the improvement of user demand for positioning accuracy, the traditional GPS positioning method is not very good to meet the use requirements of all kinds of users. Therefore, there are many kinds of GPS optimization algorithm, used to improve GPS positioning precision. On the other hand, in many occasions, we cannot only receive the satellite signals, but also the WLAN signals [1]. This paper assumes that the WLAN transmitter also integrates the function of receiving the signals of GPS satellite. We will discuss the use of WLAN to assist GPS signal. In addition, correct the positioning accuracy of receiver by comparing the measuring position and the real position of WLAN transmitter [1].

2. GPS Positioning Theory

2.1 GPS

As a positioning system developed by the United States, GPS has been widely used all around the world. It provides separate services for civil and military users. These are called the Standard Positioning Service (SPS) and the Precise Positioning Service (PPS). The SPS is designated for the civil community, whereas the PPS is intended for U.S. authorized military and select government agency users. This paper focuses on the SPS which is specified to provide accuracies of better than 100m (95%) in the horizontal plane and 156m (95%) in the vertical plane (global average; signal-in-space errors only) [2]. Obviously, because of the horizontal and vertical error, the use of GPS was limited. It will not be able to meet the industry and the application requirements of public users based on location-based services. Scientists and researchers had done many works on error sources of GPS. A systematic and effective GPS error correction theory gradually formed. The establishment of these theories improved the positioning precision of the GPS to some extent.

2.2 GPS Error Analysis

GPS utilizes the concept of TOA ranging to determine user position. This concept entails measuring the time it takes for a signal transmitted by an emitter (e.g., radiobeacon, or satellite) at a known location to reach a user receiver [2]. This time interval, referred to as the signal propagation time, is then multiplied by the speed of the signal (e.g., speed of sound or speed of light) to obtain the emitter-to-receiver distance. According to the perspective of analytic geometry, the receiver can determine its position, by measuring the propagation time of the signal broadcast from more than three emitters (i.e., navigation aids) at known locations[2]. Moreover, consider revising the clock synchronization issues between the satellite and the receiver, the receiver need one more propagation time from another emitter to solve the problem of clock synchronization [2]. Therefore, GPS location will need at least four

emitters' measurements.

Considering the physical factors from nature, the interference caused by the distance between the emitters, and the receiver internal error caused by other interfering satellite signal errors, there is always have errors of the measured distance from the emitters to the receiver. The distance of emitter and receiver which measured by signal transmitted called pseudorange. To analyse the effect of errors on accuracy, a fundamental assumption is usually made that the error sources can be allocated to individual satellite pseudoranges and can be viewed as effectively resulting in an equivalent error in the pseudorange values. The effective accuracy of the pseudorange value is termed the user-equivalent range error (UERE). Moreover, as the location of the satellite constantly changing, dilution of precision (DOP) is used to describe the error which is caused by the different relative geometric layout between receiver and satellite. The accuracy of the position/time solution determined by GPS is ultimately expressed as the product of a geometry factor and a pseudorange error factor. Loosely speaking, error in the GPS solution is estimated by the formula [2]:

(error in GPS solution) = (DOP) × (UERE) (1)

This article will analyse the accuracy in the horizontal plane when the locations of the emitters are fixed. When the geometry layout between the receiver and the emitters is variable, the geometric layout can be considered as a random variable quantity. Since the geometric factor fixed, the error of GPS solution will only determined by the pseudorange error factor. Table 1 show estimates of typical contemporary UERE budgets issued by the GPS Joint Program Office (JPO) [2]. It shows the results of pseudorange which Converted by various errors. Among them, as the biggest single error sources for SPS user, SA was set by the U.S. Department of Defence (DOD) to add error to the users' navigate solution.

TABLE 1. Estimates of typical contemporary UERE budgets [2]

No	Error Sources	GPS errors(m)
1	Satellite clock stability	3.0
2	Satellite perturbations	3.0
3	SA	32.3
4	Ephemeris prediction error	4.2
5	Troposphere delay	5.0
6	Receiver noise	1.5
7	Multipath	2.5
8	Others	1.9
9	Total(UERE)	33.3

For horizontal error, we usually use circular error probability (CEP) of error distribution (CEP) to represent. Thus, when the HDOP (Horizontal DOP) is 1.5, 95% of the error distribution CEP value can be calculated by the following formula [2]:

$$CEP_{95} = 2.0 HDOP \sigma_{UERE} = 2.0 \times 1.5 \times 33.3 = 99.9m \quad (2)$$

That is, under the conditions defined in the article, GPS error is 99.9 m in the horizontal plane.

3. WLAN / Wi-Fi

WLAN (Wireless Local Area Networks, IEEE 802.11 standard; 'Wi-Fi' is used interchangeably or as a superset of IEEE 802.11 and denotes the registered trademark of the Wi-Fi Alliance) can be used to estimate the location of a mobile device within this network. WLAN positioning can be used for indoor environment. The use of WLAN signals is a tempting approach, since WLAN access points are readily available in many indoor environments and it is possible to use standard mobile hardware devices. The range of 50 m to 100 m, which is typically covered by WLAN outreaches that of Bluetooth or RFID [1]. Another advantage of using WLAN is that line of sight is not required. ToA, TDoA or AoA methods are less common in WLAN due to the complexity of time delay and angular measurements. The most popular WLAN positioning method is to make use of RSSI (Received Signal Strength Indicators) which are easy to extract in 802.11 networks and can run on off-the-shelf WLAN hardware [3]. Therefore, WLAN positioning systems have become the most widespread approach for indoor localization.

Similar to the principle of GPS positioning, WLAN signal transmitter equivalent to GPS positioning satellites. By measuring at least three fixed positions WLAN distance between the signal transmitter, we can obtain three-dimensional coordinates of the receiver, That is measured the specific location of the receiver. Table 2 is based in part on WLAN signals indoor positioning system performance parameters [4].

TABLE 2. Localization approaches using WLAN and reported performance parameters [1]

Name	Year	Reported Accuracy(m)	Area (m2) per Access Point. For non RSS: Type of Modification	Principle
Parody	2006	3.3	1235/14 = 88	Online learning
Xiang	2004	2-5	1400/5 = 280	Fingerprint & map
Bahl	2000	5	978/3 = 326	Fingerprinting
Gansemer	2010	2.1	972/24 = 40	Fingerprinting
Koski	2010	5-7	5600/206 = 27	Fingerprint & map
Hansen	2010	2	Ca. 140/14 = 10	Fingerprinting
Chen	2005	2-4	400/5 = 80	Fingerprint & RFID
King	2006	1.6	312/20 = 156	Fingerprinting
Teuber	2006	2-3	400/5 = 80	Fingerprinting
Mazuelas	2009	4	3375/8 = 422	Multilateration
Gallagher	2010	Room level	Unknown	Fingerprinting
Ekahau	2009	7	Ca. 450/5 = 90	Fingerprinting
Skyhook	2011	30-70	‘global’ coverage	Fingerprinting
Golden	2007	1-5	Hardware mod.	Diversity
Reddy	2007	15	Hardware mod.	Time of arrival
Wong	2008	2	Hardware mod.	Angle of arrival
Gunther	2004	5-15	Software mod.	Round trip time
Ciurana	2009	1.7	Software mod.	Round trip time

4. GPS Positioning Accuracy Correction Assisted by the WLAN

From the previous introduction, we can see that the WLAN positioning accuracy is more better than GPS. In other words, we can only use WLAN signals to measure the receiver’s location where at least three WLAN signals can be detected. In this case, GPS satellite signals are not required[5].

However, we usually cannot simultaneously receive three WLAN signals at least to complete the positioning process. Meanwhile, because of the interference from the external environment, such as building block, we cannot receive GPS signals sometimes. In this case, in order to complete global positioning for receiver, and optimize the positioning accuracy as far as possible, we cannot simply rely on GPS satellites signals or WLAN signals. We should combine GPS signals and WLAN signals to measure the position of the receiver. This article will analyse the method that how to measure the receiver’s position where three GPS signals and one WLAN signal can be received at least. Further, this paper proposes a correction method of positioning accuracy based on WLAN signal [6].

4.1. Position Determined Using GPS and WLAN

According to the perspective of analytic geometry, the receiver can determine its position, by measuring the propagation time of the signal broadcast from more than three emitters (i.e., navigation aids) at known locations. Additionally, taking into account the distance offset generated by satellite clock error, another unknown parameter (satellite clock error) is needed to be introduced. Therefore, the locations of at least four emitters need to be known. It is assumed that the three-dimensional coordinates of the receiver is (x_u, y_u, z_u), the distance offset generated by satellite clock error is t_u, The spatial coordinates of three satellites are (x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3), The spatial coordinate of the WLAN signals emitter is (x_a, y_a, z_a), the pseudoranges of each emitters (satellite and WLAN transmitter) are $\rho_1, \rho_2, \rho_3, \rho_a$, The speed of light is a constant c. Thus, the equations can be generated [2]:

$$\rho_1 = \sqrt{(x_1 - x_u)^2 + (y_1 - y_u)^2 + (z_1 - z_u)^2} + ct_u \tag{3}$$

$$\rho_2 = \sqrt{(x_2 - x_u)^2 + (y_2 - y_u)^2 + (z_2 - z_u)^2} + ct_u \quad (4)$$

$$\rho_3 = \sqrt{(x_3 - x_u)^2 + (y_3 - y_u)^2 + (z_3 - z_u)^2} + ct_u \quad (5)$$

$$\rho_a = \sqrt{(x_a - x_u)^2 + (y_a - y_u)^2 + (z_a - z_u)^2} + ct_u \quad (6)$$

From the above equations, the spatial coordinates of the receiver can be solved. Then, the location of the receiver can be determined. The measurement value from receiver to satellite (WLAN transmitters) always affected by many factors. In this case, a method to solve the program is to measure more pseudoranges from more than four emitters. Then process the pseudoranges using the valuation techniques of the least square.

4.2. Pseudorange Correction Based on WLAN

The method discussed herein, not only need to use WLAN transmitter transmitting WIFI signals, should also be able to receive satellite signals. The principle of pseudorange correction based on WLAN signal is: because the effective distance of the WLAN transmitter is always between a few meters to tens of meters. Therefore, the error from the WLAN transmitter receives the satellite signals are equal to the error from the receiver receives the satellite signals. Since the actual position of the WLAN transmitter is known, we can compare the position measured by satellites to the actual position to determine the error value. Then the error can be used to correct the pseudorange of the receiver.

Given that the reported position of the i-th satellite is (x_i, y_i, z_i), and the position of the WLAN transmitter is known through a survey to be at position (x_a, y_a, z_a), the computed geometric distance, R_a^i, from the WLAN transmitter to the satellite is [2]

$$R_a^i = \sqrt{(x_i - x_a)^2 + (y_i - y_a)^2 + (z_i - z_a)^2} \quad (7)$$

On the other hand, due to the WLAN transmitter can receive the satellite signals, the WLAN transmitter then makes a pseudorange measurement, ρ_a^i , to the i-th satellite. This measurement contains the range to the satellite, along with the errors:

$$\rho_a^i = R_a^i + \varepsilon_a \quad (8)$$

Where ε_a are the pseudorange errors.
Therefore, the equations (3) ~ (6) become

$$\rho_1 - \varepsilon_a = \sqrt{(x_1 - x_u)^2 + (y_1 - y_u)^2 + (z_1 - z_u)^2} + ct_u \quad (9)$$

$$\rho_2 - \varepsilon_a = \sqrt{(x_2 - x_u)^2 + (y_2 - y_u)^2 + (z_2 - z_u)^2} + ct_u \quad (10)$$

$$\rho_3 - \varepsilon_a = \sqrt{(x_3 - x_u)^2 + (y_3 - y_u)^2 + (z_3 - z_u)^2} + ct_u \quad (11)$$

$$\rho_a - \varepsilon_a = \sqrt{(x_a - x_u)^2 + (y_a - y_u)^2 + (z_a - z_u)^2} + ct_u \quad (12)$$

Then, we can obtain (x_u, y_u, z_u) by solving Equations (7) ~ (12).

5. Results

In order to verify the method described in this article, we designed an experiment. We prepared two smart phones with GPS hardware module. One of them as WLAN signals emitter, the other as a receiver. As a test experiment, we only used three satellite signals to test, no matter more satellite signals could be received. We wrote a test program, and got more than 1000 groups of data in different locations. In the table below shows a set of data:

TABLE 3. Pseudorange errors ($\mathcal{E}_a$) and GPS error after correction

Outdoor environment			Indoor environment		
No	$\mathcal{E}_a$ **(m)**	**GPS error (in the horizontal plane) (m)**	**No**	$\mathcal{E}_a$ **(m)**	**GPS error (in the horizontal plane) (m)**
1	32.43	6.01	1	96.33	17.52
2	52.09	10.12	2	85.94	15.00
3	58.77	9.89	3	73.10	15.13
4	43.10	8.35	4	128.95	24.54
5	25.21	5.89	5	91.97	16.97
6	15.20	2.47	6	65.57	14.37
7	27.45	5.73	7	60.71	14.08
8	33.67	6.25	8	34.23	9.85
9	30.66	5.90	9	45.51	11.18
10	19.52	3.05	10	30.65	10.01

6. Conclusions

In this paper, the proposed method can effectively improve GPS positioning precision. The method and the other GPS optimization algorithm does not conflict. We can put them together to further improve the accuracy. Based on the method of this paper, we designed the test environment and test data were obtained. It can be seen that, the accuracy of GPS positioning improved by using the method of this paper. Subsequent work will focus on enhancing the stability of this method in different environments.

References

1. Mautz, R. (2012) Indoor Positioning Technologies, Institute of Geodesy and Photogrammetry, Department of Civil, Environmental and Geomatic Engineering, ETH Zurich.
2. Kaplan, E.D., Hegarty, C.J. (2006) Understanding GPS Principles and Applications second edition, the USA, Artech House.
3. Hightower, J., Borriello, G. (2001) Location systems for ubiquitous computing, *IEEE Computer*, August, 34(8), 57-66.
4. Battiti, R., Nhat, T.L., Villani, A., (2002) Location-aware computing: a neural network model for determining location in wireless LANs, Technical Report DIT-02-0083, February 2002.
5. Roos, T., Myllymaki, P., Tirri, H., Misikangas, P., Sievanan, J. (2002) A probabilistic approach to WLAN user location estimation, International Journal of Wireless Information Networks, 9(3), July 2002, 42-46.
6. Ciurana, M., Barcelo-Arroyo, F., Izquierdo, F. (2007) A ranging method with IEEE 802.11 data frames for indoor localization, WCNC 2007 proceedings, Wireless Communications and Networking Conference (WCNC), March 2007, HongKong, 2095-2098.

Received on the 20th of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 229-235
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

RESEARCH ON IMAGE INPAITING WITH ADAPTIVE DECOMPOSITION IN FREQUENCY DOMAIN

M. Wu[1], H.Wang[2], Zh. Zhang[3]

[1]School of Management
Xi'an University of Architecture and Technology, Xi'an 710055, Shaanxi, China
E-mail: 15006244@qq.com
[2]School of Information and Control Engineering
Xi'an University of Architecture and Technology, Xi'an 710055, Shaanxi, China
E-mail: whq463@263.net
[3]School of civil engineering
Xi'an University of Architecture and Technology, Xi'an 710055, Shaanxi, China
E-mail: 370331615@qq.com

This paper is concerned with image decomposition algorithm, which is good for the simultaneous filling-in of texture and structure image. The basic idea is to decompose the damaged image into two function with pule texture and pule structure image, and then reconstruct each part with pertinent method to get a batter effect than using one single algorithm. In this paper we put lots of tests using wavelet transform to get the α factor which can describe the average energy of the damaged image, so that the images with different energy could be decomposed adaptively during DCT. We present the underlying examples to show that the optical connectivity requirements of human beings can be satisfied by the inpainted effects.

Keywords: image inpainting, self-adaptive decomposition, wavelet transform, DCT

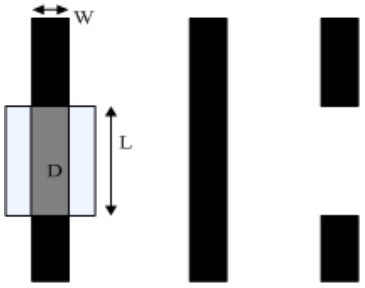
1. Introduction

Image inpainting technique is that repairs damaged area or remove areas in an image. It is used to restore the old photos, remove undesired object, and repair digital video or cultural relic images. The ultimate goal is to maintain reasonable visual perception after image inpainting. It introduced by Bertalmio et al., who proposed the famous BSCB model [1]. They adopted PDE (the partial differential equations) into image restoration techniques, using the edge information of the area to be patched. The transmission is estimated by using the isopods direction to disseminate source information to the damaged regions. BSCB is complex and slow, but it provides the basis for image inpainting.

2. The Research of Image Inpainting and Their Problems

In 2001, Chan et al. proposed TV (Total Variation) mathematical model [2] and (Curvature Driven Diffusion) CDD model [3] based on the BSCB model. TV is established by the constraint of the Euler - Lagrange equation, the algorithm retouches the damaged image for overall situation, not only can remove image noise, but also can sharpen the image edges. It is widely applied in the retouch of damaged area as cracks and scratches. This is caused by the retouch model of the algorithm itself, which main idea is to make a physics model of the heat flow partial differential equations and calculated the pixels to fill the empty linear structure, and then through a diffusion process to put the colours to the user-defined mask part, in order to achieve the effect of the complete image. We can see the demonstration in Figure 1, (a) is for the missing part of information lines, the width is w, the missing region is denoted D. When the length $L > w$ of D, the image impainting result is showed as (c), and (b) is the artificial assumption according to human visual.

We can see from the above that TV model can well maintain the edge of the image, but it cannot accord the principle of connectivity in visual theory. In order to improve it, Criminisi and Toyama et al. [4] proposed a new texture and structure synthesis algorithm based on the sample blocks in 2004. Their algorithm analysis image structure and texture characteristics and calculate priority value of the edge to choose the reasonable inpainting order, and then find the optimal sample blocks to fill the unknown area. However, it is easy to lead to the blocking effect.



(a) (b) (c)

Figure 1. Comparison of the effect between human visual and TV model

This paper presents a novel algorithm that combines the advantages of these two approaches. Firstly, we use DCT to decompose the damaged image into pure structure part and pure textures part; secondly, we use wavelet transform to calculate the energy function of texture feature, and chose the adaptive position of decomposition; thirdly, we use the exemplar-based texture synthesis algorithm to inpaint the pure texture image and use the improved CDD algorithm to inpaint the pure structure image; finally, we add the two pats to obtain the reconstructed result. It showed there is a better optical connectivity requirement of human beings by the inpainted effects.

3. Image decomposition

In 2003, Bertalmio et al. [5] proposed an inpainting algorithm based on image decomposition which core idea is to decompose an image into two parts of structure image and texture image. They use the decomposition method follows the work by Vese an Osher [6], and they use the total variation minimization of [7] for image denoising to get the structure image, and the texture image can be calculated by subtracting structure from the original image. However, this method changes the effective information, so it is not suitable for inpainting the special images such as medical image, cultural relic image and aerial image. Decomposition in frequency domain can improve it.

4 .Self-Adaptive Decomposition in Frequency Domain

4.1. DCT-Based Image Decomposition

After lots of decomposition tests such as DCT, wavelet and contour wave, we find that wavelet decomposition and contour wave decomposition are not well retain the mask image which we set the grey value of the defect area as 255. Thus, we select the two-dimensional DCT to decompose the image.

Let f(x,y) be the image function, size $N \times N$, take 2-D DCT to produce an array coefficients.

$$F(u,v)=\frac{2}{N}C(u)C(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}f(x,y)\cos\frac{(2x+1)u\pi}{2N}\cos\frac{(2y+1)v\pi}{2N} \tag{1}$$

u, v, x, y = 0, 1, 2…..N-1, among them: x, y is the spatial coordinates. u, v is the Frequency domain coordinates.

$$C(u), C(v)=\begin{cases}\frac{1}{\sqrt{2}} & for\ u,v=0\\ 1 & otherwise\end{cases} \tag{2}$$

We obtain DC coefficient and AC coefficients from the upper left corner to the lower right corner, which are low frequency, medium frequency and high frequency region, shown in Figure 2.

Image DCT coefficients by ZigZag scanned order as seen in Figure 3. We use different decomposition coefficients on the experimental image decomposition. As shown in Figure 4.

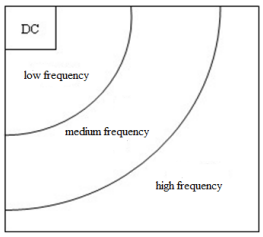
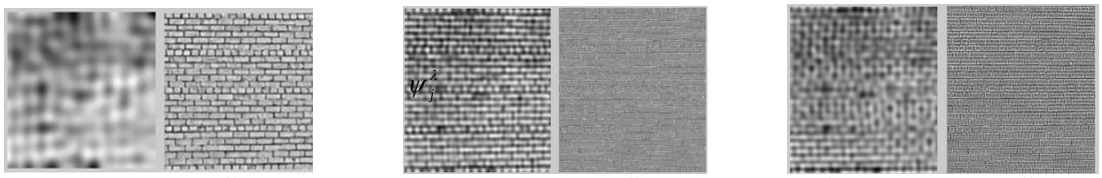


Figure 2. Band of DCT coefficients

0	1	5	6	14	15	27	28
2	4	7	13	16	26	29	42
3	8	12	17	25	30	41	43
9	11	18	24	31	40	44	53
10	19	23	32	39	45	52	54
20	22	33	38	46	51	55	60
21	34	37	47	50	56	59	61
35	36	48	49	57	58	62	63

Figure 3. ZigZag scan diagram



(d)Take the top 4% coefficient (e) Take the top 5% coefficient (f) Take the top 7% coefficient

Figure 4. Comparison of the DCT decomposition at different position

As the figure 4 shows, the choice of the position has a large effect on decomposition, and different image need different position.

4.2. Wavelet Transform for the Average Energy

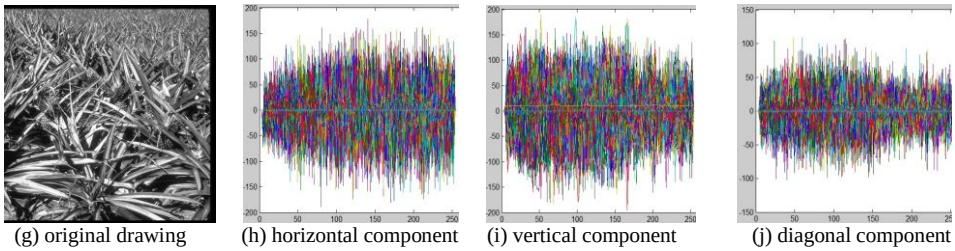
We use 2-D wavelet transform to calculate the energy of texture features of image. For the image f (x, y), wavelet transform is following:

$$W_j^i f(x,y) = \int_{R^2} \int f(u,v) \psi_j^{\lambda}(x-u, y-v) du dv \tag{3}$$

There j is the decomposition scale, λ is the one low frequency component and the other three different high frequency components. ψ_j^{λ} is the 2-D wavelet generating function under j. The wavelet transform can be decomposed into following:

$$\begin{cases} C_{j+1} = HC_jH' \\ D_{j+1}^h = GC_jH' \\ D_{j+1}^v = HC_jG' \\ D_{j+1}^d = GC_jG' \end{cases}, \ (j = 0,1,2,...J-1) \tag{4}$$

There H and G are the filter coefficient matrices which correspond to scaling function, the original image recorded as. h, v, d are horizontal, vertical and diagonal components. As showed in picture 5.



(g) original drawing (h) horizontal component (i) vertical component (j) diagonal component

Figure 5. Haar wavelet decomposition

Define the energy function of image details:

$$E_{d_1^1}(a,b)=\sum_{m=1}^{b/2}\sum_{n=1}^{a/2}D_1^h(n,m)^2$$
$$E_{d_1^2}(a,b)=\sum_{m=1}^{b/2}\sum_{n=1}^{a/2}D_1^v(n,m)^2 \quad (5)$$
$$E_{d_1^3}(a,b)=\sum_{m=1}^{b/2}\sum_{n=1}^{a/2}D_1^d(n,m)^2$$

The $E_{d_1^1}(a,b),E_{d_1^2}(a,b),E_{d_1^3}(a,b)$ indicate the energy of horizontal, vertical and the diagonal coefficients.

The size of image is $a\times b$.

$$E_{averd}(a,b)=\frac{E_{detail}(a,b)}{area} \quad (6)$$

Now standardize the detail energy of image.

$$E_{detail}(a,b)=E_{d_1^1}(a,b)+E_{d_1^2}(a,b)+E_{d_1^3}(a,b) \quad (7)$$

The area is the total pixels. In addition, the average detail energy is according to the definition of texture degree in formulas (4)-(7).

(k) (l) (m)

Figure 6. Image of different texture degrees

The energy value is below:

TABLE 1. Average energy of different texture degrees

	k	l	m
E	73.2506	110.8899	528.9812

4.3. Self-Adaptive Decomposition

The purpose of image decomposition is to decompose Image f into homogeneous part u and oscillatory v, so f is expressed as:

$$f=u+v \quad (8)$$

We chose the best decomposition positions for different image with different energies and obtained the results below:

(n) (o) (p)

Figure 7. Structure part and texture part of different energy images

In figure7, (n), (o), (p) are the structure image and texture image after DCT decomposition.

The experiment results show that the decomposition position changed with the different texture average energies of the image. Through the experiments on a large number of images average energy of the texture detail, we can obtain the relationship of best image decomposition position and texture energy in the following figure 8.

TABLE 2. Different decomposed position

	energy	position
n	84.2138	7%
o	110.8899	5%
p	281.7954	4%

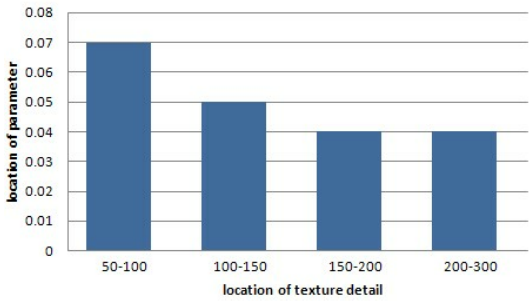


Figure 8. Relationship of the position and energy

By using this law, we design the image inpaiting algorithm with adaptive decomposition in frequency domain that includes:

1) Image wavelet transforms to calculate the average energy of the texture detail;
2) DCT coefficients zig-zag scans;
3) According to the average energy of the texture detail, the image can adaptively choose the coefficient position and be decomposed into two parts;
4) Repair the texture image using the improved template-based image inpainting algorithm and repair the structure image using the improved CDD algorithm.
5) The original image is then reconstructed adding back these two sub-images.

4.4. Image Inpainting

The part of the structure image is inpainted using the algorithm refer to our team Li Suli [8]. The algorithm proposed the CDD inpainting algorithm based on p-Laplace operator, the algorithm is the use of P values. If $l < p < 2$ the algorithms can reduce the staircase effect whereas it can still keep the sharp edges effectively, and be able to overcome the edge model introduced by the harmonic ambiguity.

The part of the texture image is inpainted using the algorithm refer to our team Yang Xiuhong [9]. The algorithm uses adaptive templates to search and match the restored area that is to say that the sizes of these templates alter to local character of the image adaptively. When updating the confidence term, in order to avoid “error match”, that algorithm adopts negative exponential function to make the updated confidence value be inversely proportional to the “cumulative error”.

5. Experimental Results

As shown in Figure 9. image (l) is the damaged image, and it decomposed into image (m) and image (n). Image (o) is the repaired structure image, image (p) is the repaired texture image, and image (q) is the finally result after reconstructed.

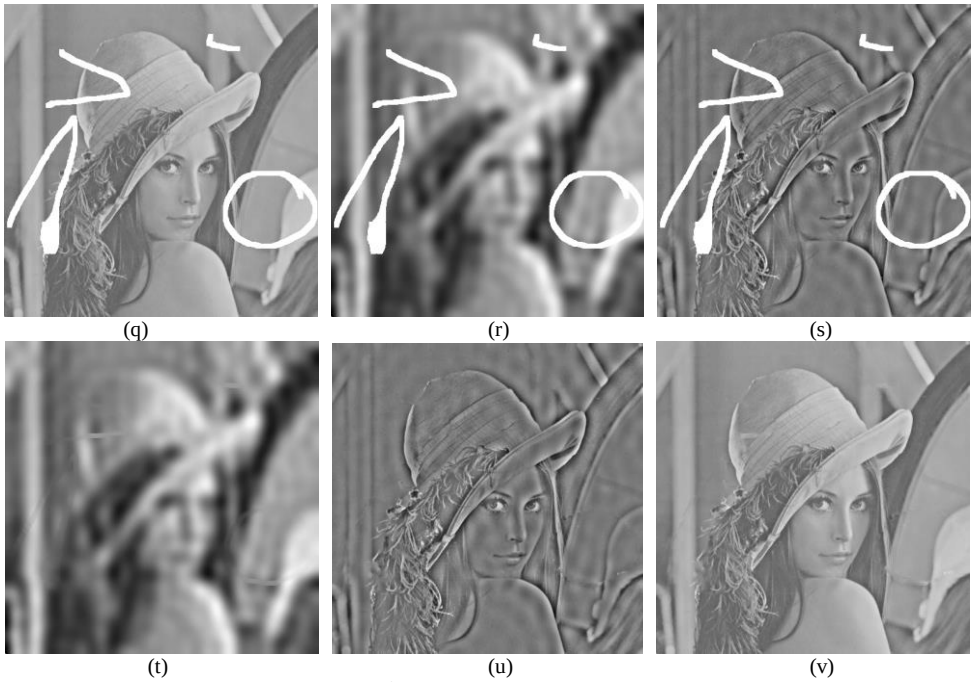


Figure 9. An example of image inpainting with adaptive decomposition

Figure 10 is the test comparison with other previous algorithms.

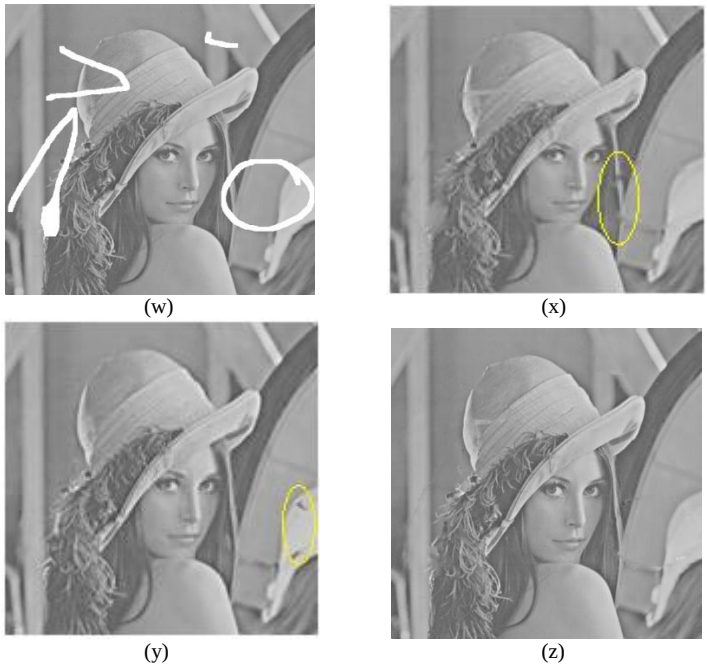


Figure 10. Comparison of experiments

Image (w) is the original image, image(x) is the experimental result of improved CDD algorithm [8], image (y) is the experimental result of improved Exemplar-Based algorithm [9], image (z) is the result of our algorithm.

6. Conclusions

In this paper, we have presented a frequency domain adaptive decomposition inpainting algorithm which can provide satisfactory visual results. It set the parameter to make the decomposition changed with different texture feature images. We first use of wavelet transform estimate the image texture details; and then classify images with different texture energy; and last, different types of image were selected the appropriate α value adaptively. These algorithm can inpaint damaged images contain structure and texture characteristics simultaneously, and its decomposition does not change the pixel from original image, so this is better for the structural texture image. Therefore, it will be efficient and superior to others.

Further work include the removal of the target combined with image segmentation technology, and an automated inpainting mask marker process.

Acknowledgements

This paper is supported by Scientific Research Foundation for Returned Scholars, Ministry of Education of China (K05055), Special fund of the Education Department of Shaanxi Province (11JK1027), Natural Science Foundation of Shaanxi Province of China (2013JM8015), and Xi'an University of Architecture and Technology Development Program (QN1022).

References

1. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C. (2000) Image inpainting [C]. In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co, SIGGRAPH, 417-424.
2. Chan, T.F., Shen, J.H. (2001) Non-texture inpainting by Curvature-Driven Diffusions (CDD), *Journal of Visual Communication and Image Representation*, 12(4), 436-449.
3. Chan, T.F., Shen, J.H. (2001) Mathematical models for local Non-texture inpainting, *SIAM Journal on Applied Mathematics*, 1019-1043.
4. Criminisi, A., Pérez, P., Toyama, K. (2004) Region filling and object removal by exemplar-based image inpainting. [J]. In: *Image Processing, IEEE Transactions on*, 13(9), 1200-1212.
5. Bertalmio, M., Vese, L., Sapiro, G., Osher, S. (2003) Simultaneous texture and structure image inpainting. In: *IEEE Transaction on Image processing*, 12(8), 882-889.
6. Vese, L., Osher, S. (2003) Modelling Texture with Total Variation Minimization and Oscillating Patterns in Image Processing [J]. In: *Journal of Scientific Computing*, 19(1), 553-572.
7. Rudin, L., Osher, S., Fatemi, E. (1992) Nonliner total variation based noise removal algorithms, In: *Physica D*, (60), 259-268.
8. Li, S.-L., Wang H.-Q., Yang X.-H. (2010) Image Inpainting Using CDD Based On P-Laplace [J]. In: *Computer Simulation*, 189-193.
9. Yang, X.,Wang, H., Li, S. (2009) Image Inpainting Algorithm Based on the Adaptive Template and Update of the Confidence Term[J]. In: *Electronic Science and Technology*, 69-72.

Received on the 20th of June 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 236-243
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

COMPRESSION OF POINT SET SURFACES: A SURVEY

Chen He, Die Wang, Xueqing Li

School of Computer Science and Technology, Shandong University,
Jinan, Shandong, P.R.China
Phone: (+86)-15966347617. E-mail: imhechen@163.com

With the increasing capability of 3D data scanning devices, point-sampled geometry has gained increasing attention as an alternative surface representation, both for efficient rendering and for flexible geometry processing of highly complex 3D models. When representing complex shapes in a raw data format, point set surfaces, consume a large amount of space. Applications calling for compact storage and fast transmission of 3D point set surfaces have motivated the multitude of algorithms developed to efficiently compress these datasets. In this paper, we survey recent developments in compression of 3D point set surfaces. We make a comprehensive study, classify various algorithms into categories, and describe the typical algorithms in each category. We also list some trends in 3D point set surface compression for the future research.

Keywords: Compression, Coding, Point sampled geometry, Point-based models, Survey

1. Introduction

In recent years, three-dimensional (3D) models have become an important digital media type besides the audio, image and video. They find applications in many fields such as gaming, animation, and scientific visualization. The dominant form of representation for 3D models has been polygonal mesh, mostly due to the native support in modern graphics cards. However, with the increasing capability of 3D data scanning devices, there is an increasing interest in point-based representation, which defines a geometric shape merely by a sufficiently dense cloud of sample points on its surface. As a result, increasingly more point set surfaces data are being produced, which gives rise to a growing demand on effective compression schemes for efficient utilization of the storage, the processing and the bandwidth resources.

Many compression schemes have been proposed for point set surfaces in the last one decade. These algorithms can be classified as single-rate and progressive encoders. A single-rate encoder usually traverses the model surface driven by the local correlation of the neighbouring points, while a progressive encoder typically encodes an initial coarse model and thereafter a sequence of local refinements to the model. In this paper, we make a comprehensive study of these algorithms and survey the main ideas and intuition behind techniques for single-rate and progressive encoders.

The rest of the paper is organized as follows. Section two describes some basic knowledge about the point-based representations. Section three gives the classification method of the point set surface compression algorithms. Section four and five survey the single-rate and progressive compression techniques respectively. Section six concludes and discusses some new trends in point set surface compression research.

2. Point-Based Representations

Points by themselves are dimensionless primitives and hence a local region of influence is typically assigned to each point. The local region of influence can be surface-based or volume-based. Surface-based point representations such as Surfels [1] and implicit surfaces [2, 3] approximate the scanned datasets well at high resolutions and are usually topology sensitive. On the other hand, volume-based representations such as spheres [4] and octree cells [5-7] are topology blind and easy to hierarchically organize.

A compact representation is beneficial for compression algorithms. In this section, we will introduce two most commonly used surface-based point representation methods: surfel set surfaces and moving least squares (MLS) surfaces. The volume-based representation methods will be described latter in the section of progressive compression algorithms.

2.1. Surfel Set Surfaces

Pfister et al. [1] propose to use surfel (surface element) as a rendering primitive. A surfel is a zero-dimensional *n*-tuple with shape (or geometry) and shade (or texture colour) attributes that locally approximate an object surface. In order to bridge the gaps between neighbouring point samples, surfels are also associated with a radius r_i, turning them into object-space circular disks. Fig. 1 is an example of surfel set surface. The left is the 'gnome' model without radius for each surfel, and the right is the model with radius.

In 2002, Zwicker et al. [8] present a system, called Pointshop3D, for interactive shape and appearance editing of 3D point-sampled geometry. By generalizing conventional 2D pixel editors, this system supports a great variety of different interaction techniques to alter shape and appearance of 3D surfel set models, including cleaning, texturing, sculpting, carving, filtering, and re-sampling. The source code and details about this software can be obtained from this website: http://graphics.ethz.ch/pointshop3d/.

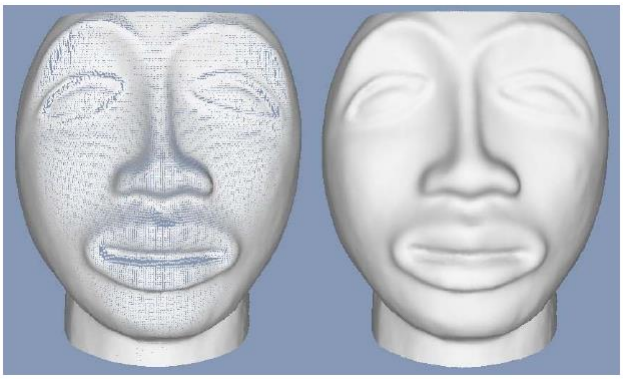


Figure 1. An example of surfel set surface. Left: surfels without radius; Right: surfels with radius. Data courtesy the website of Pointshop3D.

2.2. MLS Surfaces

The moving least squares surfaces of Levin [9, 10] provide an approximating or interpolating surface for a given set of point samples by local higher order polynomials and have first been applied to point-based methods by Alexa et al. [2]. The MLS surface S_P of a set of points $P = \{p_i\}$, $p_i \in R^3$, $i \in \{1, \ldots, N\}$ is defined implicitly by a projection operator. The projection of a point r onto S_P is split into two steps.

First, a local reference domain $H = (\mathbf{n}, \mathbf{d})$, where $\mathbf{d}$ is the origin of the plane and $\mathbf{n}$ is the normal to the plane is computed. More precisely, the local reference domain $H = \{\mathbf{x} | \langle \mathbf{n}, \mathbf{x} \rangle - \mathbf{d} = 0, \mathbf{x} \in R^3\}$, $\mathbf{n} \in R^3$, $\|\mathbf{n}\| = 1$ is determined by minimizing

$$\sum_{i=1}^{N} (< \mathbf{n}, \mathbf{p_i} - \mathbf{r} - t\mathbf{n} >)^2 e^{-\|\mathbf{p_i} - \mathbf{r} - t\mathbf{n}\|^2 / h^2}$$

in all normal directions $\mathbf{n}$ and offsets t. Here $\mathbf{d} = \mathbf{r} + t\mathbf{n}$. A local coordinate system over H is defined by taking the standard basis for R^3, e_1, e_2, e_3 and computing a rotation matrix $\mathbf{M}$ such that $\mathbf{n} = \mathbf{M} \cdot e_3$. Then the local coordinate system is defined by ($\mathbf{M} \cdot e_1$, M e_2, $\mathbf{n}$).

Then, a local bivariate polynomial is fitted over H to the point set. Let $\mathbf{q}$ be the projection of $\mathbf{p_i}$ onto H, and f_i the height of $\mathbf{p_i}$ over H, i.e. $f_i = \mathbf{n}\,(\mathbf{p_i} - \mathbf{q})$. The polynomial approximation g is computed by minimizing the weighted least squares error

$$\sum_{i=1}^{N} (g(x_i, y_i) - f_i)^2 e^{-\|\mathbf{p_i} - \mathbf{r} - t\mathbf{n}\|^2 / h^2}$$

The projection of $\mathbf{r}$ is given by $MLS(\mathbf{r}) = \mathbf{r} + (t + g(0, 0))\mathbf{n}$.

Formally, the surface S_P is the set of points that project onto themselves. In this definition, h is the anticipated spacing of the points. Thus, the point set together with an adequately chosen value for h defines the surface. Fig. 2 gives an MLS surface representation of a statue of an angel. The density of points and, thus, the accuracy of the shape representation are changing (intentionally) along the vertical direction.

3. Compression Algorithms Classification

A good number of compression algorithms have been proposed for point set surfaces since 2000. We make a comprehensive study of these algorithms and classify them into different classes.

A portion of the research on point set surface compression has been conducted in the context of efficient rendering [4, 6, 11-15], rather than for the purpose of storage and transmission. Since compact representation of surfaces is not the sole aim of these works, the compression performance they have achieved may not be sufficiently high.

Many general purpose point set surface compression algorithms have also been proposed, which do not specifically target at rendering applications. These algorithms can be classified as single-rate coders [16-18] and progressive coders [3, 19-25].

In order to obtain a high rendering efficiency, most of the compression algorithms for rendering have a progressive feature. So these algorithms can also be seen as progressive coders. In the following two sections, we will describe the point set surface compression algorithms as two categories: single-rate compression and progressive compression.

Figure 2. An MLS surface representation for a statue of an angel. The density of points and, thus, the accuracy of the shape representation are changing (intentionally) along the vertical direction. Data courtesy Alexa et al. [2].

4. Single-Rate Compression

In a single-rate, point set surface compression algorithm, all geometry data are compressed and decompressed as a whole. In the environment of internet, the model can be decoded only after the entire bit stream has been wholly received.

Compared to progressive compression, there are relatively few research works on single-rate compression for point set surfaces. All the three single-rate coders [16-18] use a spanning tree to make predictions during the encoding process. Unlike in mesh compression, where prediction techniques are commonly used, there is no connectivity information in point set surface to determine the neighbouring points that best predict a new one. In general, it is not possible to permute the points into a sequential order such that each point could simply use its predecessors for prediction. Thus a prediction tree, i.e., a spanning tree over the vertices, needs to be constructed. Having the prediction tree, a point can be predicted, using its parent and/or grandparent. While Gumhold et al. [16] only use a linear predictor, Merry et al. [18] use both a linear predictor and a lateral predictor. Chen et al. [17] do not make an initial quantization to the points' attributes and therefore can achieve a truly lossless compression.

Generally, a spanning-tree-based single-rate compression algorithm consists of the following steps.

Step 1. Constructing a spanning tree. First, a complete weighted graph G is constructed where the nodes are the vertices $v_1, v_2, \ldots, v_n$ of the point cloud and the weight on each edge is the coordinate-wise 3-tuple of differences. That is, if $v_i = (x_i, y_i, z_i)$ and $v_j = (x_j, y_j, z_j)$ then the edge (v_i, v_j) has weight

$(x_i - x_j, y_i - y_j, z_i - z_j)$. The tuples of v_i and v_j are extended similarly when attribute values are included, with the weight defined analogously. After constructing the complete weighted graph G, the spanning tree can be obtained from G using the well-known Kruskal's or Prim's minimum-spanning-tree (MST) algorithm. Fig. 3 gives an example of spanning tree for the Stanford bunny model.

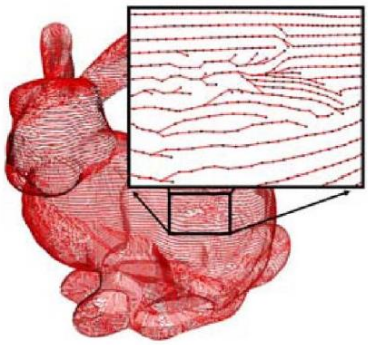


Figure 3. The spanning tree for the Stanford bunny model. Data courtesy Gumhold et al. [16].

Step 2. Selecting prediction rule(s). In these three papers [16-18], three prediction rules are mentioned: constant prediction, linear prediction and lateral prediction. In constant prediction the new point p_i is predicted from the parent point p_j alone, by placing the predicted point $\bar{p}_i$ in the same position as p_j ($\bar{p}_i = p_j$). In linear prediction, the new point is predicted based on its two generation ancestors — its parent and grandparent: $\bar{p}_i = 2p_j - p_k$ (p_k is the parent of p_j). That is to say, the difference between parent and child is predicted to be the same as the difference between the grandparent and the parent. In lateral prediction, three are three possible predictions for point p_i, as shown in Fig. 4. Suppose the difference between the grandparent p_k and parent p_j is $\delta = p_j - p_k$, the estimated normal for p_j is **n**, and the implied "left" vector $L = \mathbf{n} \times \delta$. Then the possible predictions for point p_i are $p_j + \delta$, $p_j + L$ or $p_j - L$. The one which has the smallest difference with p_i is selected as the final prediction for point p_i.

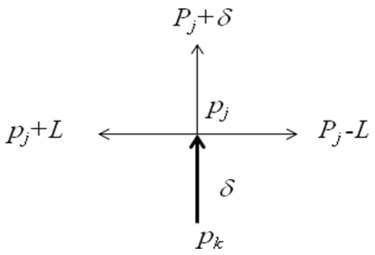


Figure 4. The lateral predictors. The bold arrow indicates the known edge, and the estimated normal **n** points out of the page. The other arrows indicate the predicted left, forward and right edges. Figure reproduced from [18].

Step 3. Encoding. There are two types of information need to be encoded: the spanning tree and the correction vectors. The spanning tree is encoded in a breadth first order. The valence is encoded for each node. As most of the valences are 0, 1 or 2 arithmetic coding allows compressing the valences to less than two bits per vertex. The vertices' position is encoded via the correction vectors $r_i = p_i - \bar{p}_i$ from the predicted position $\bar{p}_i$ to the original position p_i. The vectors r_i are represented in the global coordinate frame of the model. After quantization of the correction vectors the integer valued coordinates can be encoded by the adaptive arithmetic coder.

5. Progressive Compression

Progressive compression of 3D point set surface is desirable for transmission of complex models over networks with limited bandwidth. In progressive compression and transmission, a coarse model is transmitted and rendered first. Then, refinement data are transmitted to enhance the model representation until the model is rendered in its full resolution or the transmission task is cancelled by the user. In other words, progressive compression allows transmission and rendering of different levels of details. Progressive compression can be implemented by different methods. We will briefly introduce three typical schemes for progressive compression in the following sub-sections.

5.1. Compression Based on MLS Surfaces

Based on the properties of the MLS surfaces [2], Fleishman et al. [3] propose a progressive point set surface (PPSS) representation. Starting from a base point set, a refinement rule using the MLS projection operator can construct a PPSS from any given manifold surface. In the following, the MLS surface with respect to a point set X will be denoted as S_X.

Given a reference (input) point set $R = \{r_i\}$ defining a reference surface S_R. The point set R is reduced by removing points to form a base point set $P_0 \subset R$. This base point set defines a surface S_{P0} which differs from S_R. Next, the surface will be re-sampled by adding more points so that the difference between the surface and S_R decreases.

The base point set P_0 is refined by inserting additional points yielding the set P_1. The refinement operator first inserts points independent of the reference set R, which means $P_1 \not\subset R$. Then, the inserted points are displaced so that the difference between the surfaces decreases, i.e. $d(S_{P1}, S_R) < d(S_{P0}, S_R)$.

This process is repeated to generate a sequence of point sets P_i with increasing size and decreasing difference from the reference surface. A progressive point set surfaces is defined as the MLS surface of $\overline{P} = P_0, P_1, \ldots$. Where each point set P_i is encoded by the (scalar) displacements of inserted points, yielding a compact representation of the progressive point set surface.

5.2. Compression Based on Octree Decomposition

By octree decomposition, point set surfaces can be easily organized hierarchically. Thus, several research works on progressive point set surface compression are based on octree decomposition [21-23, 25]. Among these works, Huang et al. [23] propose a generic point cloud encoder that provides a unified framework for compressing different attributes of point samples corresponding to 3D objects with arbitrary topology. In the following, we will take this paper for example to describe the main idea of octree-based compression algorithms. The coding process can be divided into two stages: octree construction and encoding.

Octree construction. The proposed encoder recursively and uniformly subdivides the smallest axis-aligned bounding box of a given model into eight children in an octree data structure. Fig. 5 shows a simple illustration of octree cell subdivision and the tree structure. Only the nonempty child cells will be subdivided further. The part of the model within each cell is represented by its cell's attributes — the position of each cell is represented by the geometric centre of the cell, and the normal/colour of each cell is set to the average of normals/colours of contained points. The attributes of nonempty cells in each level in the octree structure yield a Level of Detail (LOD) of the original 3D model.

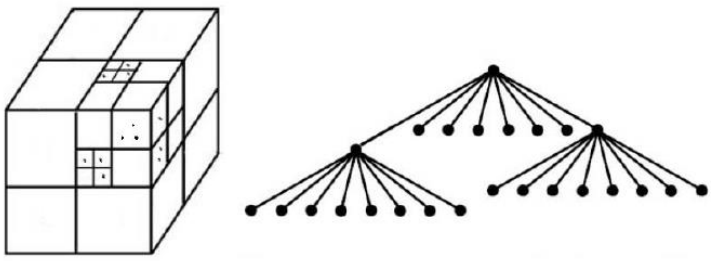


Figure 5. Octree cell subdivision and octree structure

Encoding. The efficiency of the proposed coding scheme lies in effective coding of LODs of the model represented by the octree data structure. In association with each octree cell subdivision, the position, normal and colour attributes of each nonempty child cell will be encoded. Each point in an LOD is called as a *representative*.

The position of each cell is implicit as the subdivision of a cell is uniform and the centre of the cell can be computed from the position of the parent cell. So the position can be encoded by encoding the sequence of nonempty child cells. In the proposed position coder, a 1-bit flag is used to signify whether a child cell is nonempty, with '1' indicating a nonempty child cell and '0' an empty child cell. For each octree cell subdivision, by traversing all child cells according to a fixed order, and collecting the flag bits of all child cells, an 8-bit code will be obtained. This code is called the *occupancy code*, which has to be coded. In order to reduce the entropy of occupancy codes, the '1'-bits are "pushed" toward an end by

reordering the bits in each occupancy code. The key in occupancy code reordering is probability estimation. We refer the intrigued readers to [23] to find more details about the reordering method.

The normal attribute is first quantized based on uniform subdivision of the unit sphere. When the normal information needs to be refined for an octree cell subdivision, normals of children are predicted by the normal of their parent, and their residuals are coded. On the unit sphere, quantized normals around the predicted normal are locally sorted and indexed, resulting in a reduced entropy of normal residual indices.

Before colour coding, PCA is first performed on the colour data of the model to determine a new colour frame where an oriented bounding box of colour samples is calculated. Then, the generalized Lloyd algorithm (GLA) is used to calculate the quantization ranges, or representatives, along each dimension of the oriented bounding box. This adaptive quantization reduces the number of quantization bins (thus, the number of representational bits) for a given quantization error threshold. When the colour information needs to be refined for an octree cell subdivision, each child colour is predicted to be the same as its parent colour, and the residual is encoded.

5.3. Compression Based on Height Field Approximation

Based on the height field approximation, Ochotta et al. [20, 24] propose a compression scheme that is capable of processing densely sampled surfaces. To approximate a given surface, the encoder decomposes the model into a set of patches, each of which is parameterized as an elevation map over a planar domain and resampled on a regular grid. The resulting representation consists of a set of images with arbitrary regions of support. The encoder compresses the shapes by using a binary image coder and the elevation images using a shape adaptive wavelet coder. The decoder extracts the binary images and reconstructs the elevation images by wavelet synthesis, followed by performing the deparameterization.

Surface partition. The goal of the partitioning procedure is to decompose the surface S into a set of patches, such that each patch can be parameterized as an elevation map. In [24], two partitioning methods are proposed: split-merge clustering and Generalized Lloyds algorithm (GLA). Fig. 6 shows the partition results for Igea model using these two different methods.

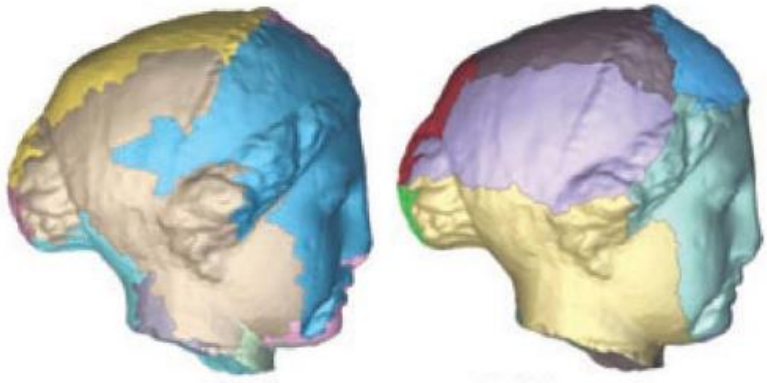


Figure 6. Surface partition results for Igea model using different methods. Left: split-merge clustering; Right: GLA algorithm. Figure reproduced from [24].

Resampling. After constructing the partition, each surface patch is resampled on a regular grid in order to apply the shape-adaptive wavelet coder. In the resampling procedure, they compute for each grid point q a corresponding elevation value $f(q)$. This value is computed iteratively with a Newton-like method using the MLS projection operator, which is described in section 2.2.

Encoding. For each patch, there are three types of data need to be encoded: side information, binary masks and the scalar height field values. Side information, consisting of the base plane rectangle and its grid sampling resolution, can be encoded using a few quantized floating point numbers. Binary masks, which define the support of the height field on each base plane rectangle, are encoded using an adaptive context-based arithmetic coder. Elevation maps can be regarded as grey scale images with irregular boundaries. They are encoded using a shape-adaptive wavelet coder.

6. Conclusion and Future Work

In this paper, we performed a survey on current 3D point set surface compression techniques by classifying major algorithms and describing main ideas behind each category. Currently, the state-of-the-art compression schemes are those based on the octree decomposition [23] and the height field approximation [24].

Future research for point set surface compression can be conducted by exploiting the semantic information such as symmetry or regularity of the shape to markedly reduce the redundancy in the 3D models. Besides, the dynamic point clouds compression will be another promising research area.

Acknowledgements

This work was supported by the Graduate Independent Innovation Foundation of Shandong University (Grant No. yzc10130).

References

1. Pfister, H., Zwicker, M., Van Baar, J., et al. (2000) Surfels: Surface elements as rendering primitives[C]//Proceedings of the 27th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., 335- 342.
2. Alexa, M., Behr, J., Cohen-Or, D., et al. (2001) Point set surfaces[C] //Proceedings of the conference on Visualization'01. *IEEE Computer Society*, 21-28.
3. Fleishman, S., Cohen-Or, D., Alexa, M., et al. (2003) Progressive point set surfaces [J]. ACM Transactions on Graphics (TOG), 22(4), 997-1011.
4. Rusinkiewicz, S., Levoy, M. (2000) QSplat: A multiresolution point rendering system for large meshes[C]// Proceedings of the 27th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., 343-352.
5. Woolley, C., Luebke, D., Watson, B., et al. (2003) Interruptible rendering [C] //Proceedings of the 2003 symposium on Interactive 3D graphics. ACM, 143-151.
6. Botsch, M., Wiratanaya, A., Kobbelt, L. (2002) Efficient high quality rendering of point sampled geometry[C]// Proceedings of the 13th Eurographics workshop on Rendering. Eurographics Association, 53-64.
7. Pajarola, R. (2003) Efficient level-of-details for point based rendering[C]//Proceedings IASTED Computer Graphics and Imaging Conference (CGIM). 141-146.
8. Zwicker, M., Pauly, M., Knoll, O., et al. (2002) Pointshop 3D: an interactive system for pointbased surface editing [J]. ACM Transactions on Graphics (TOG), 21(3), 322-329.
9. Levin, D. (1998) The approximation power of moving least-squares [J]. Mathematics of Computation of the American Mathematical Society, 67(224), 1517-1531.
10. Levin, D. (2003) Geometric modelling for scientific visualization [J]. Mesh-independent surface interpolation, Springer-Verlag, 37-49.
11. Kruger, J., Schneider, J., Westermann, R. (2005) Duodecim-a structure for point scan compression and rendering[C]//Point-Based Graphics, 2005. Eurographics/IEEE VGTC Symposium Proceedings. IEEE, 99-146.
12. Kalaiah, A., Varshney, A. (2005) Statistical geometry representation for efficient transmission and rendering [J]. ACM Transactions on Graphics (TOG), 24(2), 348-373.
13. Hubo, E., Mertens, T., Haber, T., et al. (2006) The quantized kd-tree: Efficient ray tracing of compressed point clouds[C]// Interactive Ray Tracing 2006, IEEE Symposium on. IEEE, 105-113.
14. Hubo, E., Mertens, T., Haber, T., et al. (2008) Self-similarity based compression of point set surfaces with application to ray tracing [J]. Computers & Graphics, 32(2), 221-234.
15. Schnabel, R., Moser S., Klein, R. (2008) Fast vector quantization for efficient rendering of compressed point-clouds [J]. Computers & Graphics, 32(2), 246-259.
16. Gumhold, S., Kami, Z., Isenburg, M., et al. (2005) Predictive point-cloud compression[C]//ACM SIGGRAPH 2005 Sketches. ACM, 137.
17. Chen, D., Chiang, Y.J., Memon, N. (2005) Lossless compression of point-based 3D models[C]//Pacific Graphics, 124-126.
18. Merry, B., Marais, P., Gain, J. (2006) Compression of dense and regular point clouds[C]//Computer Graphics Forum. Blackwell Publishing Ltd, 25(4), 709-716.

19. Waschbusch, M., Gross, M., Eberhard, F., et al. (2004) Progressive compression of pointsampled models[C]// Proceedings of the First Eurographics conference on Point-Based Graphics. Eurographics Association, 95-103.
20. Ochotta, T., Saupe, D. (2004) Compression of point-based 3d models by shape-adaptive wavelet coding of multi-height fields[C]// Proceedings of the First Eurographics conference on Point-Based Graphics. Eurographics Association, 103-112.
21. Huang, Y., Peng, J., Kuo, C.C.J., et al. (2006) Octree-based progressive geometry coding of point clouds[C]// Proceedings of the 3rd Eurographics/IEEE VGTC conference on Point-Based Graphics. Eurographics Association, 103-110.
22. Schnabel, R., Klein, R. (2006) Octree-based point-cloud compression[C] //Symposium on point-based graphics. The Eurographics Association, 111-120.
23. Huang, Y., Peng, J., Kuo, C.C.J., et al. (2008) A generic scheme for progressive point cloud coding [J]. Visualization and Computer Graphics, IEEE Transactions on, 14(2), 440-453.
24. Ochotta, T., Saupe, D. (2008) Image-Based Surface Compression[C] //Computer graphics forum. Blackwell Publishing Ltd, 27(6), 1647-1663.
25. Park, S.B., Lee, S.U. (2009) Multiscale representation and compression of 3-D point data [J]. Multimedia, IEEE Transactions on, 11(1), 177-183.

Received on the 20th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 244-250
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

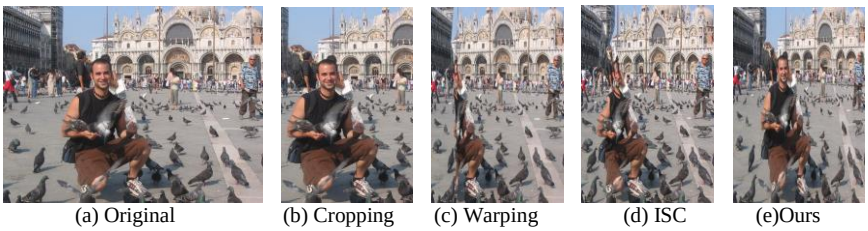
A PRACTICAL MULTI-OPERATOR METHOD FOR CONTENT-AWARE IMAGE RETARGETING

Yan Zhang[1,2], Jonathan Z. Sun[3] and Jingliang Peng[1*]

[1]School of Computer Science and Technology, Shandong University, Jinan, China
[2]School of Science, Shandong Jianzhu University, Jinan, China
[3]School of Computing, University of Southern Mississippi, Hattiesburg, USA
Phone: (+86)-13665315770. E-mail: jingliap@gmail.com

In this paper, we propose a practical multi-operator retargeting method based on optimized seam carving and scaling which can better understand the semantic meaning of the image contents. Considering the entire frequency-tuned saliency map, the face map, the centre prior map and the gradient map in the energy function, it can protect the ROIs effectively when the image is downsized or enlarged. Moreover, a bi-directional search scheme for seam carving and an adaptive switching scheme between seam carving and scaling are employed to minimize the shape distortions of the image. Experiments demonstrate that the proposed algorithm can generate more desirable resized images than cropping, uniform scaling and several other methods including the conventional seam carving approach.

Keywords: image retargeting, saliency map, face map, centre prior map, bi-directional search



(a) Original (b) Cropping (c) Warping (d) ISC (e)Ours

Figure 1. Here we show a comparison of several methods. The original image (left) is retargeted to half width of the original image using: simple cropping, warping [3], ISC [2], and our algorithm

1. Introduction

In recent years, image retargeting has become more and more popular along with the availability of so many display devices of different resolutions and sizes such as mobile phones, cameras and televisions. Traditional image retargeting methods such as cropping and homogeneous scaling all are not sufficient since they pay much attention to the display space limitation but very little attention to the preservation of the image content. To address this problem, many content-aware retargeting methods have been proposed, such as seam carving [1, 2] and non-homogeneous warping [3]. Seam carving attracts a lot of researchers' attention due to its elegancy and flexibility. It is proved that in some cases the retargeting results are more desirable since important information has been preserved. However, the seams may still go through the regions of interest (ROIs) and distort the ROIs in the image when the energy of the ROIs is low. The main drawback of seam carving is that it only considers the pixel information to make the image retargeting unperceivable but not the semantic meaning of image for users. From Figure1 (d) we can see that the shape of the man is severely damaged by seam carving when retargeting the original image to 50% width.

In this paper, we propose a practical multi-operator technique for content-aware image retargeting based on seam carving method. Gradient magnitude, saliency map, face map and the centre prior map are combined together to formulate the operator cost function. Moreover, a bi-directional search scheme is employed which can delete or insert seams in both vertical and horizontal dimensions. When only one dimension of the image is to be changed, the proposed method can change the image in both two dimensions instead of operating in a single dimension as done by the original seam carving method. Considering that single-operator methods might not work well in many cases, the seam carving is combined with homogeneous scaling and an adaptive switching scheme between seam caving and scaling is proposed to minimize the shape distortions of the image.

In summary, the main contributions of our work are as follow:

We optimize the cost function by combining all the gradient map, the face map, the centre prior

map and the saliency map together for semantic-oriented seam carving;

We propose optimized searching in both width and height dimensions for seams to delete or insert.

We combine seam carving and scaling for image retargeting and make an adaptive switch from the former to the latter and, as a result, a good balance between content loss and image distortion is achieved.

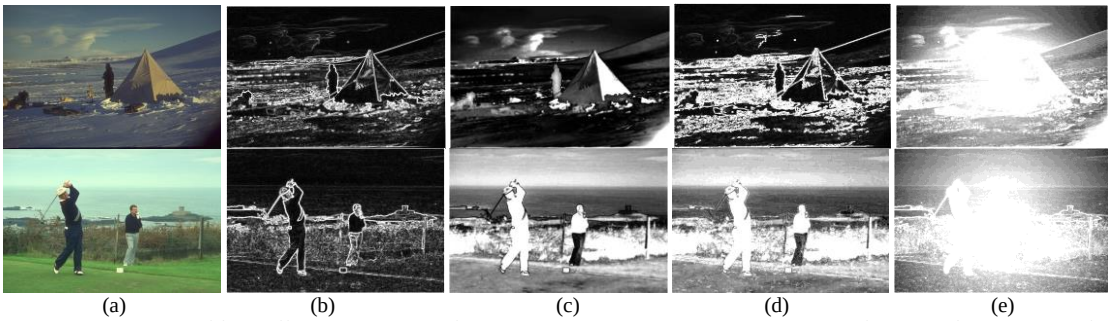
2. Related Work

The increasing need to adapt content to various displays with different sizes and/or aspect ratios has caused a surge in the number of publications dealing with image retargeting. Traditional retargeting approaches like cropping and homogeneous scaling all propagate the distortion throughout the entire image since they do not consider the image content when retargeting the images. To overcome this shortcoming, many content-aware retargeting methods have been proposed which can be broadly divided into two categories as discrete methods and continuous methods.

Warping [3] is a typical one of the continuous image retargeting methods. The main idea is to place a grid mesh onto the image and then compute a new geometry for this mesh to fit the boundaries of the target size. Wang et al. [4] introduce a "scale-and-stretch" warping method (SNS) that can preserve the important objects in the resized image while excessively distorting other objects and damaging the global spatial structure. Liang et al. introduce a patchwise scaling method (PS) in [5]. The final output of this method is satisfactory but may be with shearing around the boundary of important and unimportant objects because the patches covering a prominent feature of the input may be divided into different columns.

The discrete content-aware retargeting is demonstrated by the seam carving (SC) method [1]. However, the ROIs of low energy in the image cannot avoid from being carved out because the way to determine the least perceivable seam considers only the change of intensities around pixels within a seam. To improve the efficiency of seam carving, many attempts have been made. Rubinstein et al. [2] introduce forward energy criteria to the scheme (ISC) that can effectively remove the artificial noise generated by method addressed in [1]. For considering the semantic meaning of image contents and the global visual effect, saliency map [6, 7] and other importance measures are used to measure the importance of the pixels in the image. A human attention model (PSC) was proposed by Hwang and Chien [8]. Considering both face map and saliency map of Itti's [6] as human attention model in the energy function, it can keep perceptually important information. Achanta et al. [9] propose a frequency-tuned salient detection approach which is proved to outperform other saliency detection methods. This approach is used in [10] to compute the importance value of the pixel replacing the gradient map in the seam carving. It is proved that the importance map works well in many cases but may fail when the salient regions of the image are not detected rightly. Daniel Domingues [12] proposes a decision scheme to select one of the gradient map or the frequency-tuned saliency map to combine with other measures guiding the stream carving (STC). However, the decision process cannot utilize the capabilities of the gradient map and the frequency-tuned saliency map complementarily since only one is used.

Another way to improve the effect of image retargeting method is to combine several operators together in order to capture the best aspects of each. Rubinstein et al. [11] present a user study that concludes that users generally prefer to combine cropping, scaling and seam carving together to obtain more pleasing results, rather than relying on the use of a single operator. The methods in references [4] and [8] combine the seam carving with the scaling together by setting a threshold to decide when to switch from seam carving to scaling. Daniel Domingues [12] uses two restricting conditions including the average weight of the last seam and the comparison of the two dimensions to decide when to stop seam carving and begin homogeneously scaling. A fast multi-operator image resizing method (FMO) is developed by Dong et al. [13] that combines seam carving with scaling and cropping.



(a) (b) (c) (d) (e)

Figure 2. Comparison of four different energy maps for retargeting the original images *tent* (upper left) and *golfmen* (bottom left). (a) Original image, (b) Gradient map, (c) Saliency map in [9,10], (d) Energy map of STC [12] ,(e) Our energy map.

3. Practical Multi-Operator Retargeting

In this section, we introduce our practical multi-operator image retargeting method using optimized seam carving and scaling in detail. Our method is based on ISC [2] but makes the following improvements: optimized cost function, bi-directional seam search and adaptive switch between seam carving and scaling, as described in the following subsections.

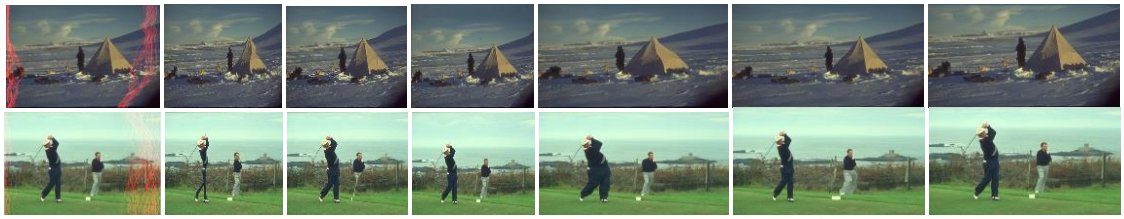
3.1. Optimized Cost Function

Considering that the gradient map cannot work well alone, we incorporate other three measures into our new cost function.

3.1.1. Frequency-tuned Saliency Map

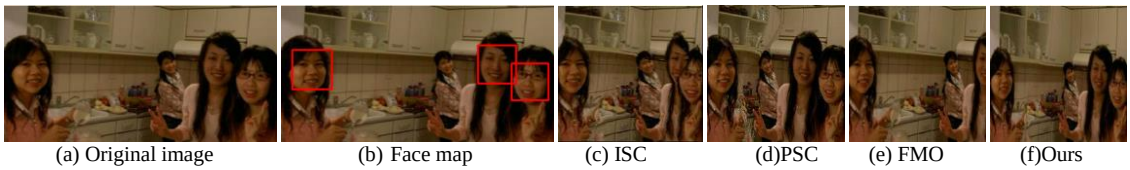
In this paper, we choose the frequency-tuned saliency detector to compute the saliency map of an image. This saliency map is a full resolution saliency map since, when computing the saliency map, the method operates on the original image without any down-sampling. We calculate distances of pixels in a blurred version of the image to the average colour of the original image in the *Lab* colour space since Euclidean distances in this colour space are approximately perceptually uniform. The saliency map *S* for an image *I* is formulated as: $S = \left\| I_\mu - I_G \right\|$, where I_μ is the mean *Lab* vector value of the input image, I_G is the corresponding image pixel vector value in the Gaussian blurred version (using a 5×5 separable binomial kernel) of the original image, and $\| \ \|$ is the L_2-norm.This saliency map has uniformly highlighted salient regions with well-defined boundaries, demonstrating both higher precision and better recall than several state-of-the-art methods. Moreover, high saliency values are not assigned just at the edge of a region, but the entire region. Once we know which pixels are less salient with respect to the original image, we can remove them without having to re-compute their importance after each removal.

However, as mentioned in Section 2, the frequency-tuned saliency detector may still sometimes fail to detect the salient regions correctly. From Figure 2 we can see that this saliency detector may not work well alone and neither may the gradient map. The gradient map of image *tent* is shown in Figure 2(b) without detecting the salient region of the tent while the saliency map from [9, 10] showed in Figure 2(c) can. On the other hand, the saliency map of the image *golfmen* from [9, 10] does not detect the club and the legs of the man in the right while the gradient map can. Daniel Domingues [12] proposed a decision scheme to select either the gradient map or the frequency-tuned saliency map (but not both) to combine with other measures. The combined measures are used to guide the seam carving process. However, the decision process cannot utilize the complementary capabilities of the gradient map and the saliency map since only one is used. The importance maps computed using method [12] are shown in Figure 2(d). We can see that they still fail to recognize the tent, the club and the legs of the man in the right as ROIs.



(a) (b) (c) (d) (e) (f) (g)

Figure 3. Image retargeting comparison for 80% and 120% of original width:
(a) is the original images, (b) and (e) are the outputs of [2], (c) and (f) are the outputs of [9, 10],
(d) and (g) are the results of our algorithm.



(a) Original image (b) Face map (c) ISC (d)PSC (e) FMO (f)Ours

Figure 4. Image retargeting comparison for 80% and 120% of original width:
(a) is the original images, (b) and (e) are the outputs of [2], (c) and (f) are the outputs of [9, 10],
(d) and (g) are the results of our algorithm.

3.1.2. Face Detector

Faces play an important role in retargeting as they usually draw the first attention of a human viewer. Just like [8], our face detection method is an implementation based on [14] and it is based on a machine learning algorithm by constructing classifiers corresponding to a small number of visual features based on AdaBoost. Moreover, this method combines classifiers with cascade so that the background is soon discarded, and thus the efficiency is improved. The outputs are squares where there are faces. The face map of image *girls* is showed in Figure 4(b) and a comparison for retargeting the original image to 50% width of four different retargeting methods is shown in Figure 5(c)-(f).

3.1.3. Centre Prior Map

From the research of Judd T. et al. [15] we can conclude that when humans watch pictures they naturally frame an object of interest near the centre of the image. For this reason, we include a feature named centre prior map, which indicates the distance of the centre to each pixel. In our algorithm, we define the distance from the centre to the corner is one. Then the distance of each pixel to the centre can normalized by the distance from the pixel to the centre to a value between 0 and 1.

3.1.4. The Optimized Cost Function

Instead of employing purely the gradient magnitudes, we use a cost function more related to human perception combining the gradient image with the saliency map, the centre prior map and the face map. Let I be an image and (x,y) be the coordinates of an arbitrary pixel, the cost function to measure the importance of each pixel in the original image is modified as:

$$e_{opt}(I(x,y)) = \alpha \cdot \sqrt{\left|\frac{\partial}{\partial x} I(x,y)\right|^2 + \left|\frac{\partial}{\partial y} I(x,y)\right|^2} + \beta \cdot s(x,y) + \gamma \cdot f(x,y) + \lambda \cdot c(x,y),$$

where $s(x,y)$, $f(x,y)$ and $c(x,y)$ correspond to the saliency map, the face map and the centre prior map, respectively. In the face map, if (x, y) is inside a detected rectangular facial region, we set $f(x, y) = 1$; Otherwise, we set $f(x, y) = 0$.

Weights α, β, γ and λ of the four attributes in the formula were empirically determined. In this paper, we set α=0.003, β=0.065, γ= 1 and λ= -1. Noting that the saliency map, the centre prior map and the face map are more accurate in the original image than in the diminished image, we calculate $s(x,y)$, $f(x,y)$ and $c(x,y)$ for each pixel in the original image only once, instead of re-calculating them whenever a seam is carved. As shown in Figure 2(d), the importance map obtained with our optimized cost function uniformly highlights the salient regions, and this cost function effectively keeps the seams to be carved out of any ROIs. As shown in Figure 4(a), our importance map prevents the seams from passing through the tent and the two men playing golf. Therefore, we can see that our method yields a much more

pleasant and accurate view when the image is downsized to 80% or enlarged to 120% width of the original in Figure (d) and (g).

3.2. Bi-Directional Seam Carving

In image browsing, studies have shown that the change of aspect ratio can have significant effects on human visual perception. If the image is to be changed in size while not in aspect ratio, the image can be retargeted to the target size by homogeneous scaling that perfectly preserves the content of the image. However, distortion may occur when the image is to be retargeted to a different aspect ratio. Moreover, when only a one-dimensional change is needed to retrieve the resized image, many approaches such as cropping and homogeneously scaling only operates on the image in one dimension. In the same case, SC algorithm only looks for seams in one direction, which causes global visual distortions.

Therefore, Daniel Domingues et al. [12] advanced an idea to reason in terms of aspect ratios and not any more in terms of dimensions. They compare the aspect ratio of the input image with the target image before searching for the least energy seams in both width and height dimensions. When the desired target ratio is smaller than the input ratio, vertical seams may be deleted for the image and horizontal seams may be added:

$$\frac{targetWidth}{targetHeight} = \frac{inputWidth - s_v}{inputHeight + s_h}$$

where s_v is the number of vertical seams to be deleted and s_h is the number of horizontal seams to be added. In the opposite case, vertical seams may be added to the image and horizontal seams may be deleted.

In the case of the seam duplication, their algorithm checks if the least energy seam crosses any important edge or a face. In the negative case, a "stream" instead of a seam may be added. However, this method cannot guarantee inserting seams with the current least energy. This may cause undesirable changes to the layout of the image content such as the image shown in Figure 5(d). In our proposed algorithm, we choose a vertical seam and a horizontal seam with the least energy costs and select the seam with less cost between them to operate at every step. We iterate this step until the desired aspect ratio is achieved. Then, we homogeneously scale the image to the desired dimensions by bi-cubic interpolation.

3.3. Adaptive Operator Switch

In order to avoid the introduction of noticeable artefacts to the image, our algorithm may switch from seam carving to scaling before the target aspect ratio is achieved. Depending on the characteristics of the concrete image, we propose to adaptively determine how many seams are deleted /inserted before switching from the seam carving to the scaling operator.

Specifically, the user sets a preferred threshold on the percentage, *p*%, of energy to be preserved in the resized image. Initially, seam operations are made on the input image. After each seam operation, we check if the total energy left in the image is below *p*%. If it is, we terminate the seam carving process and then homogeneously scale the seam-carved image to the target dimension. We empirically set *p*=85 which yields satisfactory results on our test images. As will be demonstrated in Section 4, our adaptive switching scheme can effectively prevent losing too much content of the image and protect the global visual quality of the image.

(a) Original (b) Scaling (c) ISC (d) STC (e) Ours

Figure 5. Comparison of four different retargeting methods for retargeting the original image to 60% width using Scaling, ISC [2], STC [12] and our algorithm.

(a) Original (b) ISC (c)SNS (d) PS (e) FMO (f) Ours

Figure 6. Comparison of five different retargeting methods for retargeting the original image to 50% width using ISC [2], SNS [4], PS [5], FMO [13] and our algorithm.

4. Experimental Results

The retargeted results obtained by our practical multi-operator retargeting method are compared against those of ISC [2] obtained using gradient map, as well as those of [5, 6, 9, 10, 11, 13, 14]. Figures in this paper show that outputs of our algorithm look more visually pleasing.

Figure1 demonstrates that our algorithm keeps the feature of the pigeons and the man better. When retargeting the original image to 50% width, cropping loses half content of the image. The ISC [2] deletes hundreds of vertical seams to reduce the width of the image to half of the original width, which leads to the terrible deformation of the man, while our optimization scheme chooses to delete less vertical seams and insert horizontal seams to achieve the target aspect ratio. Based on the results in Figure 4, comparing the results of image downsizing to 50% width of the original image, it is clear that our method represents an interesting alternative to ISC [2] and preserves the four girls and the global structure better than the other five approaches. As Figure 4(d) shows, the head of the girl in the distance is distorted and the cabinet above deformed severely using PSC algorithm [8]. In Figure 4(e) we can see that the girl on the right disappeared using FMO method [13]. Figure 5 shows that our algorithm can protect the content and the layout better than STC [12] and other two methods. Figure 6 shows that our algorithm preserves the arms of the bride and the shape of the couple better comparing with SNS algorithm [4], PS method [5] and other two methods.

5. Conclusions

We have presented a practical multi-operator image retargeting algorithm combining optimized seam carving and scaling for content-aware image retargeting. The proposed method uses an optimized cost function considering gradient map, frequency-tuned saliency map, face map and the centre prior map to guide the seam carving process. This new cost function can effectively detect the important regions in the image. The bi-directional searching operation prevents the retargeting process from operating excessively in a single direction of the image causing obvious visual deformation. In addition, the integration of seam carving and homogeneous scaling with a switching scheme affects the retargeting results. The optimized new cost function, the bi-directional searching operation and the adaptive switching scheme from seam carving to scaling work together to preserve the ROIs of the images. As a result, the proposed algorithm leads to more pleasant results when compared to several different image retargeting algorithms.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Grants No. 61070103 and No. U1035004), Shandong Provincial Natural Science Foundation, China (Grant No. ZR2011FZ004) and the Program for New Century Excellent Talents in University (NCET) in China.

References

1. Avidan, S., Shamir, A. (2007) Seam carving for content-aware image resizing, *ACM Transactions on graphics (TOG)*, 26(3), 10.
2. Rubinstein, M., Shamir, A., Avidan, S. (2008) Improved seam carving for video retargeting, *ACM Trans. on Graphics*, 27(3), 1–9.
3. Gal, R., Sorkine, O., Cohen-Or, D. (2006) Feature-aware texturing, //Proceedings of the 17th Eurographics conference on Rendering Techniques. Eurographics Association, 297-303.
4. Wang, Y.-S., Tai, C.-L., Sorkine, O., Lee, T.-Y. (2008) Optimized scale-and-stretch for image

resizing, *ACM Transactions on Graphics (TOG)*, 27(5), 118.
5. Liang, Y., Su, Z., Luo, X. (2012) Patchwise scaling method for content-aware image resizing, *Signal Processing*, Elsevier, 92(5), 1243–1257.
6. Itti, L., Koch, C. (2001) Computational modelling of visual attention, *Nature Reviews Neuroscience*, 2(3), 194–204.
7. Harel, J., Koch, C., Perona, P. (2007) Graph-based visual saliency, *Advances in Neural Information Processing Systems*, 545–552.
8. Hwang, D.S., Chien, S.Y. (2008) Content-aware image resizing using perceptual seam carving with human attention model, *IEEE International Conference on Multimedia and Expo*, 1029–1032.
9. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.S. (2009) Frequency-tuned salient region detection, IEEE International Conference on Computer Vision and Pattern Recognition, 1597–1604
10. Achanta, R., Susstrunk, S. (2009) Saliency detection for content-aware image resizing, ICIP, 16th IEEE International Conference on Image Processing, 1005–1008.
11. Rubinstein, M., Shamir, A., Avidan, S. (2009) Multi-operator media retargeting, *ACM Transactions on Graphics (TOG)*, 28(3), 1–11.
12. Domingues, D., Alahi, A., Vandergheynst, P. (2010) Stream carving: an adaptive seam carving algorithm, 17th IEEE International Conference on Image Processing (ICIP), 901–904.
13. Dong, W.M., Bao, G.B., Zhang, X.P., Paul, J.C. (2012) Fast Multi-Operator Image Resizing and Evaluation, *Journal of Computer Science and Technology*, 27(1), 121–134.
14. Viola, P., Jones, M. (2001) Rapid object detection using a boosted cascade of simple features, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 511–518.
15. Judd, T., Ehinger, K., Durand, F., Torralba, A. (2009) Learning to predict where humans look, IEEE 12th International Conference on Computer Vision, 2106-2113.

Received on the 20th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 251-259
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

WEB INFORMATION EXTRACTION BY USING IMPROVED HYBRID INTELLIGENT ALGORITHM AND HMM

Rong Li

Department of computer, Xinzhou Teachers' University, Xinzhou, Shanxi 034000, China
Phone: 13037007086. E-mail: lirong_1217@126.com

In order to further enhance the accuracy of Web information extraction, for the shortcomings of hidden Markov model (HMM) and its hybrid methods in parameter optimization, an improved Web extraction algorithm based on a self-adaptive hybrid intelligent algorithm and HMM is presented. This paper first builds an HMM, and then proposes a self-adaptive hybrid intelligent algorithm which organically integrates the updating mechanism of particle swarm optimization (PSO) and bacterial foraging algorithm (BFA). The hybrid intelligent algorithm is adopted for the approximate global optimal solution and then Baum-Welch algorithm (BW) is adopted for the local modification, which not only solves the BW dependency on initial values and the trapped local optimum problem, but also makes full use of the global search ability of the hybrid intelligent algorithm and local development ability of BW. Finally, the improved model uses Viterbi algorithm to decode. Compared with the existing HMM optimization methods, the comprehensive $F_{\beta=1}$ value is averagely increased by 7.5%, which shows that the improved algorithm can effectively enhance optimization performance and extraction accuracy.

Keywords: information extraction, hidden Markov model, hybrid intelligent optimization algorithm, particle swarm optimization, bacterial foraging algorithm

1. Introduction

With the development of Internet technology, Web resources have revealed the increase trend in massive and unstructured information. How to recognize those interesting data for users from unstructured or semi-structured Web information and turn them into a more structured and clearer semantic format, which is the technology problem of Web information extraction. A large number of experts and scholars applied statistical machine learning methods to this field. Typical statistical methods mainly include Hidden Markov Model (HMM) [1] and its hybrid methods [2, 3, 4, 5]. For instance, Lin et al. [2] proposed text information extraction method based on maximum entropy and HMM, which used the weighting sum of observation text feature to adjust HMM transition probability. Xiao et al. [3] employed genetic algorithm (GA) to optimize HMM parameters and obtained the extraction effect superior to traditional HMM, but the approach still reflected the precocious shortcoming of GA. Zou et al. [4] proposed Web information extraction based on simulated annealing (SA) and HMM, but the method did not consider HMM context features. Wang et al. [5] presented a web extraction algorithm using improved PSO and HMM, its improvement embodied in inertia weight and the mutation of part particles. By Analysis of these HMM literatures, there is still lots of room for improvement in parameter optimization and extraction performance.

Inspired by this, this paper proposes a self-adaptive hybrid intelligent optimization HMM algorithm for Web citation extraction. After constructing an HMM, a self-adaptive hybrid intelligent optimization algorithm based on BFA and PSO (or ABSO for short) is put forward. The ABSO algorithm realizes the dynamic nonlinear self-adaptive adjustment in parameters of PSO and BFA, respectively. After bacterial complete chemotaxis with variable step-length, the ABSO algorithm uses the optimization updating mechanism of PSO to continue to update bacteria's position, so as to help BFA jump out of local optimal. Then this paper maps the approximate global optimal solution, found out by ABSO, as the initial model of BW algorithm and adopts BW to continue to modify locally parameters. Finally, the improved model uses Viterbi algorithm to decode for the optimal state sequences. Experimental results indicate that the ABSO-HMM algorithm greatly improves the accuracy of Web information extraction, which proves the feasibility and effectiveness of the ABSO-HMM algorithm.

2. Web Information Extraction Based on HMM

An HMM may be viewed as a five-tuple (S, O, Π, A, B), where 1) S is a state set containing N

states, denoted as $S=\{S_1,S_2,...S_N\}$; 2) O is a symbol set including M output symbols, denoted as $O=\{O_1,O_2,...O_M\}$; 3) Π is the initial state probability matrix, denoted as $\pi=\{\pi_i\}, \pi_i=P(q_1=S_i), 1\le i\le N, 0\le\pi_i\le 1, \sum_{i=1}^{N}\pi_i=1$; 4) A is the state transition probability matrix; 5) B is symbol output probability matrix, denoted as $B=\{b_j(0_k)\}, b_j(0_k)=P(0_t=v_k|q_t=S_j), 0\le b_j(0_k)\le 1, \sum_{k=1}^{M}b_j(0_k)=1, 1\le j\le N, 1\le k\le M$.

When HMM is applied to information extraction, model observation layer is the text sequence to be observed, hidden layer is state sequence composed of state domains such as <Author>, <Title> and <Journal>. Extraction process may be described as follows: given the HMM model $\lambda=(\Pi,A,B)$ and the observe text sequence $O=\{O_1,O_2,...O_T\}$, initialize HMM parameters randomly, and use BW algorithm for HMM training so as to build HMM, finally adopt Viterbi algorithm to find out the state domain sequence $q^*=(q_1,q_2,...,q_T)$ with the maximum probability $P(q|O)$.

3. Self-Adaptive Hybrid Intelligent Optimization Algorithm

Particle swarm optimization (PSO) algorithm is a global optimization algorithm simulating the movement behaviour of bird swarm [6, 7, 8]. The solution of each problem may be seen as a particle in search pace. Particles update their velocity and position by tracking the individual optimal solution x_{pbest_i} and the global optimal solution x_{gbest}, the updating equations are as follows:

$$v_i(t)=\omega v_i(t-1)+c_1r_1(t)(x_{pbest_i}-x_i(t))+c_2r_2(t)(x_{gbest}-x_i(t))$$

$$x_i(t)=x_i(t-1)+v_i(t)$$

Among them, $x_i(t)$ and $v_i(t)$ represent the location and velocity of the i^{th} particle in the t^{th} generation, ω represents non-negative inertia weight, c_1 and c_2 are non-negative learning factors, and r_1,r_2 are random numbers in [0,1].

Bacterial foraging algorithm (BFA) is a kind of bionics random search algorithm based on bacterial foraging behaviour, including three procedures: chemotaxis, reproduction and elimination-dispersal [9, 10, 11]. The chemotaxis process is the core operation of BFA, the updating equation of bacterial tumble is as follows:

$$\theta^i(j+1,k,l)=\theta^i(j,k,l)+c(i)r(i)/\sqrt{r^T(i)r(i)}$$

Where $\theta^i(j,k,l)$ represents the position in which bacterium i is in the i^{th} generation chemotaxis, the k^{th} reproduction and the l^{th} elimination-dispersal operation, $c(i)$ refers to step-length and $\Delta(i)$ indicates tumble direction.

3.1. Self-adaptive Adjustment of Inertial Weight in PSO

The value of inertia weight ω has important influence on the PSO optimization search. in order to obtain better algorithm performance, traditional self-adaptive method makes ω linearly decrease with the increase of iterations, the equation is as follows [12]:

$$w=w_{\max}-(w_{\max}-w_{\min})\times\left(\frac{iter}{iter_{\max}}\right) \quad (4)$$

However, this kind of PSO algorithm cannot effectively reflect the complicated non-linear behaviour in the particle swarm's actual search process, so the convergence speed and convergence precision is still not ideal. In addition, the slope of which ω linearly decreases is still problem-dependency, there is no universal optimal change slope for all optimization problems.

This paper adopts a non-linear function to describe the dynamic change rule of ω in iteration process. The ω value of each iterative step is determined by the following exponential function formula:

$$w(iter) = w_{init} \times \exp\left(-\left(\frac{iter_{now}}{iter_{\max}} \right)^n \right) \tag{5}$$

Where n is a control power exponent of non-linear change rule, particularly, when $n = 2$, (5) is often referred to as probability curve function. Figure 1 shows the ω iteration change curve with different n value.

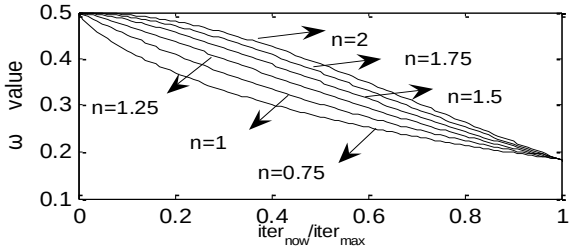


Figure 1. The ω iteration change curve with different n value

As shown in Figure 1, the greater the n value is, the longer the global search duration of particle swarm is, while the smaller the n value is, the longer the local search duration of particle swarm is. By using d-dimension spherical function $f(x) = \sum_{i=1}^{d} x_i^2 \quad (d = 6,\ x_i \in [-5.12, 5.12])$ to validate parameter values, the result shows that when ω_{init} value is set in [0.2,0.5], and n value is set in [0.5,2], the algorithm has excellent performance.

3.2. Self-adaptive Adjustment of Chemotaxis Step-length in BFA

As a very important parameter in BFA chemotactic link, the appropriate selection of chemotaxis step-length can determine population diversity, accelerate convergence and enhance convergence precision. Original BFA adopts fixed step-length to solve problem, which is not good for the convergence of algorithm [9]. A large number of experiments show that in early phase the algorithm maintains a larger initial step-length for a wide range of coarse search, and in the later stage maintains a smaller step-length for a small range of precision search, which can not only speed up convergence, but also improve convergence precision to a certain extent. Hence, this paper employs the following dynamically adjusted step-length equation:

$$C(k,l) = C_{init} / m^{k+l-1} \tag{6}$$

Where $C(k,l)$ is the chemotaxis step-length of the k^{th} copy and the l^{th} elimination-dispersal, C_{init} is the initial chemotaxis step-length and m is the parameter of controlling step-length's reduction gradient. $C(k,l)$ is the function which declines with the increase of elimination-dispersal events. So when $k+l$ is smaller, $C(k,l)$ is larger, thereby avoiding consuming too much time in the local search range, and when $k+l$ gradually increases, $C(k,l)$ decreases, accordingly the bacterial local search ability near the global optimal point increases, and this guarantees that the algorithm eventually approaches the global optimal point.

3.3. Self-adaptive Hybrid Optimization Algorithm

The general idea of self-adaptive hybrid optimization algorithm based on PSO and BFA (ABSO) is under the self-adaptive adjustment strategy for learning factors and inertial weight in PSO and BFA chemokines step-length, after finishing BFA chemotaxis with variable step-size, the new algorithm continues to adjust the bacterial position with the self-adaptive PSO iteration updating formula. This method can make up for the blindness of BFA random tumble, thereby helping BFA jump out of local optimal and balance the local search and global development optimization ability. The specific procedure of ABSO is embodied in Figure 5 in section 4.1.

4. Web Information Extraction Based on ABSO-HMM

4.1. Parameters Training Based on ABSO-BW

1) Coding scheme

When ABSO is applied to train HMM for Web information extraction, a particle (or bacterium) corresponds to an HMM. We define that elements of particle's location vector *X*, in turn, are the linear array of four types of parameters of the improved HMM model $\lambda = \{\pi_i\}, \{a_{ij}\}, \{b_{jk}\}$, as can be expressed in the matrix in Figure 2. Mapping all the parameters in the matrix to ABSO search space can compose a dimensionality, which contains N probabilities π_i, N^2 probabilities a_{ij}, $N*M$ probabilities b_{jk}, totals $N*(1+N+M)$ parameters. Therefore, each particle is a real number coding string with $N*(1+N+M)$-dimension.

π_1	π_2	...	...	π_N
a_{11}	a_{12}	...	...	a_{1N}
...				
a_{N1}	a_{N2}	...	...	a_{NN}
b_{11}	b_{12}	...		b_{1M}
...				
b_{N1}	$b_{N2}\ B$	...		b_{NM}

Figure 2. Coding scheme of a particle

2) Fitness function

In this hybrid algorithm, we use the logarithmic probability mean of observation sequences to measure quality of model, the purpose is to make all sequences have the greatest probability values, so that the model can better explain observation sequence, the fitness function is as follows:

$$f(\lambda_i) = \frac{1}{L}\sum_{k=1}^{L} \ln\left(p\left(o^k|\lambda_i\right)\right) \tag{7}$$

Among them, λ_i is the composite HMM corresponding to the i^{th} particle, L is observation sequence number, O^k, one of the observation sequences, whose length is T and whose structure form is a tuple $(O_{11}, O_{12}, ..., O_{1T})$. The probability value $p(O|\lambda_i)$ can be calculated by the forward-backward algorithm of HMM, taking the logarithm of probability is in order to avoid the underflow of probability multiplication.

3) Constrained normalization

HMM training is a constrained optimization problem, as shown in chapter 2, parameter values of model λ are all probabilities, whose constraint requirements are non-negative, [0,1], and sum of all probabilities should be one. While ABSO is an optimization algorithm without constraints, so when ABSO is used to optimize HMM, those particles which do not meet the above constraints in every generation need to be normalized. We illustrate the normalization process with the example of a row of state transition matrix A. Correction is divided into two steps: first all the probabilities of negative value in the matrix row are set to zero, so as to meet the requirement of non-negative probability; then all the probabilities a_{ij} are replaced by $a_{ij} = a_{ij} / \sum_{j=1}^{N} a_{ij}$ to meet the requirement that the sum of all probabilities should be one. The parameters of Vector π and matrix B are normalized in the same way.

4) HMM training algorithm based on ABSO-BW

Wang and Chen et al. [5] first adopted BW to estimate HMM initial parameters, and then took each estimated sub-optimal HMM as a particle of PSO. Different from this method, this paper first adopts ABSO for an intermediate result close to the global optimum, and uses the result as the initial model of BW, and then modifies BW parameters for the final result θ_{gbest}.

The ABSO-BW specific training procedure is as follows:

*Step 1: Init*ialize parameters S, N_C, N_{re}, N_{ed}, P_{ed}, N_S, C_{init}, m, a, ω_{init}, n, R_1, R_2 and BW termination iteration threshold ε.

Step 2: Randomly initialize bacterial position $\theta^i(1,1,1),\ i = 1,2...S$.

Step 3: Calculate the initial fitness of each bacterium: $J^i(1,1,1) = Fun\left(\theta^i(1,1,1)\right)$, and the initialize individual optimal value θ_{pbest_i} of each bacterium and the global optimal value θ_{gbest}.

Step 4: For $(l = 1; l < N_{ed}; l++)$// l is elimination-dispersal generation

Step 5: For $(k = 1; k < N_{re}; k++)$// k is reproduction generation

Step 6: For $(j = 1; j < Nc; j++)$// j *is chemotaxis generation*

Step 7: For $(i = 1; i < S; i++)$

Step 7.1: Compute the bacterial fitness, $J^i(j,k,l) = Fun\left(\theta^i(j,k,l)\right)$.

Step 7.2: Save the current optimal individual to J_{last}, $J_{last} = J^i(j,k,l)$;

Step 7.3: Tumble randomly with self-adaptive step-size according to the ABSO equation (3) and (6), *and update bacterial location* $\theta^i(j+j,k,l)$.

Step 7.4:Constrained normalization for HMM parameters; // Constraint modification for bacterial location

Step 7.5: Compute the bacterial fitness $J^i(j+j,k,l)$ with $\theta^i(j+j,k,l)$.

Step 7.6: While $(m1 < Ns)$

1) If $J^i(j+j,k,l) < J_{last}$, *let* $J_{last} = J^i(j+j,k,l)$ *and continue to swim forward with variable step-size in this same direction,* normalize bacterial location and then calculate the fitness-updating value.

2) *Else terminate the swim, and save the local optimal value of bacterium i to Jlast..*

3) End if

Step 7.7: End while

Step 8: End for //End for the traverse to bacterial individual

Step 9: Calculate the new velocity and position of each bacterium according to the self-adaptive PSO equ*ation (1),(2) and (5).*

Step 10: Constrained normalization for HMM parameters;// Constraint modification for bacterial location

Step 11: Calculate the updated fitness value of each bacterium.

Step 12: Update the individual optimal position θ_{pbest_i} and the global optimal position θ_{gbest}

Step 13: End for // End for chemotaxis

Step 14: Calculate the health degree of each bacterium, $J_{health}^{i} = \sum_{j=1}^{N_c+1} J^{i}(j,k,l)$, and sort bacteria in order of ascending health degrees.

Step 15: Split and elimination: eliminate the latter half bacteria with poorer health degree, and split the same amount of bacterial colony from the first half bacteria with better health degree.

Step 16: End for // End for reproduction

Step 17: Elimination and dispersal: For the given dispersal probability P_{ed} , if a bacterium individual meets the dispersal condition, re-generate a new individual according to random algorithm.

Step 18: End for // End for elimination-dispersal

Step 19: BW iteration, map θ_{gbest} to initial parameters of BW $\lambda_0^{'} = (\pi_0^{'}, A_0^{'}, B_0^{'})$, and perform BW algorithm for the HMM corresponding to the optimal particle.

Step 20: If the model growth rate of two successful iterations is less than threshold, namely $\frac{f(\lambda_{iter+1}) - f(\lambda_{iter})}{f(\lambda_{iter}) - f(\lambda_{iter-1})} < \varepsilon$, then the algorithm terminates.

The flowchart of the ABSO-BW algorithm is illustrated in Figure 3.

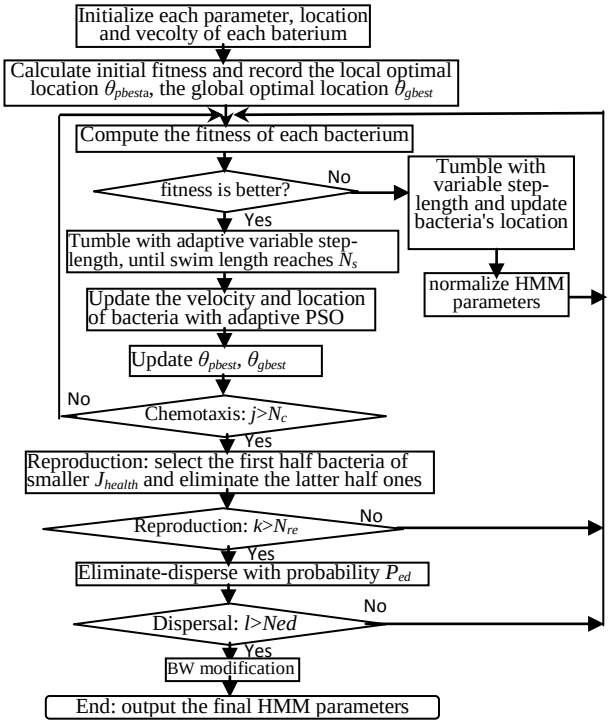


Figure 3. Flowchart of the ABSO-BW parameter training

4.2. Web Information Extraction Based on ABSO-HMM

This paper builds an improved ABSO-HMM model. By Using citations in Web research papers as treatment objects, the model extracts state domains in reference such as <Author>, Book>, <Title>, <Journal>. The extraction process of the improved model is as follows:

1) Information preprocessing

This article first uses the peweb tool from www.jaist.ac.jp/~hieuxuan/softwares/peweb website for Web citation record extraction; and then utilizes delimiter such as punctuation and text features for information chunking pretreatment. Among them, the deterministic text features are characteristic word "Journal" corresponding to journal state domain <Journal>, characteristic words "Conference", "Proceedings" and "Symposium" corresponding to conference proceedings state domain <Conference> and characteristic words "Press", "Publishers" corresponding to press state domain < Press > and so on.

2) Model training.

After initialling HMM parameters randomly, the ABSO algorithm is adopted to optimize HMM parameters and then BW algorithm is used to modify HMM parameters locally, which builds an improved HMM.

3) Information extraction

Viterbi algorithm is employed to obtain the optimal state sequence of test sample.
The specific extraction process is shown in Figure 4.

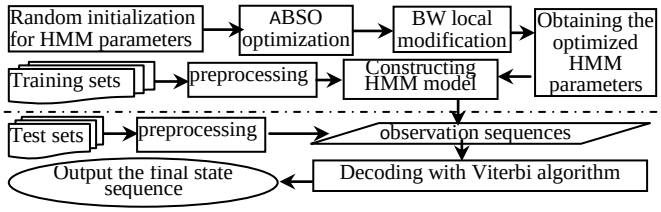


Figure 4. Web extraction process

5. Experimental Results and Analysis

2500 unlabelled research paper citations are used as experimental samples, a part of which are 500 citation data sets (http://www.cs.cmu.edu/~kseymore/ie.html) from the United States Carnegie Mellon University (CMU), the other part of which are 2000 literature records from 387 research papers extracted randomly from online journal database. We select 1800 citation records as training sets, totalling 44,283 words, the other 700 citation records as open test sets, totalling 15,073 words.

In HMM optimization training process, the hybrid training parameters are as follows: the population size $S=30$, the maximum dispersal generation $N_{ed}=2$, the maximum chemotaxis generation $N_c=200$, the maximum reproduction generation $N_{re}=5$, the biggest swim steps $N_s=4$, the dispersal probability $P_{ed}=0.25$, the ABSO initial step-length $C_{init}0.15$, the control parameter of variable step-length $m=2$, the control parameter of self-adaptive learning factor $a=7$, the initial inertia weight $\omega_{init}=0.5$, the control parameter of self-adaptive inertia weight $n=1.25$, BW iteration threshold $\varepsilon=1e-5$, the HMM state number $N=11$, the initial value of λ is selected randomly. We use GA-HMM [3], SA-HMM [4] and ABSO-BW algorithm in this paper to train HMM and analysis the convergence of three algorithms, their sampling error formula is defined in (8).

$$\varepsilon\left(p\left(O|\lambda\right)\right)=\sqrt{\frac{1}{N\times M}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\hat{p}\left(O|\lambda\right)-p\left(O|\lambda\right)\right)^2} \tag{8}$$

Where $\left(\hat{p}\left(O|\lambda\right)\right)$ is the probability of training model, $\left(p\left(O|\lambda\right)\right)$ is the probability of sample generation model. The comparison results are illustrated in Figure 5.

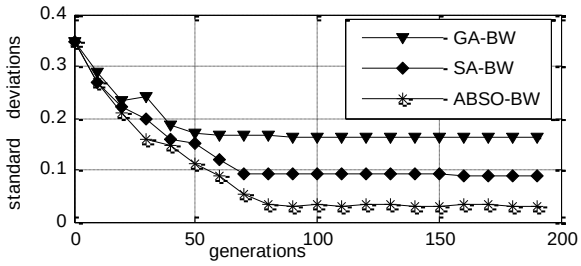


Figure 5. Standard error comparison of three algorithms

As shown in Figure 5, the standard error of GA-HMM and SA-HMM begin to converge close to 0.16 and 0.09, respectively, while that of ABSO-BW begins to converge only close to 0.03, its standard error reduces respectively by about 13% and 6% compared with the previous two algorithms. It proves that the improved algorithm has stronger search ability, convergence speed and very low error, can more accurately train HMM model, so as to improve system quality. At the same time, Figure 5 shows the improved algorithm has better stability.

The information extraction results of three optimization algorithms are shown in Table 1.

TABLE 1. Extraction precision and recall comparison of three algorithms

State	GA-HMM		SA-HMM		ABSO-HMM	
	Precision	Recall	Precision	Recall	Precision	Recall
Author	0.91367	0.91413	0.91945	0.92219	0.9698	0.95342
Title	0.75854	0.80937	0.83156	0.84101	0.9074	0.92167
Book	0.80480	0.79412	0.83167	0.80467	0.84823	0.89582
Journal	0.79430	0.81583	0.81042	0.86312	0.90267	0.92551
Conference	0.76784	0.78756	0.82357	0.85382	0.85991	0.92491
Press	0.85668	0.87337	0.85945	0.88057	0.91108	0.94702
City	0.85493	0.85568	0.85462	0.85998	0.87672	0.88523
Volume	0.82558	0.85801	0.82962	0.81034	0.91945	0.93792
No.	0.89330	0.91080	0.90357	0.88382	0.91516	0.92359
Year	0.86318	0.90436	0.88164	0.89314	0.92871	0.95909
Pages	0.87008	0.85989	0.86245	0.87158	0.91455	0.94965
Average	0.83663	0.85301	0.85527	0.86220	0.90488	0.929439

TABLE.2. Average $F_{\beta=1}$ comparison of three algorithms

	GA-HMM	SA-HMM	ABSO-HMM
$F_{\beta=1}$	0.84474	0.858725	0.916995

As shown in Table 1, 2, the extraction precision and recall of ABSO-HMM are all much higher than the previous two algorithms. Measuring from the average comprehensive index $F_{\beta=1}$ value, ABSO-HMM increases respectively by 7.2% and 5.8% than the previous two. The precision and recall of states <Journal>, <press> and <conference> increase significantly, mainly because the combination of deterministic feature information with hybrid optimization model enhances extraction performance. The interference of state fields such as <book> versus <title>, <Journal> versus <conference> is strongest, but due to text features and bi-order decoding ,the precision rates of states <title>, <Journal> and the recall rates of states <book>, <conference> are greatly enhanced. While the interference of <Author> is smaller, its accuracy rate reached 0.96980. Synthesizing the above data, it can be seen that the validity of ABSO-HMM algorithm.

6. Conclusions

In view of the defects of the traditional HMM hybrid method for Web information extraction, this paper proposes a self-adaptive hybrid intelligence training algorithm based on ABSO-BW for citation extraction. The ABSO-BW training algorithm adjusts adaptively parameters of PSO and BFA, combines their optimization mechanism and controls the global search direction and speed, and then uses BW to revise parameters locally, which considers the influence of the information contained in training sequence and social information on HMM global optimization. Therefore, the new hybrid algorithm

enhances the probability of model global optimization, effectively overcomes premature, quickly converges with extremely low error and has stronger optimization ability. Experimental results show that compared with the traditional SA-HMM and GA-HMM optimization method, ABSO-HMM reflects the strong advantage in optimal performance and extraction accuracy, which proves the effectiveness of the improved algorithm. Future researches can focus on: use new hybrid intelligent algorithms to development HMM optimization method with better performance and lower complexity, and then apply it to the actual intelligent information processing system.

Acknowledgments

We would like to thank to the reviewers for their helpful comments. This work was financially supported by the Natural Science Foundation of China (#1072166), the Higher School Science and Technology Development Project in Shanxi Province of China (#2013147).

References

1. Marcinczuk, M., Piasecki, M. (2010) Study on named entity recognition for polish based on hidden Markov models,Proceedings of Text, *Speech and Dialogue-13th International Conference* (TSD 2010), Springer Berlin Heidelberg, 142-149.
2. Lin, Y.P., Liu, Y.Z, Zhou, S.X., Chen, Z.P., Cai, L.J. (2005) Using hidden Msrkov model for text information extraction based on maximum entropy, *Acta Electronica sinica*, 33(2), 236-240.
3. Xiao, J.Y., Zou, L.M., Li, C.Q. (2008) Hybrid genetic algorithm and hidden Markov model for web information extraction, *Computer Engineering and Applications*, 44(18), 132-135.
4. Zou, L.M., Gong, X.J., Xiao, F., Ma, S.P. (2011) Web information extraction based on simulated annealing algorithm and hidden Markov model, *Journal of University of South China*, 25(1), 70-74.
5. Wang, C., Duan, D.Q., Wang, X.D. (2010) An improved PSO and HMM algorithm for web information extraction, *Journal of Henan Normal University (Natural Science)*, 38(5), 65-68.
6. Yang, S.J., Wang, S.W., Tao, J., Liu, X. (2012) Multi-objective optimization method based on hybrid swarm intelligence algorithm, *Computer simulation*, 29(6), 218-222.
7. Motiian, S., Soltanian-Zadeh, H. (2011, September). Improved particle swarm optimization and applications to hidden markov model and ackley function. *In Computational Intelligence for Measurement Systems and Applications (CIMSA), 2011 IEEE International Conference on IEEE*, Ottawa, Canada, 1-4.
8. Najkar, N., Razzazi, F., Sameti, H. (2010) A novel approach to HMM-based speech recognition systems using particle swarm optimization, *Mathematical and Computer Modelling*, 52(11), 1910-1920.
9. Bakwad, K.M., Pattnaik, S.S., Sohi, B.S. (2009) Hybrid bacterial foraging with parameter free PSO, *NaBIC 2009: World Congress on Nature & Biologically Inspired Computing*, Piscataway, NJ: IEEE Press, 1077 - 1081.
10. Anant, B., Chauhan, V.S., Jayabarathi, T. (2011) Application of PSO, Artificial Bee Colony and Bacterial Foraging Optimization algorithms to economic load dispatch: An analysis, *International Journal of Computer Science Issues*, 8(4), 467-470.
11. Sathya, P.D., Kayalvizhi, R. (2011) Optimal multilevel thresholding using bacterial foraging algorithm, *Expert Systems with Applications*, 38(12), 15549-15564.
12. Shi, Y., Eberhart, R.C. (2001) Fuzzy self-adaptive particle swarm optimization, *In: Proceedings of the IEEE Congress on Evolutionary Computation*, Piscataway, NJ: IEEE Service Centre, 101-106.

Received on the 20th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 260-269
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

DYNAMIC ANALYSIS ON TREND OF STOCK PRICE BASED ON ELASTICITY COEFFICIENT MODEL

Zhang Jilin

Fujian University of Technology, Fuzhou, 350108, China
E-mail: jilinzhang1976@126.com

In this paper, the quartic non-linear regression model of stock price and trading volume with respect to time has been set up. Based on this, we built a dynamic elasticity coefficient model to forecast the stock price and the stock market trend. With the detailed study methods given by these two models, we made an empirical research on daily data of three stock samples namely China Railway Construction Company (601186), Xiamen Tungsten Industry (600549) and Beijing Capital Development Holding (Group) Co. Ltd. (600376), to check the efficiency of elasticity predictive models.

Keywords: Stock Price; trading volume; Non-linear Regression; Elasticity coefficient.

1. Introduction

Now days with the blooming development of market economy, financial system speeds up and acts as the major player of modern market economy. As a result, numerous factors affecting the stock prices arise from time to time in this complex stock market. Therefore, it is hard to explain the trend in stock price. The essential factors for technical analysis are stock price, trading volume, time and place, and the relationship among different factors in market; the most basic being the relationship between stock prices and trading volume. In general trading volume is the best indicator of stock price in stock market. There are some questions about the causality between them, which are finally clarified by empirical evidence that the absolute value of trading volume and stock price is positively correlated. Trading volume directly reflects the market status. Two platitudes have constantly been quoted in the Wall Street, one is that trading volume drives a high stock price, and the other is that one should sell a bull market short during downfall. The former one tells that the trading volume and stock price are positively corrected and trading volume is the leading indicator of stock price. While the latter one means the change of stock price will influence trading volume. In past 22 years of domestic stock market, it has been observed that roughly every peak of stock index was also the peak of trading volume and every trough of stock index was the valley of trading volume. This indicates the possibility of direct relationship between stock price and trading volume.

Since past it has been believe that for a stable stock market, there should be good coordination between stock index and trading volume, and any divergence may probably lead to a market inversion. Usually a change in trading volume precedes the change in stock prices. Numerous researches on relations between stock price (return rate) and its fluctuation have achieved great results. However, studies or files on trading volume, by contrast, seem to be less. There are several reasons to investigate into the relationship between trading volume and stock prices. Some of which are:

- It gives a closer look into the micro-construction of financial market
- Studying on relations between price and volume is of great meaning to event study
- Structures of relationship between price and volume decide the empirical distribution of speculative price
- Basic presentation of market behaviour is price and volume, while the focal spot of technical analysis is exactly market behaviour itself.
- Results of study on price and volume may prove out to be very helpful in study of financial derivatives, for example, futures.

Researchers interested in studying the relationship between stock prices and trading volume are mainly fascinated by its implication in theory and practice. It has huge application in real life situations. In 2000, Mr. Wang Peixun fitted the quartic nonlinear regression equation in sections by least-square method. Analysing from a technical point of view, he revealed the inherent laws of interactions between stock price and trading volume through elasticity analysis of regression equation. Ms. Huang Yuanyuan

developed a quartic nonlinear regression equation based on elasticity coefficient in 2006. She analysed the relations between stock price and trading volume as well as elasticity numerical sequence of northeast-expressway from Jan. 9th 2006 to Jan. 13th 2006 (600003), however she did not make any predictions on it with the help of this model. Both of the above studies are static analysis for stock price. Analysis has been done by developing a quartic non-linear regression model of trading volume and stock price with respect to time. However, with the developments of global economy and domestic stock market, more and more factors affecting the stock market arise from time to time. Therefore, it seems too simple to build a nonlinear regression of trading volume and price about time only by quartic model.

In this paper, we will see a quartic nonlinear regression model of volume and price about time established by static software, on which elasticity model is built to predict stock price or market. 2012 daily data from three enterprises, China Railway Construction Company (601186), Xiamen Tungsten Industry (600549) and Beijing Capital Development Holding (Group) Co. Ltd. (600376), have been taken as empirical samples, the model could be effective because the predictions it made are more or less similar to stock trend observed in practice.

2. Nonlinear Regression Model of Correlations between Stock Price and Trading Volume

Technically speaking, for traditional models, trading volume is simply for reference, what researchers really care about, is time sequence of stock price. While the fundamental principle for stock price and trading volume is that, the former one results in the latter one. In other words, we can say that trading volume is the real power that drives the change of stock price. Researchers have framed many principles stating relationship of stock price and trading volume, considering the fact that stock prices can be influenced by big organization but trading volume still remains same.

In this paper, an attempt has been made to draw a curve of stock price and trading volume about time *t*, and to build an elasticity coefficient model of correlations between stock price and trading volume to predict the stock price trend by nonlinear regression model.

Elasticity coefficient is identified as the ratio of fractional changes of the two indexes related to each other in time. It can be used to check the effects of percentage change in one economic index over another, which may reflect the correlations between trading volume and stock price. We assume that daily trading volume of a stock during K days is $v_1, v_2, ..., v_k$ and daily closing price is $u_1, u_2, ..., u_k$. According to wave theory, we can see from trading time that the stock price and trading volume are wavy distributes in spots. This is the reason why time is taken as an argument to fit the nonlinear regression curve of daily closing price and trading volume.

A part of stock price pattern similar to non-linear curve has been chosen from graph of daily K line, and it is assumed that the daily trading volume on the nth transaction day is $v_1, v_2, ..., v_n$. With the help of polynomial regression, a non-linear regression curve of highest significance level has been fitted. The quartic polynomial is taken as an example, and the nonlinear model of daily trading volume *v* is shown below:

$$v = h(t) = b_4 t^4 + b_3 t^3 + b_2 t^2 + b_1 t + b_0 \tag{1}$$

It can be transformed into multi-linear model:

$$v = h(t) = b_4 x_4 + b_3 x_3 + b_2 x_2 + b_1 x_1 + b_0 \tag{2}$$

b_4, b_3, b_2, b_1, b_0 are undetermined coefficients and t is time variable $t = 1,2,...,n$. For $t = n$, the corresponding v will be v_n. Matlab can be used to get the undetermined coefficient of b_4, b_3, b_2, b_1, b_0 which will be denoted by $b_4', b_3', b_2', b_1', b_0'$.

The nonlinear regression model of daily trading volume about time t is:

$$V = h(t) = b_4' t^4 + b_3' t^3 + b_2' t^2 + b_1' t + b_0' \tag{3}$$

Similarly, the stock price in the above assumed n days is written as $u_1, u_2, ..., u_n$ respectively, in the same way, a non-linear regression curve of stock price about time can be built, and the model is as follows:

$$u = s(t) = a_4 t^4 + a_3 t^3 + a_2 t^2 + a_1 t + a_0 \tag{4}$$

And it can be transformed into multi-linear model:

$$u = s(t) = a_4 x_4 + a_3 x_3 + a_2 x_2 + a_1 x_1 + a_0$$

a_4, a_3, a_2, a_1, a_0 are undetermined coefficients, while t is time variable $t = 1,2,...,n$. When $t = n$, u will be u_n. With matlab, we will get the undetermined coefficient of a_4, a_3, a_2, a_1, a_0, which will be denoted by $a_4^{'}, a_3^{'}, a_2^{'}, a_1^{'}, a_0^{'}$ counted by multi-linear regression, and the nonlinear regression model of stock price about time t is:

$$U = s(t) = a_4^{'} t^4 + a_3^{'} t^3 + a_2^{'} t^2 + a_1^{'} t + a_0^{'} \tag{6}$$

3. Elasticity Coefficient Prediction Model

We take an assumption that the trading volume and stock price at t is v and u. Consider a small increments of Δt, Δv and Δu at time t, volume v and price u respectively. The fractional change in trading volume and stock prices are $\Delta v / v$ and $\Delta u / u$ respectively. As a common principle in economics elasticity coefficient refers to the fractional (percentage) change in one index with respect to fractional (percentage) change in the other. Or in other words growth rate of the two indexes is connected through elasticity. Elasticity coefficient here may reflect the correlations of trading volume and stock price. The elasticity coefficients of trading volume and stock price can be treated as the limit ratio of price change $\Delta u / u$ and volume change $\Delta v / v$ as Δt approaches 0:

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{\Delta u}{u} / \frac{\Delta v}{v} \tag{7}$$

When $h'(t) \neq 0$

$$\begin{aligned} \alpha(t) &= \lim_{\Delta t \to 0} \frac{\Delta u}{u \cdot \Delta t} / \frac{\Delta v}{v \cdot \Delta t} = \lim_{\Delta t \to 0} \frac{\Delta u}{\Delta t} \cdot \frac{1}{u} / \frac{\Delta v}{\Delta t} \cdot \frac{1}{v} = \frac{v}{u} \cdot \frac{u'}{v'} = \frac{h(t)}{s(t)} \cdot \frac{s'(t)}{h'(t)} \\ &= \frac{b_4^{'} t^4 + b_3^{'} t^3 + b_2^{'} t^2 + b_1^{'} t + b_0^{'}}{a_4^{'} t^4 + a_3^{'} t^3 + a_2^{'} t^2 + a_1^{'} t + a_0^{'}} \cdot \frac{4a_4^{'} t^3 + 3a_3^{'} t^2 + 2a_2^{'} t + a_1^{'}}{4b_4^{'} t^3 + 3b_3^{'} t^2 + 2b_2^{'} t + b_1^{'}} \end{aligned} \tag{8}$$

In above equations:

$$V' = \frac{\Delta V}{\Delta t} \tag{9}$$

$$U' = \frac{\Delta U}{\Delta t} \tag{10}$$

With elasticity coefficient $\alpha(t)$ we listed four possible situations to analysis the stock trend:

Situation 1): $\alpha(t) > 1$

When $h'(t) > 0$ that is trading volume increases and $s'(t) > 0$ that is stock prices increases then in this situation the ratio of fractional changes in stock prices and trading volume is greater than one. Which implies that percentage increase in stock prices (u'/u) is more than percentage increase in trading volume (v'/v). That is in such situation a larger increase in stock prices is preceded by comparatively smaller increase in trading volume and this marks the beginning of bull market that is increasing trend of stock prices. Investors eagerly wait for such situations, as this is the time when they can earn huge profits. This acts as a signal of good future of stock market.

When $h'(t) < 0$ that is trading volume decreases and $s'(t) < 0$ that is stock prices decreases then in this situation the ratio of fractional changes in stock prices and trading volume is greater than one. Which implies that percentage decrease in stock prices (u'/u) is more than percentage decrease in trading volume (v'/v). In such situation a larger decrease in stock prices is preceded by comparatively smaller decrease in trading volume. This may just be the beginning of the downfall in the stock prices and this decline may continue for long time. In addition, if this has been continued for long then it may stabilize in near future. Investors may earn profits through short sell in such situations.

Situation 2): $0 < \alpha(t) < 1$

When $h'(t) > 0$ and $s'(t) > 0$ that is stock prices and trading volume both increases then in such situation the fractional change in stock prices is less than that of trading volume. Stock prices rises with increase in trading volume. This is a signal for investors to buy more stocks.

When $h'(t) < 0$ and $s'(t) < 0$ that is stock prices and trading volume both decreases but the fractional change in stock price is less. This may just be the beginning of the downfall in the stock prices and this decline may continue for long time. Moreover, if this has been continued for long then it may stabilize in near future. Investors may earn profits through short sell in such situations. Investors like to earn through short sell in such situations.

Situation 3): $-1 < \alpha(t) < 0$

When $h'(t) > 0$ and $s'(t) < 0$ in this situation stock price and trading volume moves in opposite direction but the magnitude of fractional change in stock price is less than that of trading volume. After continuously falling stock prices, the trading volume may move very sharply the opposite way. This can be regarded as the bottom most point of the stock price fall. If the fall in prices had just begun then it may be expected to continue declining in future.

When $h'(t) < 0$ and $s'(t) > 0$ then also it's the same case as above and the magnitude of fractional change in stock prices is more than that of fractional change in trading volume. This phenomenon of trading volume decline and stock price increase mostly occurs at the end of stock price up going period and in some exceptional cases occurs during the rebounding process in middle of the decline. However, the research results on it will be different for such market situation.

Trading volume shrinkage acts as a check on market and prevents it from increasing stock price indefinitely. In order to do so many chips in circulation may be locked by it. In such situations, stocks are intentionally sold at high rates if the trading volume seems to be amplified and is expected to follow an uptrend. In this case, we see that trading volume and stock price deviate from each other.

Sometimes volume rebounds a rare incidence. It is observed when there is a rise in prices after consistently falling prices. Generally there is rapid decline in stock prices because of this some of the securities remain unsold. The big players of the market know that in such situations mostly investors will remain out of stocks. Hence, the big players spend some money to raise the stock price to give investors an illusion of rising prices. Therefore, with an expectation of rising prices, in the future investors buy the stocks and big players are able to save themselves. The big players create an illusion to achieve their final target of valuably selling their securities at hand.

Therefore, investors should have a clear idea about the market situation of trading volume shrinkage and price rise. As there are high chances of falling into traps of big players of the market. The best solution is to hold up the money and see the best choices.

Situation 4): $\alpha(t) < -1$

In this situation $h'(t) > 0$ and $s'(t) < 0$ that is prices and trading volume move oppositely and fractional change in prices is more than fractional change in trading volume. This low price high volume phenomenon may be an indication of a rebound force building up in the market. This may probably be the floor of the stock and prices may rise in the near future. Stock prices may burst out in the near future and investors should pay attention to it to find out the right time to invest.

The other case is that $h'(t) < 0$ and $s'(t) > 0$. It may be just one day blooming if the upward movement has just begun.

4. Empirical Analysis

In this study I have transformed the non-linear multiple regression equation into linear multiple regression equation. First the coefficients $b_4^{'}, b_3^{'}, b_2^{'}, b_1^{'}, b_0^{'}$ and $a_4^{'}, a_3^{'}, a_2^{'}, a_1^{'}, a_0^{'}$ of multiple regression equation are calculated u using matlab and then elasticity coefficient $\alpha(t)$ is calculated by elasticity coefficient equation.

In this paper, the data of trading volume and the real and fitted values of closing price for the three researched objects namely Shanghai A share of China Railway Construction Company (601186), hereinafter referred as CRCC, from 7th-18th May, 2012, Shanghai A share of Xiamen Tungsten Industry (600549), hereinafter referred as XTC 2,from 19th-30th Mar. 2012, and Shanghai A share of Beijing Capital Development Holding (Group) Co. Ltd.(600376), hereinafter referred as BCDC, from 12th-25th Oct. 2011, are listed. A prediction for the movements of stock price has been made according to elasticity coefficient $\alpha(t)$. The following results were obtained:

4.1. Stock price analysis for CRCC (601186)

The data of trading volume and the real value of stock price of CRCC (601186) from 7th-18th May 2012 is shown as Table 1:

TABLE 1. The data of trading volume and the real value of stock price of CRCC (601186)

Stock CRCC (Unit: Ten Thousand Hand, RMB Yuan)					
Project Name	Trading Volume *v*	Closing Price *u*	Project Name	Trading Volume *v*	Closing Price *u*
T=1	27.29	4.42	T=6	13.17	4.29
T=2	28.07	4.44	T=7	14.78	4.26
T=3	24.83	4.34	T=8	14.82	4.17
T=4	20.93	4.3	T=9	11.84	4.25
T=5	16.07	4.27	T=10	15.86	4.22

Multiple regression coefficients $b_4^{'}, b_3^{'}, b_2^{'}, b_1^{'}, b_0^{'}$ calculated by matlab are 22.1317, 9.0027, -4.0799, 0.5176, -0.020 correspondently, and $a_4^{'}, a_3^{'}, a_2^{'}, a_1^{'}, a_0^{'}$ are 4.4492, -0.0032, -0.0143, 0.0021, -0.0001.

Its non-linear regression curve of daily trading volume about time *t* is:

$$v = 22.1317 + 9.0027t - 4.0799t^2 + 0.5176t^3 - 0.0207t^4 \tag{11}$$

Moreover, its non-linear regression curve of stock price about time *t* is:

$$u = 4.4492 - 0.0032t - 0.0143t^2 + 0.0021t^3 - 0.0001t^4 \tag{12}$$

With matlab, the related coefficients of non-linear regression curve of daily trading volume about time *t* are figured out, R^2=0.9566, F=27.5286, and the corresponding probability of F is 0.0013<0.005, which means that it is highly significant; and the related coefficients of non-linear regression curve of stock price about time *t* are R^2=0.8687, F=8.2680. The corresponding probability of F is 0.0198<0.05,

which also highly significant. Elasticity coefficients $\alpha(t)$ and trading volume in crement $h'(t)$ are calculated respectively by equations (2.3) and (2.6), and the fitting value is shown as Table 2:

TABLE 2. The fitting value of Elasticity coefficient and *h'(t)* of CRCC (601186)

Stock CRCC (Unit: Ten Thousand Hand, RMB Yuan)					
Project Name	Elasticity coefficient *α(t)*	Trading Volume Increment *h'(t)*	Project Name	Elasticity coefficient *α(t)*	Trading Volume Increment *h'(t)*
T=1	-0.0695	2.3129	T=6	0.0408	-1.9192
T=2	0.1355	-1.7675	T=7	0.1067	-0.3958
T=3	0.0638	-3.7346	T=8	-0.0228	0.7596
T=4	0.0464	-4.0848	T=9	-0.0029	1.0509
T=5	0.0389	-3.3143	T=10	0.3729	-0.0182

Predicting analysis: elasticity coefficient *α(1)*=-0.0695, and *h'(t)*=2.3129>0. According to current market situation, Stock prices just starts falling, which will be going on in the near future. Elasticity coefficient *α(10)*=0.3729, and *h' (t)* =-0.0182<0, stock price stops to rise, and the investors will get a medium to long-term lucrative profits. The above study result is more or less consistent with the trend of CRCC daily K line in Figure 1:

Figure 1. K line Graph of CRCC

4.2. Stock Price Analysis for XTC 2 (600549)

The data of trading volume and the real value of stock price for XTC 2 from 19th-30th Mar. 2012 is shown as below Table 3:

TABLE 3. The real value of trading volume stock price of XTC 2 (600549)

Stock XTC 2 (Unit: Ten Thousand Hand, RMB Yuan)						
Project Name	Trading Volume *v*	Closing Price *u*	Project Name	Trading Volume *v*	Closing Price *u*	Project Name
T=1	20.55	44.79	T=6	12.45	43.8	T=1
T=2	19.62	45.7	T=7	16.53	45.19	T=2
T=3	33.05	47.77	T=8	15	44.39	T=3
T=4	17.59	45.25	T=9	14.41	42.17	T=4
T=5	12.42	44.76	T=10	11.24	41.21	T=5

Multiple regression coefficients $b_4^{'}, b_3^{'}, b_2^{'}, b_1^{'}, b_0^{'}$ worked out by matlab are equaled to -2.7825, 31.8658, -11.4730, 1.4880, -0.0646, and $a_4^{'}, a_3^{'}, a_2^{'}, a_1^{'}, a_0^{'}$ are 40.0017, 6.6249, -2.2249, 0.2831, -0.0126.

Its non-linear regression curve of trading volume about time *t* is:

$$v = -2.7825 + 31.8658t - 11.4730t^2 + 1.4880t^3 - 0.0646t^4 \tag{13}$$

Its non-linear regression curve of stock price about time *t* is:

$$u = 40.0017 - 6.6249t - 2.2249t^2 + 0.2831t^3 - 0.0126t^4 \tag{14}$$

With matlab, the related coefficients of non-linear regression curve of daily trading volume about time *t* are figured out, R^2=0.6115, *F*=1.9677, and the corresponding probability of *F* is 0.2379>0.005, which means that it is not highly significant; and the related coefficients of non-linear regression curve of stock price about time *t* are R^2 =0.8268, *F*=5.9676. The corresponding probability of *F* is 0.0383<0.05, which is highly significant. From this, we can conclude that non-linear regression curve of daily trading volume about time t of XTC 2(600549) is not highly significant. It can be easily found that the third point is abnormal with the help of residual and confident interval of matlab as well as scatter Figure 2:

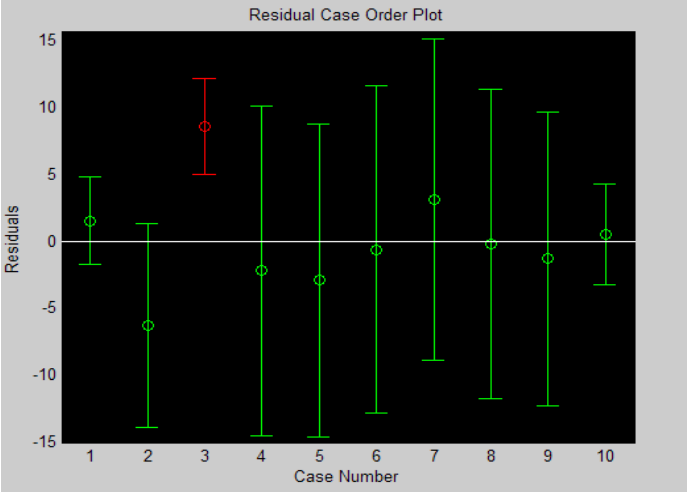


Figure 2. Residual Scatter Diagram

To recheck the multiple regression coefficients after deducting the third point, we get a new result that $b_4^{'}, b_3^{'}, b_2^{'}, b_1^{'}, b_0^{'}$ are 17.6944, 6.5902, 4.2174, 0.7539, -0.0418. The updated non-linear regression curve of XTC 2 (600549) is:

$$v = 17.6944 + 6.5902t - 4.2174t^2 - 0.7539t^3 - 0.0418t^4 \tag{15}$$

The relevant coefficients R^2=0.8586, *F*=6.0746, and its results become more significant after transformation.

Respectively, the elasticity coefficient $\alpha(t)$ and trading volume increment $h'(t)$ are worked out by equations (2.3) and (2.6). It's fitting value is shown as below Table 4:

TABLE 4. The fitting value of trading volume and *h'(t)* stock price of XTC 2 (600549)

Stock XTC 2 (Unit: Ten Thousand Hand, RMB Yuan)						
Project Name	Elasticity coefficient *α(t)*	Trading Volume Increment *h'(t)*	Project Name	Elasticity coefficient *α(t)*	Trading Volume Increment *h'(t)*	Project Name
T=1	5.5343	0.25	T=6	-0.0936	1.3005	T=1
T=2	-0.1168	-2.5697	T=7	-0.0645	1.0409	T=2
T=3	Abnormal Point	Abnormal Point	T=8	0.0851	-1.7153	T=3
T=4	0.1508	-1.6586	T=9	0.043	-7.971	T=4
T=5	-3.0837	0.0664	T=10	-0.0091	-18.729	T=5

Predictive analysis: Elasticity coefficient $\alpha(1)$=5.5343, and $h'(t)$=0.25,which indicates that the stock price will go up with the increase of trading volume, and the upward trend is fixed, investors are encouraged to buy more stocks; $\alpha(4)$=0.1508, and $h'(t)$= -1.6586, if price is about to go down, the decline will be going on for a period; $\alpha(9)$=0.0430 and $h'(t)$=-7.9710 means that the bear's power is sapped, while the bull starts to strike back, which prevents the price from going down, so investors are suggested to buy. The above study results well tally with the trend of daily K line of XTC 2 in Figure 3.

Figure 3. K line Graph of XTC

4.3. Stock Price Analysis for BCDC (600376)

The data of trading volume and the real value of price for BCDC from 12th-25th Oct. 2012 is shown as below Table 5:

TABLE 5. The real value of trading volume and turnover of price for BCDC

Stock BCDC (Unit: Ten Thousand Hand, RMB Yuan)						
Project Name	Trading Volume v	Closing Price u	Project Name	Trading Volume v	Closing Price u	Project Name
T=1	19.75	8.25	T=6	7.23	7.78	T=1
T=2	13.83	8.45	T=7	11.95	7.46	T=2
T=3	8.45	8.27	T=8	5.69	7.44	T=3
T=4	7.69	8.21	T=9	11.34	7.64	T=4
T=5	8.15	7.9	T=10	15.19	8.03	T=5

Multiple regression coefficients $b_4', b_3', b_2', b_1', b_0'$ worked out by matlab are equalled to 34.9367, -19.2013, 5.0560, -0.5818, 0.0248, and $a_4', a_3', a_2', a_1', a_0'$ are 7.9150, 0.4759, -0.1331, 0.0069, 0.0002.

Its non-linear regression curve of trading volume about time t is:

$$v = 34.9367 - 19.2013t + 5.0560t^2 - 0.5818t^3 + 0.0248 \tag{16}$$

Its non-linear regression curve of stock price about time t is:

$$u = 7.9150 + 0.4759t - 0.1331t^2 + 0.0069t^3 + 0.0002t^4 \quad (17)$$

With matlab, we see that the related coefficients of non-linear regression curve of daily trading volume about time t are R^2=0.8428, F=6.6998, and the corresponding probability of F is 0.0305<0.05, which means that it is highly significant; and the related coefficients of non-linear regression curve of stock price about time t are R^2=0.9797, F=60.3601. The corresponding probability of F is 0.0002<0.05, which indicates that the fitting degree is highly significant. To use equations (2.3) and (2.6), the value of elasticity coefficient $\alpha(t)$ and trading volume increment $h'(t)$ are worked out, whose fitting value are shown as below Table 6:

TABLE 6. The fitting value of trading volume and *h'(t)* stock price of BCDC

Stock BCDC (Unit: Ten Thousand Hand, RMB Yuan)						
Project Name	Elasticity coefficient *α(t)*	Trading Volume Increment *h'(t)*	Project Name	Elasticity coefficient *a(t)*	Trading Volume Increment *h'(t)*	Project Name
T=1	-0.0527	-10.7354	T=6	-2.3822	0.1008	T=1
T=2	-0.0091	-5.1642	T=7	-1.0045	0.144	T=2
T=3	0.0683	-1.8914	T=8	0.0475	0.8707	T=3
T=4	0.663	-0.3205	T=9	0.1328	2.8773	T=4
T=5	-1.7671	0.1448	T=10	0.1667	6.76	T=5

Predictive analysis: elasticity coefficients *α(3)*=0.0683, and *h'(t)*= -1.8914, *α(4)*=0.6630, and *h'(t)*=-0.3205, price starts falling and will keep going down; after few days decline, elasticity coefficients *α(6)*=-2.3822, and *h'(t)*=0.1008, *α(7)*=-1.0045, and *h'(t)*=0.1440, stock is now at a low rate, and trading volume increases. What's more, a fresh main force is building its position, and it means that the floor of stock price is almost formed if trading volume turns up in the coming days. The study results above are coincided with the daily K line's trend of BCDC in Figure 4:

Figure 4. Analysis Graph based on K line

5. Conclusion

According to quartic non-linear regression model of stock price and trading volume about time, a predictive model of elasticity coefficient for stock price is built based on relationship between stock price and trading volume. The model is taken as a significant technical analysis indicator to be used in studying the change of stock price and trading volume. The predictive model of elasticity coefficient is proved to be effective through the empirical analysis and examination on the three sample stocks above, China

Railway Construction Company (601186), Xiamen Tungsten Industry (600549) and Beijing Capital Development Holding (Group) Co. Ltd. (600376).

Acknowledgment

This research is financially supported by Fujian Soft Science Funds (2012R0003) and Fujian Social Funds (2013B206).

References

1. Qin, P., Wang, X. (2013) Research on the Performance Evaluation of Financial Management with Triangular Fuzzy Linguistic Information, *JDCTA: International Journal of Digital Content Technology and its Applications*, 7(2), 430-437.
2. Kumar, S., Managi, S., Matsuda, S. (2012) Stock prices of clean energy firms, oil and carbon markets: A vector autoregressive analysis, *Energy Economics*, 34(1), 215-226.
3. Ni, S.X., Pearson, N.D., Poteshman, A.M. (2005) Stock price clustering on option expiration dates, *Journal of Financial Economics*, 78(1), 49-87.
4. Zhai, A.M., Zhou, T. (2011) The Stock Market Price Volume Relation Analysis Based on Behavioural Assumption for Market Actor, *Chinese Journal of Management Science*, 19(4), 31-37.
5. Security Association of China. Basic Knowledge of the Stock Market, *China Financial and Economic Publishing House*, Beijing, 2010.
6. Zhang, X.D. (2007) Basic Knowledge of Econometrics III, *Nankai University Press*, Tianjin.
7. Wei, W.X., Gao, F.M. (2001) An Empirical Analysis on the Interaction of Price and Trading Volume of Shanghai's Stock Market.
8. Huang, Y.Y. (2006) Application of Non-linear Regressive Model in Stock Price Forecast, *Journal of Xiaogan University*, 26(3), 62-64.
9. Di, A.M., Wang, X.F. (2010) Model of Relations Between Stock Price and Trading Volume Based on Theoretical Plasticity Model, *Statistics and Decision*, 3, 144-146.
10. Epps, T.W., Epps, M.L. (1976) The stochastic dependence of security price changes and transaction volumes: Implications for the mixture-of-distributions hypothesis, *Econometrica*, 44(2), 305-321.
11. Li, S.L. (2005) Analysis on Behavioural Finance and Security Investment Behaviour, *Social Sciences Academic Press (China)*, Beijing.
12. Guo, W.Y. (2007) Analysis on Heterogeneity of Investors, *Capital University of Economics & Business Press*, Beijing.
13. Fang, Z.M., Wang, C.F. (2007) Analysis on Relationship between Stock Price and Trading Volume under the Perspective of Trading Frequency, *Journal of Systems Engineering*, 22(4), 386-393.
14. Song, J., Wu, C.F. (2001) Analysis on Causes and Countermeasures for Herding in Financial Market, *The Theory and Practice of Finance and Economics*, 22(6), 46-48.
15. Song, J., Wu, C.F. (2001) Analysis on Herding in Financial Market Based on Dispersion, *Economics Research*, 36(11), 21-27.
16. Sun, P.Y., Shi, D.H. (2002) Analysis on Herding in China Security Market Based on CAPM, *Economics Research*, 37(2), 64-70.
17. Yang, C.P., Wu, C.F., Chen, M. (2005) Behavioural Finance: Perceived Risk and Perceived Expected Return, *Chinese Journal of Management Science*, 13(3), 15-18.

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 270-276
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

GTS LÈVY MODEL ON TECHNOLOGY OF FINANCIAL ENGINEERING

Ch. Wang [1]*

[1]Math Dept. China JiLiang Univ, HangZhou, 310018, China
** Corresponding author E-mail address: xsmdoctor@163.com*

Classic stable Lèvy models have intrinsic drawbacks when applied on datas from engineering or financial technology. The generalized tempered stable Lèvy (GTS) models overcome the drawbacks. The parameters in GTS capture many characteristics and features of financial datas (such as big or small, positive or negative jumps of financial assets). GTS models are flexible for different periods of financial markets. Then C-GMM estimation method is applicable to obtain the parameters. From the parameters, we can see the difference of many aspects between the datas from different market periods.

Keywords: GTS Lèvy model; C-GMM estimation; Characteristic function.

1. Introduction

Exponential Lèvy processes have long been used in mathematical finance. Osborne (1959) proposed the exponential e^{B_t} of Brownian motion as a stock price model. The process e^{B_t} called exponential or geometric Brownian motion was also use by Samuelson (1965) in a more systematic manner.

One of the first to propose an exponential non-normal Lèvy process was Mandelbrot (1963). He observed that the logarithm of stock price changes on financial markets exhibit a long-tailed distribution. He pointed that the B_t in e^{B_t} should be replaced by symmetric $\alpha-$stable Lèvy motion with index $\alpha<2$. Recall that the normal distribution is 2-stable. A few years later, an exponential Lèvy process with a non-stable distribution was proposed by Press (1967). Lu (2005), Wu and Tang (2007), Liu and Han (2007) used stable distributions to model China stock markets.

In recent years, Madan etc. (1987, 1990, 1998), Barndorff-Nielsen (1995, 1997, 1998) have proposed several non-stable Lèvy processes such as VG(Variance Gamma) and NIG(Normal Inverse Gauss) process respectively to model financial market. VG and NIG processes are both subordinations (time changed Brownian motion) of Brownian motion in which the subordinators are gamma process and inverse Gaussian process respectively. (On details of subordinations, see Sato 1999). A wider class that contains VG and NIG is CGMY process, which is described in detail by Koponen (1995) and Carr, Geman, Madan, and Yor (2002). Liu (2009) applied Truncated Levy (TL, also called Levy Flight) and CGMY on stock markets.

Definition 1.1 (Cont, 2004). A generalized tempered stable process (GTS process) is a Lèvy process with no Gaussian component and a Lèvy density of the form

$$v(x)=\frac{c_-}{|x|^{1+\alpha_-}}e^{-\lambda_-|x|}1_{x<0}+\frac{c_+}{|x|^{1+\alpha_+}}e^{-\lambda_+x}1_{x>0}, \tag{1}$$

with $\alpha_+<2$ and $\alpha_-<2$.

Let X_t be a GTS Lèvy process with drift γt, the CF (characteristic function) of X_1 (denoted by $\Psi(s)$) is given in Cont (2004). For example, if $\alpha_\pm\neq1$ and $\alpha_\pm\neq0$, the CF is

$$\psi(s)=\exp\left\{is\gamma+\Gamma(-\alpha_+)\lambda_+^{\alpha_+}c_+\left[\left(1-\frac{is}{\lambda_+}\right)^{\alpha_+}-1+\frac{is\alpha_+}{\lambda_+}\right]+\right.$$
$$\left.\Gamma(-\alpha_-)\lambda_-^{\alpha_-}c_-\left[\left(1+\frac{is}{\lambda_-}\right)^{\alpha_-}-1+\frac{is\alpha_-}{\lambda_-}\right]\right\} \quad (2)$$

Remark 1.1 GTS process is obtained by taking a 1-dimensional stable process and multiplying the Lèvy measure with a decreasing exponential on each half of the real axis. After this exponential damping, the small jumps keep their initial stable-like behaviour whereas the large jumps become much less violent.

Remark 1.2 (Cont, 2004): GTS contain compound Poisson process as a special case when $\alpha_\pm$ are both negative. GTS has trajectories of finite variation if $\alpha_\pm$ are both less than one. In the limiting case $\alpha_+=\alpha_-$, the GTS becomes VG if $c_+=c_-$. The GTS, which can be represented as a time changed Brownian motion, coincides with CGMY. Therefore, GTS has greater freedom than other processes above to model financial data.

The parameters $\lambda_\pm$ determine the tail behaviour of the Lèvy measure, they tell us how far the process may jump, and from the viewpoint of a risk manager, this corresponds to the amount of money that we can lose (or gain) during a short period of time. The parameters $c_\pm$ determine the overall and relative frequency of upward and downward jumps; of course, the total frequency of jumps is infinite, but if we are interested only in jumps larger than a given value, then these two parameters tell us how often we should expect such events. Finally, $\alpha_\pm$ determine the local behaviour of the process (how the price evolves between big jumps). When $\alpha_\pm$ are close to two, the process behaves much like a Brownian motion, with many small oscillations between big jumps. On the other hand, if $\alpha_\pm$ are small, most of price changes are led by big jumps during a period of time. We see GTS dividedly describes positive and negative (big or small) jumps, so it is more suitable for financial market during different periods.

In order to use GTS to model the China stock market, in section 2, we first review the China stock market during 1990-2011, the C-GMM method for estimating parameters, and then calculate the necessary quantity in our problem in order to apply the C-GMM method.

2. China Stock Market and Estimating Methods of Parameters

2.1. Outline of China Financial Market

China economy has changed from a centrally planned economy (CPE), which was introduced in 1949, to a more market orientated economy since 1978 and is currently a significant participant in the global economy. Many inherent shortcomings of the CPE is not suitable to global trend and block the economic growth. This led to the reforms in the late 1970's, which started with the de-collectivisation of agriculture, the gradual liberalization of prices, a diversified banking system, development of the stock markets, the growth of non-state sector and the opening to foreign trade and investment.

In mainland China, the stock market which appeared before 1949 reappeared in the 1980's and has experienced a lot of growth ever since. In 1990, the first and second exchange authorized by the State Council, Shanghai Stock Exchange and ShenZhen Stock Exchange were established by the People's Bank of China. These two are most major exchanges in China.

China capital market has been stagnant from 2001 to 2005. In this period, Shanghai and Shenzhen Composite Indexes range from about 1500 to about 2500 and from about 3000 to about 6000 respectively. However, in 2006 and 2007, China stock market experienced surprisingly rapid and large rise, Shanghai and Shenzhen Composite Indexes reaching more than 6000 and more than 19600 respectively.

From 2008, China stock market has been undergoing dramatic rise and fall of big size frequently until now. Therefore, we separately study the relatively stagnant period 2001 to 2005 and the fluctuating period 2006 to 2013. For each period, we use GTS process to model the log-index movement and apply the method in next section to estimate parameters, and compare the parameters in two periods to see the market movement difference between two periods.

2.2. C-GMM Estimation

We use C-GMM (Generalized moment method under a continuum of moment conditions) method proposed by Carrasco and Florens (2007) to estimate the parameters in our model. C-GMM has several advantages: (i) it does not need to know probability density function (in most Lèvy model, the p.d.f has not an explicit formula); (ii) it avoids the singularity of covariance matrix; (iii) it allows the data are weakly dependent and correlated moment functions.

We first outline the theory of this approach. Let $\{X_t\}$ be a p -dimension stochastic process indexed by parameter $\theta \in \Theta \subset R^q$. Singleton (2001) propose an estimating method under CCF (conditional characteristic function). The CCF of $\{X_{t+1}\}$ given $\{X_t\}$ is

$$\Psi_\theta(s|X_t) = E^\theta\left(e^{isX_{t+1}}|X_t\right) \tag{3}$$

From the property of conditional expect, we have

$$E^\theta\left[\left(e^{isX_{t+1}} - \Psi_\theta(s|X_t)\right)A(X_t)\right] = 0 \quad \forall s \in R^p, \tag{4}$$

where $A(X_t)$ is an arbitrary instrument. The equation (3) is called moment condition. The expression $\left(e^{isX_{t+1}} - \Psi_\theta(s|X_t)\right)A(X_t)$ is called moment function. Besides being a function of X_t, A may be a function of an index r equal to or different from s. Two types of moment functions are used often: Single Index (SI) and Double Index (DI) (Singleton 2001).

SI moment function is of the form $h_t(s;\theta) = \left(e^{isX_{t+1}} - \Psi_\theta(s|X_t)\right)A(s, X_t)$, where $s \in R^p$.

DI moment function is of the form $h_t(\tau;\theta) = \left(e^{isX_{t+1}} - \Psi_\theta(s|X_t)\right)A(r, X_t)$, where $\tau = (r,s)' \in R^{2p}$.

Carrasco and Florens (2007) use DI instrument:

$$A(r, X_t) = e^{irX_t} \tag{5}$$

Denote $(r,s)' \in R^{2p}$ by τ, then the corresponding moment function is

$$h_t(\tau;\theta) = \left(e^{isX_{t+1}} - \Psi_\theta(s|X_t)\right)e^{irX_t}. \tag{6}$$

In the next subsection, we will calculate the moment function in our problem The C-GMM estimator is based on the arbitrary set of moment conditions:

$$E^{\theta_0} h_t(\tau;\theta_0) = 0, \ \forall \tau. \tag{7}$$

Let $\hat{h}_T(\tau;\theta_0) = \sum_{t=1}^T h_t(\tau;\theta_0)/T$. It is an estimator of $E^\theta h_t(\tau;\theta_0)$, the problem is reduced to find the estimator $\hat{\theta}$, which best satisfies the equation $\hat{h}_T(\tau;\theta_0) = 0$ for all τ.

Adapting operator theory and probability theory, they obtain the following essential and nice results under some regularity conditions in their paper.

Proposition 2.1 Let B be a one-to-to one bounded linear operator defined on $L^2(\pi)$ and B_T be a sequence of random bounded linear operators converging to B as $T \to \infty$. The C-GMM estimator

$$\hat{\theta}_T = \arg\min_{\theta\in\Theta} \left\| B_T \hat{h}_T(\theta) \right\| \tag{8}$$

has the following properties:

$\hat{\theta}$ is consistent and asymptotically normal such that

$$\sqrt{T}\left(\hat{\theta}_T - \theta_0\right) \xrightarrow{L^1} N(0,V) \tag{9}$$

with

$$\begin{aligned} V &= \left\langle BE^{\theta_0}\left(\nabla_\theta h\right), BE^{\theta_0}\left(\nabla_\theta h\right)\right\rangle^{-1} \times \left\langle BE^{\theta_0}\left(\nabla_\theta h\right), \left(BKB^*\right)BE^{\theta_0}\left(\nabla_\theta h\right)\right\rangle \\ &\times \left\langle BE^{\theta_0}\left(\nabla_\theta h\right), BE^{\theta_0}\left(\nabla_\theta h\right)\right\rangle^{-1} \end{aligned} \tag{10}$$

Theorem 2.1 A good estimate is the solution of the following problem

$$\min_\theta \underline{w}(\theta)\left[\alpha_T I_T + C^2\right]^{-1}\underline{v}(\theta) \tag{11}$$

where C is a $T\times T$ -matrix with (t,l) element $c_{tl}/(T-q)$, $t,l=1,...,T$, I_T is the $T\times T$ identity matrix, $\underline{v}=\left[v_1,...,v_T\right]'$ and $\underline{w}=\left[w_1,...,w_T\right]'$ with

$$v_t(\theta) = \int \bar{h}_t\left(\tau;\hat{\theta}_T^1\right)\hat{h}_T(\tau;\theta)\pi(\tau)d\tau, \tag{12}$$

$$w_t(\theta) = \left\langle h_t\left(\tau;\hat{\theta}_T^1\right), \hat{h}_T(\tau;\theta)\right\rangle, \tag{13}$$

$$c_{tl} = \int \bar{h}_t\left(\tau;\hat{\theta}_T^1\right)\hat{h}_l\left(\tau;\hat{\theta}_T^1\right)\pi(\tau)d\tau \tag{14}$$

Carrasco and Florens (2007) shows the resulting C-GMM estimator will be as efficient as the MLE provided that the full continuum of moment conditions indexed by τ is used. This methodology has two advantages: (i) the instrument e^{irX_t} is of simple form; (ii) in contrast to Singleton's method, it does not depend on discretization grid involved in numerical implementation of integration over τ.

2.3. Calculation of the Moment Function

Let $\{S_t\}_{t\geq 0}$ denote the stock index process. The log-index $X_t = \ln(S_t)$ is assumed to be a Lèvy process as usual. We write $\Psi_\theta(s)$ for unconditional characteristic function $\Psi(s)$ in section 1, where θ denotes the parameter vector $(\alpha_+,\alpha_-,\lambda_+,\lambda_-,c_+,c_-,\gamma)'$. Due to the independent increments of X_t and the properties of conditional expect, we have

Proposition 2.2. The moment function (4) for X_t is

$$h_t(\tau;\theta) = \left(e^{isX_{t+1}} - e^{isX_t}\Psi_\theta(s)\right)e^{irX_t}. \tag{15}$$

with $h_t(\tau;\theta)$ in the above proposition, the formulas in Theorem 2.1, programming in Matlab, we can obtain parameters in GTS model for China market in the next section.

3. Empirical Study and Analysis

From the analysis in section 1, we divide China stock market into two periods: 1st Jan 2001-31st Dec 2005 and 1st Jan 2006-31st Aug 2013. The data we used are Shanghai Composite Index and Shenzhen Composite Index (from Yahoo Finance).

We first compare X_t with Brownian motion, the difference is large (see figure 1, 2, 3, the empirical data exhibits an obvious fat tail and skewness). The process $\ln X_t$ has many jumps, positive and negative jumps having distinct features. Therefore, we select GTS to model $\ln X_t$ to reflect the difference. We use C-GMM based on CF (characteristic function), coding Matlab 7.6 programme to obtain the parameter estimates. The results are in Table 1 and Table 2, where the numbers in parentheses are the corresponding standard errors.

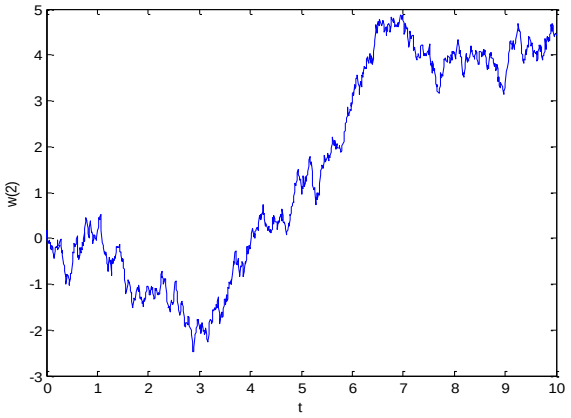


Figure 1. Simulation of Brownian motion

Figure 2. Data from real market

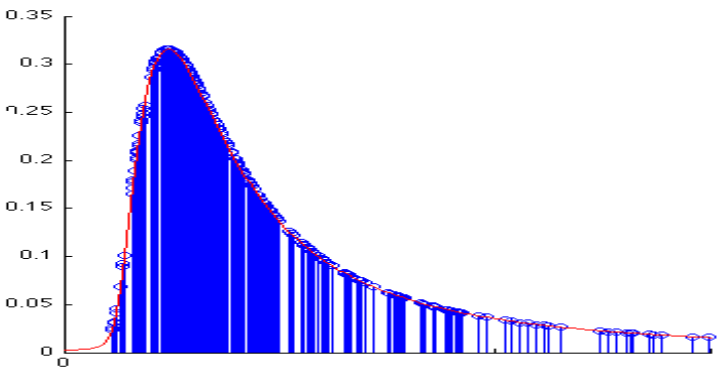


Figure 3. Fitting of log-return from real market

From the results in the Tables, we see the following facts:

(1) Using the estimating parameters to calculating the cumulants (Cont 2004) such as the mean value, variance, kurtosis, skewness, we find they are almost consistent with the empirical results. This indicates the GTS Lèvy process can capture these important characteristics very well.

TABLE 1. Parameter estimates for Shanghai Composite Index

Period	α_+	α_-	λ_+	λ_-	c_+	c_-	γ
2006.1-2013.8	0.2341(0.0170)	0.3546(0.0215)	42.1073(5.7843)	32.2510(2.2007)	0.1205(0.0218)	0.2519(0.0310)	0.0012 (0.0018)
2001.1-2005.12	0.6284(0.1003)	0.4356(0.0528)	62.5170 (8.0120)	45.7208 (3.1043)	2.0824(0.2100)	1.1983(0.0681)	0.0063 (0.0250)

TABLE 2. Parameter estimates for Shenzhen Composite Index

Period	α_+	α_-	λ_+	λ_-	c_+	c_-	γ
2006.1-2013.8	0.2671(0.0315)	0.4012(0.0807)	65.2750(8.2009)	57.2108(4.7901)	0.7489(0.2062)	2.1005(1.2108)	0.0204 (0.033)
2001.1-2005.12	0.8250(0.1120)	0.4916(0.0413)	81.0970(7.9020)	67.2290(6.3215)	4.3154(0.8870)	2.7408(0.5201)	1.8942(1.1895)

(2) The parameters $\frac{c_1}{c_1+c_2}$ and $\frac{c_2}{c_1+c_2}$ reflect the relative frequencies of positive and negative jumps respectively. For Shanghai Composite Index, we see the upward jumps frequency is higher than downward jumps frequency in 2001-2005, while the case is the opposite way in 2006-2013. The results of Shenzhen Composite Index are similar.

(3) For the two Indexes, λ_+ is greater, than λ_-, which shows the decay rate of right tail is faster than that of left tail. That leads to the left-skewness of the distribution. In addition, $\lambda_+(\lambda_-)$ in 2001-2005 is bigger than $\lambda_+(\lambda_-)$ in 2006-2013, which means the big jumps in 2001-2005 are less often than those in 2006-2013. This corresponds to the market's situation.

(4) The parameters α_+ and α_- capture the features of small jumps. For the two indexes, $\alpha_+ < \alpha_-$ in 2006-2013, while $\alpha_+ > \alpha_-$ in 2001-2005. This shows small upward jumps are less than small downward jumps in 2006-2013, while the period 2001-2005 is the opposite way. From the trend figure of stock market, we can intuitively see this fact. In addition, $\alpha_+(\alpha_-)$ in 2001-2005 is larger than $\alpha_+(\alpha_-)$ in 2006-2013, showing small jumps in 2001-2005 are more frequent than those in 2006-2013, which is opposite to big jumps' feature in (3). The process $\ln X_t$ in 2001-2005 is much closer to continuous change than that in 2006-2013.

4. Discussion

Compared with geometric Brownian motion and classic Lèvy stable (stable) process, GTS Lèvy stable process is more suitable for asset price movement. It has finite moments, describing kurtosis and skewness very well. Seven parameters, each representing one feature of the price movement make it more flexible for modelling the process.

Our study of the most important stock indexes of China, Shanghai Composite Index and Shenzhen Composite Index, shows jump is the intrinsic feature of financial asset movement, positive and negative jumps happening often with different frequency. The parameters for the two indexes during two periods are all in the interval (0, 1) which means the GTS Lèvy process for stock markets has infinitely jumps (also called infinitely activity by researchers), but has finite variation. From, we see the indexes movement has scarce big jumps, yet with many small jumps between big jumps.

Acknowledgment

This research was supported by a grant from National Natural Science Foundation of China (No. 10901137).

References

1. Barndoff-Nielson, O.E. (1995) Normal inverse Gaussian distribution and the modelling of stock returns. Technical report, Research Report NO. 300, Department of Theoretical Statistics, Aarhus University.
2. Barndoff-Nielson, O.E. (1997) Normal inverse Gaussian distribution and stochastic volatility modelling. *Scandinavian Journal of Statistics*, 24, 1-13.
3. Barndoff-Nielson, O.E. (1998) Processes of normal inverse Gaussian type. *Finance and Stochastics*, 2, 41-68.
4. Carrasco, M., Chernov, M., Florens, J. (2007) Efficient estimation of general dynamic models with a continuum of moment conditions. *Journal of Econometrics*, 140, 529–573
5. Carrasco, M., Florens, J.P. (2000) Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 16, 797–834.
6. Carr, P., Geman, H., Madan, D., Yor, M. (2002) The fine structure of asset returns: an empirical investigation. *Journal of Business*, 75, 75-83
7. Cont, R., Tankov, P. (2004) Financial modelling with jump processes. Chapman and Hall/CRC.
8. Koponen, I. (1995) Analytic approach to the problem of convergence of truncated Lèvy flights towards the Gaussian stochastic process. *Physical Review*, 52, 1197-1199.
9. Liu, X.H., Han, Q.H. (2007) The stable distributions of Shenzhen Composite Index and Shanghai Composite Index. *Mathematical statistics and Management*, 27(2), 341-345.
10. Liu, Z.D., Chen, X.J. (2009) Applying Levy TS and C-GMM on financial asset returns. *Chinese Management Science*, 17(3), 18-26.
11. Lu, F.Y. (2005) A case study for the distribution of China stock market. *Chinese Management Science*, 13, 172-175.
12. Mandelbrot, B. (1963) The variation of certain speculative process. *The Journal of Business*, 36, 394-419.
13. Madan, D.B., Seneta, E. (1987) Chebyshev polynomial approximations and characteristic function estimation. *Journal of the Royal Statistical Society Series*, B49(2), 163-169.
14. Madan, D.B., Seneta, E. (1990) The V.G. model for share market returns. *Journal of Business*, 63, 511-524.
15. Madan, D.B., Carr, P.P., Chang, E.C. (1998) The variance gamma process and option pricing. *European Finance Review*, 2, 79-105.
16. Osborne, M.F.M. (1959) Brownian motion in the stock market. Operations Research, 7, 145-173.
17. Press, S.J. (1967) A compound events model for security prices. *Journal of Business*, 40, 317-335.
18. Samuelson, P. (1965) Rational theory of warrant pricing. *Industrial: Management Review*, 6, 13-32.
19. Singleton, K. (2001) Estimation of affine pricing models using the empirical characteristic function. *Journal of Econometrics*, 102, 111–141.
20. Wu, D., Tang, Y.C. (2007) Stable distribution and its application on finance. *Applied Probability and Statistics*, 23(4), 434-445.

Received on the 17th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 277-288
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

MEASUREMENT ABOUT INFORMATION ASYMMETRY DEGREE OF CHINESE MEDICAL SERVICE MARKET

Xu Wei[1,2,*], Ding Yunlong[1], Liu Yang[1]

[1]School of Management, Harbin Institute of Technology, Harbin, 150001, China
[2]The Fourth Affiliated Hospital of Harbin Medical University, Harbin, 150001, China
[*] Corresponding author E-mail address: xuwei_6666@163.com

The article constructed a measurement model about information asymmetry degree of Chinese medical service market, and performed demonstrating measurement to the information degree of both sides of physician and patient in medical service market and its impacting effect to the eventual medical service price based on the microcosmic individual survey data in "Chinese Health and Nutritional Survey" (CHNS). The article has the policy implication that it emphasized that whether the reform ideas of strengthening the modulating effect of market mechanism in medical service by introducing competence is suitable to China is deserved to rethink. Resolving the problem of medical service price virtual height and making it return the commonweal of medical service need our government to participate more and more and exert the function such as price regulation, market surveillance, and externality correction and so on effectively.

Keywords: Medical Price; Information Asymmetry; Bilateral Stochastic Frontier Analysis; Benchmark Price.

1. Introduction

Chinese government set out a series of healthcare reform packages aimed to increase the accessibility of basic medical services obviously, relieve the resident medical expense burden effectively, and settle the problem of "difficulty of receiving medical service and high cost of getting medical treatment". Obviously, among numerous factors affecting the residents' medical services demands, medical service price is the most critical [1]. Hence, it is particularly urgent and have great practical significance to analyse the problem about the actual medical expense burden in medical service market.

Yet, the actual price reached in the competitive medical service market seemingly existed great discrepancy which couldn't be explained completely by the difference in the character of the physician or patient and service quality, the information asymmetry is absolutely one of the factors of considerable importance. The correlation research in prophase showed that the medical service market would form an effective medical service price automatically under the strictly condition. Research by Wolinsky (1993) found that consulting different medical experts could decrease even eliminate the deceit by medical staffs, but the assumed condition implied in this article is that the information gathering cost is zero. It is proved by Alger & Salanie (2006) that consumer can obtain the information about lowest treatment price at low cost and the specialist deceit is eliminated under the hypothesis of perfect competition for full information of medical market. Huangtao and Yantao (2009) constructed signal gambling model to analyse the excessive treatment and pointed out that it had the restraint function to search the decision and the specialist punishment mechanism by introducing the consumer knowledge under specified conditions from the point of view of trusting commodity [2].

However, medical service market has very significant information asymmetric characteristic (Phelps, 1997). Because the side of medical information supplier (hospital) has information superiority which will induce the moral hazard of the side of medical service supplier-induced demand (Feldstein, 1970) [3]. Rice (1983) found that the medical service quality improved while the physician service cost decreased. The patient structure data analysis result about seven important surgeries arranged by doctors of Ticino state in Switzerland by Emons (1997) revealed that the proportion of common patients is 33% over the patients whose identity is physician or the household of the physician [4]. Based on the induced demand hypothesis, Yip (1998) inferred that the physician would make up the decrease in income by increasing the amount of service once the physician profession was affected by income effect. Huanpeng Zhu (2007) pointed out that the market information asymmetry in the medical service market was particularly serious, and the in accordance and even conflict economic benefit between physician and patient provided the possibility and space for the morality risk behaviour of the physician: The physician could pursue self economic benefit regardless of the patient's benefit, and lead the patients to consume

excessive medical service and drug by taking advantage of the information superiority [5]. The incentive mechanism in the institutional arrangement made the physician's income be closely related to the diagnosis and treating income, and they could induce more medical demands, increase the health expenditure of patients or medical system by making the best use of the information asymmetry and uncertainty [6].

In view of the important role of information factor in medical service pricing making, it is necessary to intensify the regulating function of government by starting with researching the market behaviour for resolving the medical service price virtual height problem practically [7]. Although we hope urgently to remit the information imperfection of medical service market and further increase the medical service accessibility to patients, we have not sought out the research literature about measuring Chinese medical service market information asymmetry degree [8].

2. The Information Asymmetry Degree Measurement Model in Medical Service Market Formation

It is presumed that there are many medical service suppliers and demanders in a typical medical service market, and physicians and patients both master certain information [9]. The eventual price of medical service (P) can be formulated as following:

$$p = \underline{p} + \eta\left(\overline{p} - \underline{p}\right) \tag{1}$$

There into, $\underline{p}$ is the lowest medical service price the physician can accept, $\overline{p}$ is the highest medical service price the patient would like to pay. $\eta\left(0 \le \eta \le 1\right)$ is used for weighing the information degree mastered by the physician during the pricing, hence $\eta\left(\overline{p} - \underline{p}\right)$ reflects the surplus seized by the physician during the medical service price formation.

It is needed to further resolve the formula (1) for embodying the degree of the information mastered by the physician and patient simultaneously during the pricing course in the model. Firstly we describe the "fair" medical service price $\mu\left(x\right) = E\left(\frac{\theta}{x}\right)$ under condition that the individual character x is given, here θ is subsistent but cannot be informed, and almost meets with that: $\underline{p} \le \mu\left(x\right) \le \overline{p}$. Hence $\left(\overline{p} - \mu\left(x\right)\right)$ represents the patient's anticipating surplus during the medical service price formation; $\left(\mu\left(x\right) - \underline{p}\right)$ represents the physician's anticipating surplus. In addition, which side can "seize" more surplus will depend on the degree of the mastered information and the bargaining power based on it (Osbourne & Rubinstein, 1990). We can express the formula (1) again as follow using the definition of these surpluses:

$$\begin{aligned} p &= \mu\left(x\right) + \left[\underline{p} - \mu\left(x\right)\right] + \eta\left[\overline{p} - \mu\left(x\right)\right] - \eta\left(\underline{p} - \mu\left(x\right)\right) \\ &= \mu\left(x\right) + \eta\left[\overline{p} - \mu\left(x\right)\right] - \left(1 - \eta\right)\left[\mu\left(x\right) - \underline{p}\right] \end{aligned} \tag{2}$$

Formula (2) indicates that the physicians can increase the medical service price by seizing part of patient's anticipating surplus, and the scale of the surplus seized by the physician is $\eta\left[\overline{p} - \mu\left(x\right)\right]$; and similarly, the patients can realize decreasing the medical service price by seizing part of surplus of the physicians, the scale of the surplus seized by the patient is $\left(1 - \eta\right)\left[\mu\left(x\right) - p\right]$. The surplus which can be seized by the physicians depends on the degree of the mastered information by the physicians η and the total anticipating surplus of patients $\overline{p} - \eta\left(x\right)$. It means that physicians can increase price depending on the degree of the mastered information; and the amount of surplus seized by patient depends on the degree of the mastered information $\left(1 - \eta\right)$ by the patients and the total anticipating surplus $\mu\left(x\right) - \underline{p}$ of the physicians, and the patients can reduce the medical service price by means of the degree of the mastered

information by them [10].

The pricing equation (2) is composed of three parts: The first part denotes the "fair" medical service price which is named benchmark price by us under the condition of the given individual character; the second part $\eta\left[\overline{p}-\mu(x)\right]$ embodies the seized surplus by the physicians by means of the degree of the mastered information; the third part $(1-\eta)\left[\mu(x)-\underline{p}\right]$ is the obtained surplus by the patients by means of the degree of the mastered information. The net surplus $NS=\eta\left[p-\mu(x)\right]-(1-\mu)\left[\mu(x)-\underline{p}\right]$ can be used for describing the combined effect of the information asymmetry degree [11].

Hence according to the framework of this model, physician information factor has a positive effect to the reached medical service price while patient factor has a negative effect, in another word, the information factor has bilateral influence to the medical price formation, and we can simplify the medical service price model (2) as the following form:

$$P_i=\mu(x_i)+\xi_i,\ \xi_i=w_i-u_i+v_i \tag{3}$$

The model is a typical bilateral stochastic frontier model (Kumbhakar & Christopher, 2009). There into, $\mu(x)=x_i\beta$, β is a parameter vector to be estimated, x_i is the sample individual feature, and the article includes the condition, health value expectation and the characteristic factor in other aspects: $w_i=\eta_i\left[p_i-\mu(x_i)\right]\geq 0$; $u_i=(1-\eta_i)\left[p_i-\mu(x_i)\right]\geq 0$; v_i is a stochastic disturbance item in general meaning. The physicians can increase the medical service price by means of seizing the anticipating surplus of the patients which can be embodied by w_i, and the patients can decrease the medical service price described with u_i and paid by means of obtaining part of physician's surplus. And the scale of the surplus obtained by seizing depends on the degree of the information mastered by both sides of physician and patient η, patient anticipating surplus $\overline{p}-\mu(x)$ and the physician anticipating surplus $\left(\mu(x)-\underline{p}\right)$.

For measuring the β parameter vector and the seizing surplus by both sides of physician and patient simultaneously, we use the maximum likelihood estimation (MLE) method for estimating model (3). Based on the previous statement and the model (3), we can know that the disturbance item w_i and u_i both have one-sided distribution characteristic, therefore, we hypothesize that the both sides comply to index disturbance, and it means that $u_i\sim i.i.d.Exp\left(\sigma_u,\sigma_u^2\right)$, $w_i\sim i.i.d.Exp\left(\sigma_w,\sigma_w^2\right)$. For disturbance item v_i, it is hypothesized to comply with normal distribution, i.e. $v_i\sim i.i.d.N\left(0,\sigma_v^2\right)$. Meanwhile, we hypothesize that v_i, u_i and w_i are independent of one another, and even independent of the individual character x_i. Based on the above hypothesis, the probability density function for the complex disturbance item ξ_i is inferred as follow:

$$f\left(\xi_i\right)=\frac{\exp(a_i)}{\sigma_u+\sigma_w}\Phi(c_i)+\frac{\exp(b_i)}{\sigma_u+\sigma_w}\int_{-h}^{\infty}\phi(z)dz=\frac{\exp(a_i)}{\sigma_u+\sigma_w}\Phi(c_i)=\frac{\exp(b_i)}{\sigma_u+\sigma_w}\phi(h_i) \tag{4}$$

There into, $\phi(\bullet)$ and $\Phi(\bullet)$ are the probability density function and cumulative distribution function of standard normal distribution respectively, and other parameters are set as follows:

$$a_i=\frac{\sigma_v^2}{2\sigma_u^2}+\frac{\xi_i}{\sigma_u} \tag{5}$$

$$b_i=\frac{\sigma_v^2}{2\sigma_w^2}+\frac{\xi_i}{\sigma_w} \tag{6}$$

$$h_i = \frac{\xi_i}{\sigma_v} + \frac{\sigma_v}{\sigma_w} \tag{7}$$

$$c_i = \frac{\xi_i}{\sigma_v} - \frac{\sigma_v}{\sigma_w} \tag{8}$$

For the sample containing n observed values, the log-likelihood function can be formulated as follows:

$$\ln L(X;\theta) = -n\ln(\sigma_u + \sigma_w) + \sum_{i=1}^{n} \ln[e^{a_i}\Phi(c_i) + e^{b_i}\Phi(h_i)] \tag{9}$$

There into, $\theta = [\beta, \sigma_v, \sigma_u, \sigma_w]$. We can obtain the maximum likelihood estimation of all parameters by maximizing the log-likelihood function [12]. The article emphasized on the surplus obtained by the patients and physicians by means of the degree of the mastered information, for this reason, we need to further infer the conditional distribution of u_i and w_i, which are recorded as $f\left(\frac{u_i}{\xi_i}\right)$ and $f\left(\frac{w_i}{\xi_i}\right)$ respectively, then it can be inferred that:

$$f\left(\frac{u_i}{\xi_i}\right) = \frac{\lambda \exp(-\lambda u_i)\Phi(u_i/\sigma_v + h_i)}{\Phi(h_i) + \exp(a_i - b_i)\Phi(c_i)} \tag{10}$$

$$f\left(\frac{w_i}{\xi_i}\right) = \frac{\lambda \exp(-\lambda w_i)\Phi(w_i/\sigma_v + c_i)}{\exp(b_i - a_i)[\Phi(h_i) + \exp(a_i - b_i)\Phi(c_i)]} \tag{11}$$

There into, $\lambda = 1/\sigma_u + 1/\sigma_w$. Based on the condition distribution determined in formula (10-11), the condition expectation value u_i and w_i during the formation course of medical service price can be inferred respectively, and we give the estimated formula for the two values directly:

$$E\left(1 - \frac{e^{-u_i}}{\xi_i}\right) = 1 - \frac{\lambda}{1+\lambda}\frac{[\Phi(h_i) + \exp(a_i - b_i)\exp(\sigma_v^2/2 - \sigma_v c_i)\Phi(c_i - \sigma_v)]}{\Phi(h_i) + \exp(a_i - b_i)\Phi(c_i)} \tag{12}$$

$$E\left(1 - \frac{e^{-w_i}}{\xi_i}\right) = 1 - \frac{\lambda}{1+\lambda}\frac{[\Phi(c_i) + \exp(b_i - a_i)\exp(\sigma_v^2/2 - \sigma_v h_i)\Phi(h_i - \sigma_v)]}{\exp(b_i - a_i)[\Phi(h_i) + \exp(a_i - b_i)\Phi(c_i)]} \tag{13}$$

Further, the net surplus r during the bargaining process can be expressed as:

$$NB = E\left(1 - \frac{e^{-w_i}}{\xi_i}\right) - E\left(1 - \frac{e^{-u_i}}{\xi_i}\right) = E\left(e^{-u_i} - \frac{e^{-w_i}}{\xi_i}\right) \tag{14}$$

Here, it must be emphasized especially that, because the parameter σ_u only appears on a_i and c_i, and σ_w only appears on b_i and d_i, so the above two values can be identified. Therefore, during the follow-up checkout procedure, we need not to suppose the relative size of the degree of the mastered information by both sides of physician and patient, which is determined according to the estimated result

completely. This is the fundamental advantage of the difference between the analysis method of this article and tradition regression analysis method [13].

3. Data and Indexes

3.1 Data Resource

The data of this paper were obtained from the databases of "China Health and Nutritional Survey" (CHNS), This database is founded and surveyed by the National Institute for Nutrition and Food Safety of Chinese Centre for Disease Control and Prevention (CCDC) and Medical Reform Office of the State Council [14].Eventually, 1806 pieces of observed values were selected whose distribution characteristics are shown in Table 1.

TABLE 1. Distribution of Samples (overall sample size: 1806 cases).

		Weights of Observed	Samples (%)	Rural-urban distribution (%)		Sex distribution (%)		Medical insurance (%)	
				Urban	Rural	Male	Female	Yes	No
Geographic Distribution	Liaoning	269	16.39	8.03	8.36	6.76	9.63	6.37	10.02
	Heilongjiang	139	7.70	4.60	3.10	3.49	4.21	3.16	4.54
	Jiangsu	254	14.06	5.76	8.31	5.92	8.14	8.53	5.54
	Shandong	108	5.89	2.93	3.05	2.82	3.16	2.33	3.65
	Henan	166	9.19	4.60	4.60	3.99	5.20	1.88	7.31
	Hubei	178	9.86	5.48	4.37	5.26	4.60	3.27	6.59
	Hunan	213	11.79	5.43	6.37	5.70	6.09	3.77	8.03
	Guangxi	258	14.29	5.15	9.14	6.37	7.92	3.16	11.13
	Guizhou	194	10.74	4.37	6.37	4.71	6.04	1.88	8.86
Year Distribution	2000	56	3.10	0.61	2.49	1.33	1.77	0.00	3.10
	2002	73	4.05	1.00	3.05	1.50	2.55	0.89	3.16
	2004	138	7.64	2.93	4.71	3.99	3.65	1.66	5.98
	2006	187	10.35	5.09	5.26	5.15	5.20	3.60	6.76
	2008	324	17.94	7.20	10.74	7.42	10.52	3.93	14.01
	2010	618	34.22	16.39	17.83	15.17	19.05	11.35	22.87
	2012	410	22.70	13.12	9.58	10.47	12.24	12.90	9.80
Total		1806	100%	46.35	53.65	45.02	54.98	34.33	65.67

Note: All of the weights in this table are ratios to the total samples. The sum of weights may be not equal to the total weights (value 1) because of rounding.

3.2. Selection of Variable Index

Firstly, we select the index of "health care and medical services utilization" in adult survey (residents survey over 18 years of age), and then pick out questions (No. M39) which are targeted for patients who had been suffered from sick or injury in the past four weeks, or long-term chronic diseases and acute diseases of "How much money have you spent on the treatment of this disease or injury" as an index which was focus on. This question can largely measure the prices of medical services those patients paid, so it is set to be the variable index to measure the price of medical services [15].

To calculate the "benchmark price" $\mu(x)$ in formula (2), the individual characteristic variables we chose are as follows:

(1) Severity of condition (Symptoms)

The question M25 in the survey of "severity of condition" is used to get information of the individual condition, which can present the urgency status for patients to consume medical service. The bigger this index's value is the more urgency is for medical service consumption.

(2) Expectation of health and life value (Endurance)

As the individuals' expectation of health and life value is different, their selections of medical services are not the same. The question M26 “what would you do if you feel uncomfortable?” can describe well the influence of this index. For residents with higher expectation of health and life value, they would be more likely to go for medical treatment, while residents with lower expectation of those would be likely to be treated by themselves or ignore those discomforts. Dummy variable method is applied in measuring how individuals with lower expectation of health and life value influence the price of medical services [16].

(3) Other variables

Factors which may affect the consumptive preferences of medical services are also taken into consideration, including age, gender, living region, marital status, education level, stable job, and medical insurance, etc. Finally, the geographic factor of different provinces and time factor of different years are under controlled. Here are some additional notes: 1) In the questionnaire, there are two indexes for identification of whether the individual came from rural or urban. One is "Household Register "in A8b1, the other one is questionnaire number T2 as "rural survey points / urban survey point”. We use the T2 index to measure the difference of medical services between rural and urban. 2) Both assignment method and dummy variable method are used in measuring the variable of education level. Under the assignment method, values of 1-6 are assigned sequentially to primary school, middle school, high school, secondary vocational school, college/undergraduate and postgraduate or above. Under the dummy variable method, two dummy variables were set as Education 1 and Education 2. In Education 1, value of one is assigned to the education level that is equal or less high school, value of 0 to the others; while in Education 2, value of one is assigned to the education levels of secondary vocational school and college/undergraduate, value of 0 to the others. The definite judgement methods and basic statistics for variables mentioned above are listed in Table 2.

TABLE 2. Descriptive Statistic of variables

Variable	Name of Variable	Questions in Questionnaire	Mean	Maximum	Minimum	Standard Deviation
1 Price of medical services						
How much money has been spent on treatment?	Price	M39	239.86	9999.0	1.0000	824.29
2 Patients' condition						
Urgency of Consumption (Value of 1, 2 and 3 was assigned sequentially to not severe, moderate, extreme severe.)	Symptoms	M25	1.6368	3.0000	1.0000	0.6547
3 Preference of health and life value						
What would you do if you feel uncomfortable? (1 treating by oneself or ignore the discomfort; 0 going to see the local medical personnel or doctors.)	Endurance	M26	0. 6346	1. 0000	0. 0000	0. 4817
4 Basic characteristics of individuals						
Age (Calculated and rounded according to solar calendar)	Age	U1	52. 01	93. 00	18. 00	15. 96
Sex (1 Male, 2 Female)	Sex	U1B	0. 4502	1. 0000	0. 0000	0. 4976
Rural or Urban (1 Urban, 2 Rural)	Urban	T2	0. 4635	1. 0000	0. 0000	0. 4988
Current marital status (1 Single, 2 Married, including divorced)	Married	A8	0. 0825	1. 0000	0. 0000	0. 1752
Education Level (by assignment method)	Education	A12	1. 8721	6. 0000	0. 0000	1. 1977
Have a job or not (1 Yes, 0 No)	Job	B2	0. 5199	1. 0000	0. 0000	0. 4997
Have medical insurance or not (1 Yes, 0 No)	Insurance	M1	0. 3433	1. 0000	0. 0000	0. 4749

Variable	Name of Variable	Questions in Questionnaire	Mean	Maximum	Minimum	Standard Deviation
5 control variable						
Geographic Area factors: Provinces in surveyed (9 provinces)	Area	T1	—	—	—	—
Time factors: Year of survey (7 years)	year	T7	—	—	—	—

4. Analysis of Empirical Results

In this part, based on the model specification and analysis of benchmark price influential factors, we decompose the total variance of model, measure the scale of surplus extracted by doctors and patients because of the difference of bargaining capacity and information asymmetry, and further analyse the different roles of each of those factors.

4.1. Model Specification and Benchmark Price Influential Factors

Based on the pricing mechanism of medical service market price under the information asymmetry and methods of quantity measurement, we analyse how information mastering level in both doctor and patient affects the medical service pricing. Bilateral stochastic frontier approach is applied in the measurement and its estimating regression result is shown in the Table 3.

TABLE 3. Bilateral stochastic frontier approach is applied in the measurement and its estimating regression result.

Dependent Variable	Ln price						
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Lnage	1. 018… (7. 603)	1.056… (7. 850)	1.007… (7. 755)	0.751… 5.337	0.796… 6.809	0.839… 7.334	0.598… 5.228
Symptoms	0. 867… (14.350)	0.897… (14.625)	0.849… (14.378)	0.770… (13.495)	0.771… (13.534)	0.755… (13.437)	0.746… (13.600)
Urban	0. 287… (3. 506)	0.296… (3. 54)	0.275… (3. 457)	0.260… (3. 279)	0. 260… (3. 267)	0. 245… (3. 133)	0. 205… (2. 693)
Sex	0. 010 (0. 127)	-0.001 (0. 011)	0.005… (0. 069)	0. 050 (0. 665)	—	—	—
Married	0. 160… (0. 964)	0. 185 (1. 121)	0. 187 (1.161)	-0.091 (-0.559)	—	—	—
Education	0.100… (2.809)	0.112… (3.064)	0.106… (3.074)	0.114… (3.267)	0.117… (3.401)	0.099… (2.886)	0.063· (1.815)
Job	—	—	—	-0.386□(-4.518)	-0.367… (-4.512)	-0.318. -3.988…	-0.293. -3.767…
Endurance	—	—	—	-0.783… (-9.948)	-0.780… (-9.925)	-0.819… (-10.032)	-1.026… (-11.804)
Insurance	—	—	—	0.207·· (2.438)	0.209·· (2.464)	0.142 (1. 641)	0.099 (1.155)
Constant	-2.025… (-3.726)	-2.302… (-4.221)	-2.908… (-5.490)	-1.174·· (-1.973)	-1.356… (-2.700)	-1.493… (-2.994)	-1.368… (-2.538)
Area dummies	—	—	—	—	—	Control	Control
Year dummies	—	—	—	—	—	—	Control
adj-R^2	0. 151	—	—	—	—	—	—
Log likelihood	—	-3490.63	-3462.30	-3405.83	-3406.17	-3358.07	-3311.41
LR(chi2)	—	—	56.64	169.59	168.92	265.11	358.42
P-value	—	—	0.000	0.000	0.000	0.000	0.000
N	1806	1806	1806	1806	1806	1806	1806

Note: ·, ··and … represent a significant difference at the level of 10%, 5% and 1% respectively. The value inside brackets is t values.

In Table 3, Model 1 uses OLS estimation method, Model 2-6 all use bilateral stochastic frontier approach, meanwhile, constrained condition $\ln\sigma_u = \ln\sigma_w = 0$ is attached to Model 2. In Model 3 and Model 4, the condition of factors' fitting improves significantly after considering the increasing jobs, medical insurance and expectancy value of medical care. Further, excluding the non-significant factors as Sex and Marital status in Model 5-7, with gradually increasing control of factors as Area and Year, the significant improvement is also seen in model fitting. The subsequent analysis may be performed mainly based on the variances and the distance measure results in the Model 7. The estimation results indicate: Factors of Age, Urban, Disease status, Education and Medical insurance have positive effect on the medical service price. People who are elder or living in city or with more urgent disease condition, higher education level or holding medical insurance are more likely to accept a higher medical service price. While people who have a lower expectancy on health and life value and stable job are more inclined to accept the relative lower medical service price.

4.2. Variance Decomposition: The Explanation Capacity of Information Level Measurement Model on Medical Service Pricing

Table 4 reports the analysis results of the effect of information mastering level factor. We have found out that the information mastering degree plays a very important role on medical service pricing. Doctors master information better than patients, which may cause that the doctor-patient information factor has a positive comprehensive effect on medical service pricing. $E(w-u)=\sigma_w-\sigma_u=0.6954$ indicates: Comparing with the benchmark price, bargaining will lead to a higher price. Meanwhile, Ln price cannot explain why partial overall variance $\sigma_v^2+\sigma_u^2+\sigma_w^2$ is 0.2395, 46.1% of which is contributed by doctor-patient information factor. In the total effect of information factor on medical service price, doctor is in an absolute dominate position (91.38%) versus patient (8.62%). This reveals that, the medical service pricing still depends on the doctor, though the patient holds certain information and bargaining capacity. In order to analyse the surplus and net surplus for specific doctor-patient extracted in the bargaining, we have made a further analysis on both doctor and patient by unilateral effects estimation.

TABLE 4. Effects analysis of the bargaining capacity factor on medical service price

Bargaining Mechanism	Meaning of Variable	Symbol	Measure coefficient
	Random errors	σ_v	1.1353
	Bargaining capacity of patient	σ_u	0.3083
	Bargaining capacity of doctor	σ_w	1.0037
Variance Decomposition	Random term total protocol	$\sigma_v^2+\sigma_u^2+\sigma_w^2$	2.3915
	The weight of the bargaining factor in the total variance	$(\sigma_u^2+\sigma_w^2)/(\sigma_v^2+\sigma_u^2+\sigma_w^2)$	46.1%
	The weight of bargaining capacity of patient	$\sigma_v^2/(\sigma_v^2+\sigma_w^2)$	8.62%
	The weight of bargaining capacity of doctor	$\sigma_w^2/(\sigma_u^2+\sigma_w^2)$	91.38%

4.3. Estimation of the Surplus Extracted by Patients and Doctors

(1) Result of overall samples estimation

We focus on the estimation of surplus, which is gained by both doctor and patient under information asymmetry in this part. That is $E(u/\xi)$ and $E(w/\xi)$, the relative estimation equation is (7a) and (7b) respectively. It's a ratio between surplus mentioned above and changes of benchmark price $\ln\hat{p}=x_i^{'}\hat{\beta}$. Table 5 reflects the estimation of the overall samples. In average, the surplus gained by doctor under information asymmetry may lead to medical service price 50.17% higher than benchmark price. While the surplus gained by patient can only lower the medical service prices by 23.56%. This relative great variance in the degree of information mastering between doctor and patient makes actual

price of medical services 26.61% higher than benchmark price. That is, because of the information asymmetry and bargaining capacity difference, patients have to pay 126.61 Yuan for a medical service that is worth 100 Yuan in the equal market.

The last three rows (Q1-Q3) present the distribution characteristic of doctor-patient surplus in detail, which reveals a great heterogeneity in the degree of information mastered by the doctor and patient. However, almost all of the patients are at a disadvantage in the doctor-patient bargaining. Specifically, the statistic result of Q1 indicates that there is about one out of four patients may gain a medical service price that less than 12% higher than benchmark price after bargaining with doctors. However, the statistic result of Q3 shows that there is another one out of four patients may gain a medical service price that up to 40% higher than benchmark price after bargaining with doctors.

TABLE 5. The total surplus gained by doctor and patient in the bargaining

Variable	Mean (%)	Standard deviation (%)	Q1 (%)	Q2 (%)	Q3 (%)
Doctor: $E\left(1-e^{-w}/\varepsilon\right)$	50.17	16.73	37.44	46.21	59.87
Patient: $\hat{E}\left(1-e^{-u}/\varepsilon\right)$	23.56	4.07	20.50	22.61	25.49
Net surplus: $\hat{E}\left(e^{-u}-e^{-w}/\varepsilon\right)$	26.61	20.16	11.95	23.60	39.37

Note: Q1=the 1st Quartile= 25th percentile; Q2=the 2nd Quartile= 50th percentile; Q3= the 3rd Quartile= 75th percentile.

Based on the frequency distribution of the surplus of the physicians and patients ,we got to know that the distribution of the surplus either of the physicians or of the patients demonstrated the feature of trailering rightwards; it means that only minority physician or patient have the bargaining power in powerful position absolutely. Based on the distribution feature of the net surplus in the net surplus frequency distribution, it is observed that not all patients are in the absolute disadvantage during the bargaining process. Our statistical analysis indicated that the net surplus of about not exceeding 10% of the patient is less than zero. It means that they grasped the information to the stronger degree and brought down the medical price in fact. Meanwhile it means that above 90% of the patients are compelled to accept the fact of the unfair prices. On the whole, our analysis showed that the physician have the stronger information superiority with respect to the patient in the physician-patient bargaining process in the medical service market, and implement the high-price strategy depending on this advantage during the process of deciding the medical service price at last [17].

Whether the relative information asymmetry degree of the physician-patient will improve along with the progress of the medical service institutional reform? For this purpose, we made the statistics for the distribution feature of the physician and patient net surplus yearly. It is found surprisingly that the physician and patient net surplus is almost around 26% from 2000 to 2006, and it means that the introduction of the competition mechanisms by marketization reform to the medical service haven't carried out the anticipated function of reversing the information disadvantage of the patients and strengthening the bargaining capacity of the patients effectively. It also indicated that the thinking of introducing the market factor purely into the medical service mechanism to promote the function of the competitive mechanism is unable to resolve the problem of the information asymmetry of the medical service market and the virtual high price effectively.

TABLE 6. Annual distribution characteristic of net surplus extracted by doctors and patients

Year	Mean (%)	Standard deviation (%)	Q1 (%)	Q2 (%)	Q3 (%)
2000	26.66	20.07	12.68	20.14	44.06
2002	26.89	20.87	14.12	22.33	40.47
2004	26.59	20.02	10.52	23.84	40.08
2006	25.36	16.99	12.70	22.88	37.46
2008	26.18	19.02	11.67	24.88	40.70
2010	27.45	22.19	11.18	23.58	39.83
2012	26.18	19.13	12.08	23.15	37.30

TABLE 7. The role of urban-rural factor on surplus extracted by doctors and patients

Variable	Mean (%)	Standard deviation (%)	Q1 (%)	Q2 (%)	Q3 (%)
Rural area (urban=0)					
Doctor: $\hat{\hat{E}}\left(1-e^{-w}/\varepsilon\right)$	50.10	16.65	37.43	46.03	60.25
Patient: $\hat{\hat{E}}\left(1-e^{-u}/\varepsilon\right)$	23.53	3.85	20.47	22.65	25.49
Net surplus: $\hat{\hat{E}}\left(e^{-u}-e^{-w}/\varepsilon\right)$	26.57	19.97	11.94	23.38	39.79
Urban (urban=1)					
Doctor: $\hat{\hat{E}}\left(1-e^{-w}/\varepsilon\right)$	50.24	16.82	37.69	46.30	59.49
Patient: $\hat{\hat{E}}\left(1-e^{-u}/\varepsilon\right)$	23.60	4.31	20.54	22.59	25.37
Net surplus: $\hat{\hat{E}}\left(e^{-u}-e^{-w}/\varepsilon\right)$	26.64	20.38	12.32	23.71	38.95

(2) The Role of Individual Characteristic on degree of mastering information

We have discovered the great heterogeneity in the degree of mastering information between doctors and patients. For understanding its root cause, we have made a further analysis on the distribution characteristic of surplus extracted by doctors and patients, by group statistical analysis, from aspects of urban and rural variance, medical insurance stations, degree of job stability and age, as well as education background.

The statistic results in the Table 7 show: Comparing with the rural doctors and patients, both doctors and patients in urban have stronger capacity in mastering information, but there is no obvious difference in the net surplus of doctor-patient bargaining between urban and rural. Almost all of the patients are forced to accept an equilibrium price, which is about 26.6% higher than the benchmark price. Moreover, patients from different quartiles face different risen ranges of the price.

Table 8 reflects the difference in surplus and final net surplus extracted by doctors and patients from different education backgrounds. The results show that patients with universities or higher education background from different quartiles are all faced with the lowest risen range of the price. On one hand, this kind of patients can obtained relative more surpluses through the information mastering and their bargaining capability. On the other hand, doctors "plundered" relative fewer surpluses from this kind of patients than other patients. It means that patients can "self-preserve" in bargaining with their knowledge, which can help them mastering more information. Meanwhile, doctors implement a strategy with relatively lower medical services price when they are faced with highly educated patients [18].

In the same way, we have conducted the group comparative analysis on those heterogeneity characteristic factors, such as whether the patient has medical insurance, whether the patient has a stable job and the patient's age, etc. All of the results show that, patients are forced to accept an equilibrium price without exception, but patients from different quartiles face different risen ranges of the price. Thus, to a certain extent, comparing to the information advantage of doctors, the patients' heterogeneity capacity has limited impact on the medical service pricing.

5. Conclusions and Policy Recommendations

This paper has built a measurement model for the degree of information asymmetry in medical service market. It bases on the micro individual data in the "China Health and Nutritional Survey (CHNS)", measures empirically on the degree of mastering information of both doctor and patient and its total effect on the medical service price. The results of this empirical measurement demonstrate that:

(1) The factor of information mastered by doctors and patients has an important impact on the medical service pricing. In addition, doctors master more information and have stronger bargaining capacity than patients. The impact of information asymmetry on the final medical service price is positive (value 0.6954), contributing to the formation of a higher price than benchmark price.

(2) Through the unilateral effects analysis of the overall samples for both doctors and patients, we know that: In average, depending on the information they owned, doctors may elevate the medical service price to 50.17% higher than benchmark price, while patients can just lower 23.56% of it. Doctors versus

patients, medical services price will be 26.61% higher than benchmark price. The further quartile analysis indicates that: Almost all of patients will be forced to accept a higher price than the benchmark price. However, the risen ranges of the price may be different in patients from different quartiles.

(3) In the period from 2000 to 2012, physicians with their information advantage "controlled" determination of price, which was approximately 26% higher than the fair benchmark price. That means China's health care reform did not effectively help ease the problem of information asymmetry in medical services market. What's more, the bargaining price is about 26% higher than the fair benchmark price at average. Thus, the health care reform cannot completely resolve the issue of "difficulty of receiving medical service and high cost of getting medical treatment".

TABLE 8. Effect of education background on surplus extracted by both doctors and patients

Variable	Mean (%)	Standard Deviation (%)	Q1 (%)	Q2 (%)	Q3 (%)
Education≤1	Primary and illiteracy				
Doctor: $\hat{E}\left(1-e^{-w}/\varepsilon\right)$	50.42	17.38	36.71	45.52	61.42
Patient: $\hat{E}\left(1-e^{-u}/\varepsilon\right)$	23.63	4.10	20.35	22.78	25.83
Net surplus: $\hat{E}\left(e^{-u}-e^{-w}/\varepsilon\right)$	26.79	20.89	10.87	22.74	41.07
1<Education≤4	Middle school				
Doctor: $\hat{E}\left(1-e^{-w}/\varepsilon\right)$	49.94	15.71	38.19	46.80	58.44
Patient: $\hat{E}\left(1-e^{-u}/\varepsilon\right)$	23.38	3.72	20.65	22.48	25.15
Net surplus: $\hat{E}\left(e^{-u}-e^{-w}/\varepsilon\right)$	26.57	18.86	13.04	24.32	37.79
4<Education≤6	University or higher				
Doctor: $\hat{E}\left(1-e^{-w}/\varepsilon\right)$	49.57	17.64	38.89	45.20	58.75
Patient: $\hat{E}\left(1-e^{-u}/\varepsilon\right)$	24.21	5.64	20.62	22.86	24.86
Net surplus: $\hat{E}\left(e^{-u}-e^{-w}/\varepsilon\right)$	25.36	22.08	14.03	22.34	38.13

(4) After further analysis of the effect of variance factors on the final pricing, including urban-rural factor, medical insurance, job, age and education background, we can make such a conclusion that almost all of patients are forced to accept a higher medical care price in different degrees and doctors can run a discriminatory pricing strategy successfully.

The conclusions mentioned above in this paper are all pointing to one thing that: Chinese health care reform can neither relieve the information asymmetry of medical care market effectively, nor resolve the issue of "difficulty of receiving medical service and high cost of getting medical treatment". In other words, the over emphasis on such a reform approach that strengthens the regulating effect of market mechanism on medical care market via introducing marketing competition is worth being reflected on whether it is fit for China. To resolve the outstanding issues those are widespread in china, such as unreasonable higher medical care price and tension doctor-patient relationship, the following two strategies are required: 1. Return the nature of public welfare to health care; 2. More government participation to regulate price, supervise market and correct the marketing from outside.

References

1. Acemoglu, D., Shimer, R. (2000) Wage and Technology Dispersion, *Review of Economics Studies*, 67, 585-607.
2. Akin, J.S., Griffin, C.C., Guilkey, D.K., Popkin, B.M. (1986) The Demand for Primary Health Care Services in the Bicol Regio of the Philippines, *Economic Development and Cultural Change*, 34(4), 755-782.
3. Alger, I., Salanie, F. (2006) A Theory of Fraud and Over-treatment in Experts Markets, *Journal of Economics and Manageme Strategy*, 15(4), 853-881.
4. Flinn, C. (2006) Minimum Wage Effects on Labour Market Outcomes under Search Matching and

Endogenous, *Contact Rates Econometrica*, 74, 1013-1062.
5. Gupta, S., Verhoeven, M., Tiongson, E.R. (2002) The Effectiveness of Government Spending on Education and Health Care Developing and Transition Economics, *European Journal of Political Economy*, 18(4), 717-737.
6. Kumbhakar, S.C., Parmeter, C.F. (2009) The Effects of Match Uncertainty and Bargaining on Labour Market Outcomes: Eviden from Firm and Worker Specific Estimates, *Journal of Productivity Analysis*, 31(1), 1-14.
7. Kumbhakar, S.C., Lovell, C.A.K. (2000) *Stochastic Frontier Analysis*, Cambridge University Press, New York, 90.
8. Gaynor, M., Polachek, S.W. (1994) Measuring Information in the Market: An Application to Physician Services, *Southern Economic Journal*, 60(4), 815-831.
9. Feldstein, M.S. (1970) The Rising Price of Physicians' Services, *Review of Economics and Statistics*, 52, 121-133.
10. Rice, T. (1983) The Impact of Changing Medicare Reimbursement Rates on Physician-induced Demand, *Medical Care*, 21, 803-815.
11. Polachek, S.W., Yoon, B.J. (1987) A Two- Tiered Earnings Frontier Estimation of Employer and Employee Information in the Labor Market, *Review of Economics and Statistics*, 69(2), 296-302.
12. Polachek, S.W., Yoon, B.J. (1996) Panel Estimates of a Two- Tiered Earnings Frontier, *Journal of Applied Econometrics*, 11(2), 169-178.
13. Osbourne, M.J., Rubinstein, A. (1997) Bargaining and Markets (Chapter 5), Academic Press, San Diego. Phelps C. E.
14. Health Economics, New York, NY: Addison-Wesley Educational Publishers Inc. Winand Emons, 1997.
15. Credence Goods Monopolists, Berkeley Olin program in Law & Economics, Working Paper Series, 1990.
16. Wolinsky, A. (1993) Competition in a Market for Informed Experts' Services, *Rand Journal of Economics*, 24, 380-398.
17. Yip, W.C. (1998) Physician Response to Medicare Fee Reductions：Changes in the Volume of Coronary Artery Bypass Graft (CABG) Surgeries in the Medicare and Private Sectors, *Journal of health Economics*, 17, 675-699.
18. Jiangnan, C. (2012) Information insufficience is the pain of medical care. *Chinese social sciences today*, 7(6), 1-2.

Received on the 17th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 289-295
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

SIMULATION AND OPTIMIZATION OF ENTERPRISE TRANSFORMATION STRATEGY IN DIFFERENT PHASES OF BUSINESS CIRCLE

Lei Hua[1,*,2], Chunjie Yu[2], Wenping Wang[1]

[1]School of Economics and Management, Southeast University, Nanjing 211189, China
[2]School of Business, Linyi University, Linyi 276000, China
[*]Corresponding author E-mail address: 230109338@seu.edu.cn

This paper establishes a simulation model of enterprise transformation to optimize the transformation strategy of enterprise in different phases of business circle. The result shows that in the growth phase of business circle, enterprises should reduce their transformation effort, and concentrate on elevating their present capacities. In the decline phase of business circle, enterprise should increase their transformation effort, in order to find more promising strategic options. Further analysis indicates three reasons of the results. First, present capacity is more important in growth phase of business circle while finding better opportunity is more important in decline phase of business circle. Second, enterprise's capacity declines as the degree of transformation rises. Third, the change of economic trend also erodes enterprise's capacity.

Keywords: business circle; enterprise transformation; capability; strategy

1. Introduction

Rapid changes of customer demands and new technologies make enterprise transformation a lasting theme of enterprise management. It is not easy for enterprises to transform successfully, because the process of transformation is full of risks and challenges. Numerous enterprises that fail to transform are sifted out by the rises and falls of the economy, including some leaders of industries. At the same time, there are some enterprises transformed successfully, and some enterprises have gone through many times of transformations with the constantly changing economy. So the problem of how could enterprises transform successfully becomes a great concern in both theory and practice. Therefore, the area of enterprise transformation is focused on by many researchers.

Levy and Merry [1] indicate that enterprise transformation means great changes of organizational phase, including the purposes, goals, structures, and enterprise culture, are made by enterprise to adapt to the change of external environmental change. Shaheen [2] defines enterprise transformation as the changes on organizational value, shape, altitude, skill, and manner, so as to response quickly to environmental change. Klein [3] defines enterprise transformation as new operation mode of enterprise, which is designed by shifting the goal or strategy of enterprise, so as to survive or break through some bottleneck when the external environment is changed. As to the question of what should enterprise do to deal with the environmental change and transform successfully, Prahalad and Hamel [4] put forward the core competence theory. They indicate that enterprise is a system of competences, and the core competence will decide the competitive advantage and performance of enterprises. The long term competitive advantage of enterprise comes from lower cost and faster innovations, and this core competence is compatible, valuable, and hard to imitate. Teece et.al [5] put forward the dynamic capability theory. They indicate that in the changing external environment, successful competition comes from constantly developing and rebuilding the system of enterprise resources, and the capability to process these special resources is very important to enterprises.

As we can see from the existing literatures, enterprise transformation is the change of strategies made by enterprises to adapt to external environmental change. The related researches on microcosmic level are mainly based on enterprise capacities, and the progresses of the theories are made from static to dynamic. It is widely accepted that enterprises should adjust their develop strategy to chase the changes of environment to make their transformation successful. However, the problem is that the environment is complex. In different phases of business circle, the economy could be booming or declining, and the changing rates of the economic trends are also could be different. The phase of business circle will influence the strategy of enterprise transformation inevitably. So, in different phases of business circle,

how should enterprises decide their transformation strategy to optimize their interests? The answer is still not given.

In this paper, we study this problem by simulation method. This paper is organized as follows. Section 2 discusses the relations between business circle and enterprise transformation, and establishes the simulation model of enterprise transformation with dynamic environment settings that are corresponding to different phases of business circle. Section 3 provides the numerical analysis of the model and analyses the optimistic transformation strategies in different phases of business circle. Section 4 provides further analyses on reasons of the result. Section 5 gives the robust test of the results of this model. Finally, conclusions are formulated by section 6.

2. The Model

The intuition of the enterprise transformation model of this paper comes from an analogy with the multi-armed bandit model, which is widely developed and used in economics and statistics [6-8], computer science [9-11], and organizational learning [12, 13].

Assume an enterprise facing with N possible strategy choices. Each choice represents a different developing direction in the future, and could bring different incomes to the enterprise, which is decided both by the latent value of the specific choice and the possibility of the enterprise to carry it out successfully. The latent value of each choice is decided by the external environment faced by the enterprise, such as the condition of the whole economy, the trends of technology innovation, and competitions of the market. The possibility to success of each choice is decided by the capability of the enterprise on this specific direction. The higher is the capability, the more likely the choice is carried out successfully. The transformation strategy is embodied by which specific choice the enterprise will select.

Based on this background, we describe the transformation strategy as below. If the enterprise makes its choice totally depend on its own capability, and select the choice with the highest capacity, then it is taking non-transform strategy. On the contrary, if the enterprise makes its choice totally depend on external environment, and select the choice with the highest latent value, then it is taking absolute transform strategy. As it should be, the enterprise could also balance between these two extreme strategies, and taking different degrees of transform strategies, and we introduce the details of these settings below.

To model these discussions above, we build an enterprise transformation model with a sequential selection problem. At time t, an enterprise need to pick up one strategy choice between N of them. If the enterprise selects option i and carries it out successfully, it will get a positive income $p_{i,t}$; if it fails to carry it out, it will get zero income. So the external environment of time t could be denoted by the payment vector of these N choices $P_t = \left[p_{1,t},...p_{N,t}\right]$. In typical bandit models, environment is not changing with time, i.e., for every t, $P_t = P$. We relax this assumption here to make the model be able to describe the turbulence of environment. In accordance with the definitions of environmental turbulence in existing literatures [14-17], we define the change of environment as a shock of the payment vector. On every time t, if the shock has the same probability distribution, which means the payment vector before and after the shock both have the same mean value, then the economy is stable. In this model, we define the growth rate of economy IR as the change of mean value of the environment vector P_t in each time step. So the value of IR could represents the phase of business circle. If $IR > 0$, the economy is booming and the enterprise is in the growth phase of business circle; if $IR < 0$, the economy is falling and the enterprise is in the declining phase of the business circle; $IR = 0$ represents the situation of stable economy.

The capability on each strategic choice is defined by the probability to successfully carry out this choice by the enterprise. Set $q_{i,t}$ as the capability of the enterprise corresponding to option i at time t, and the capacity vector of the enterprise could be denoted by $Q_t = \left[q_{1,t},...,q_{N,t}\right]$, $0 \le q_{i,t} \le 1$. At the initial time $t = 0$, the enterprise owns an initial capacity vector $Q_t = \left[q_{1,0},...,q_{N,0}\right]$, and the payment vector of the environment is $P_0 = \left[p_{1,0},...,p_{N,0}\right]$. At every time moment, the enterprise selects one of its choices between N options. If option i is selected, the enterprise will get an income r_t form the environment, of which the expected value is

$$E(r_t) = p_{i,t} \times q_{i,t} \tag{1}$$

That means the expected income is decided both by external environment and the capacity of the enterprise. This income then is added to the total asset of the enterprise, i.e.,

$$S_t = \sum_{j=0}^{t} r_j \tag{2}$$

At the same time, if the enterprise selects option i at time t, it will improve its experience and knowledge on this direction, and the corresponding capacity $q_{i,t}$ will increase for a certain quantity. We use the update algorithm introduced by Bush and Mosterller [18], let

$$q_{i,t+1} = q_{i,t} + \alpha \times (1 - q_{i,t}) \tag{3}$$

and α denotes the rate of capacity growth. For the other non-selected options, the corresponding capacity will degenerate for a certain quantity. Take option j for example, we have

$$q_{j,t+1} = q_{j,t} - \beta \times q_{i,t} \tag{4}$$

where β denotes the rate of capability degeneration.

Facing with the vectors of external environment $P_t = [p_{1,t},...p_{N,t}]$ and capacity $Q_t = [q_{1,t},...,q_{N,t}]$, there are different transformation strategies for the enterprise to choose. First, as Hannan hand Freeman [19] introduced in their organizational biological study, enterprises are limited by some structural and conventional ingredients, and so they will have inertia in organizational changes. Subjected to this inertia, the enterprise is apt to choose the option with highest corresponding capacity, i.e., potion i with $\max(q_{1,t},...,q_{N,t})$. This choice is corresponding to non-transform strategy of the enterprise. It will let the enterprise to further increase its existing preponderant capacity, but it also risk the enterprise to miss other more promising options. On the other hand, some adventurous enterprises are also apt to chasing optimistic perspectives. Under this circumstance, the enterprise will choose the option with best environmental condition, i.e., option i with $\max(p_{1,t},...,p_{N,t})$. This choice is corresponding to absolute transform strategy. It will the enterprise a more promising perspective, but it is hard for the enterprise to achieve this goal, because the enterprise may not be good at this option. Therefore, the risk of failure is high, and the short-term income will be badly influenced. Mostly, enterprises will try to balance between these two extreme strategies, and make their choice on consideration of both external environment and corresponding capabilities. Therefore, we let $c \in [0,1]$ to denote the degree of enterprise transformation. If $c = 0$, the enterprise is adopting non-transform strategy; if $c = 1$, the enterprise is adopting the absolute transform strategy; as c increases from 0 to 1, the enterprise is more and more apt to make its choice depending on the external environment, so embodies the increasing transformation degree. Whatever strategy the enterprise takes, the purpose of the enterprise is to maximize its asset S_t.

To run this simulation process, we could investigate how the phases of business circle affect the enterprise's total asset S_t under different degrees of transformation, so as to find out the optimistic transformation strategy in different phases of business circle.

3. Simulation Results

In the numerical experiment, the enterprise faces with 10 different strategic choices. So the payment vector decided by the external environment is $P_t = [p_{1,t},...p_{10,t}]$, and the capability vector of the enterprise is $Q_t = [q_{1,t},...,q_{10,t}]$. At time 0, $p_{i,0}$ and $q_{i,0}$ are initialized by a uniform distribution from 0 to 1. At every simulation time step, the external environment is changed by a given shock. At time 0, the

shock is defined by reset the payment of each option $q_{i,t}$ by the uniform distribution mentioned above, of which the mean value is $M_0 = 0.5$. At time t, the level of the shock is decided by the increment rate of economic growth IR, and the payment of each option $q_{i,t}$ is reset by a uniform distribution with the mean value of

$$M_t = M_{t-1} \times IR \tag{5}$$

This model of enterprise transformation is run by 500 times to get generate margins of error sufficient to support statistical tests. Without losing generality, the rates of capacity growth (α) and decline (β) are all set as 0.01; the initial asset of the enterprise is $S_t = 0$; and the simulation process lasts for 600 periods.

As running this model, me change the increment rate of economic growth IR from $IR = -20\%$ to $IR = 20\%$ and observe the total asset of the enterprise S_t. The transformation degree c that leads to largest total asset under certain increment rate of the shock IR, which is denoted by c^*, is defined as the optimal transformation strategy under this increment rate of the shock.

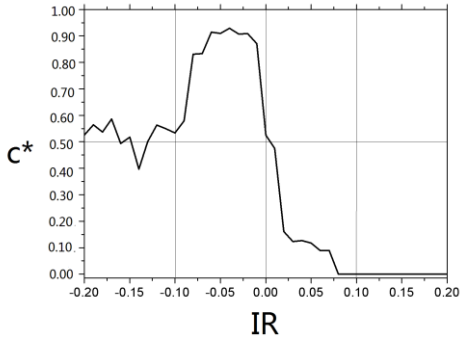


Figure 1. Optimal Transformation Degrees of Different Economic Growth Rates

As the result of this simulation process, figure 1 shows the relationship between IR and c^*. The value of c^* is obtained by fitting the data of S_t and c by cubic polynomial under each IR value. As we can see from figure 1, when the economy is stable ($IR = 0$), the value of c^* is near 0.5, which means that the optimal transformation degree of the enterprise is of medium level. In the decline phase of the business circle ($IR < 0$), when the economy is declining slowly ($IR > -0.1$), the value of c^* rises up to near 0.9, of which means that the optimal transformation degree of the enterprise is high level, and when the economy is declining fast ($IR < -0.1$), the value of c^* decline to near 0.5 again, of which means that the optimal transformation degree of the enterprise is medium level. In the growth phase of the business circle ($IR > 0$), when the economy is growing slowly ($IR < 0.1$), the value of c^* declines as IR rises, of which means that the optimal transformation degree of the enterprise declining with the rise of economic growing speed, and when the economy is growing fast ($IR > 0.1$), the value of c^* decline to zero, of which means that the optimal transformation strategy of the enterprise is not to transform.

4. Further Analyses

To explain the reasons of the result, we pick up five different values of transformation degree, which are $c = 0;\ 0.25;\ 0.5;\ 0.75;\ 1$, and investigated how the different phases of business circle affects the enterprise's capacity of the chosen option with different c values. As at every simulation step, the enterprise has to choose one option from all 10 options and the capacity of the chosen option decided the

possibility of carrying out this option successfully, we define present capability q as the mean value of the capabilities of chosen options over all time steps.

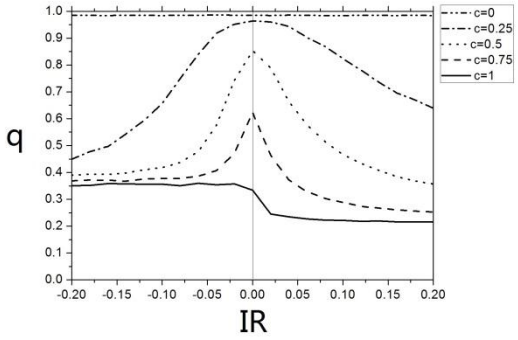


Figure 2. Present Capacities of Different Economic Growth Rates

As the setting of equation (1), when the economy is growing, no matter what strategic option the enterprise chooses, the payment value $p_{i,t}$ will be rising with time, so the possibility of carry it out successfully (present capacity q) becomes decisive. Figure 2 shows the relations between present capacity and economic growth rate. From figure 2, we can see two law of present capacity. First, present capacity declines as the degree of transformation rises. Second, economic growth and decline could erode the present capacity, and this effect rises with the absolute value of growth rate. Therefore, when the growth rate is positive and rising, the enterprise should lessen its transformation degree to mitigate the erosion effect on its present capacity to get more income. In addition, when the economy is growing very fast ($IR > 0.1$), the enterprise should not transform.

On the other hand, when the economy is declining ($IR < 0$), optimal degree of transformation rises near 0.9 and falls to near 0.5 subsequently as the rate of decline rises. First, when the economy declines slowly ($IR > -0.1$), whatever strategic the enterprise chooses, the payment value $p_{i,t}$ will be declining with time, so the possibility of carry it out successfully (present capacity q) becomes less important and finding better payment $p_{i,t}$ becomes more important. Therefore, the decline of the economy will lead to the rising of optimal transform degree. Second, when the economy declines very fast ($IR < -0.1$), the payment value of any strategic option $p_{i,t}$ will be declining very fast. This will make the income of the enterprise r_t declines quickly. We can see it from figure 3. If this happens, the asset of the enterprise will be decided by the income of the first few time steps. As the enterprise could get largest short term income if they adopt the transform strategy of medium level, the optimal transform degree falls back to near 0.5.

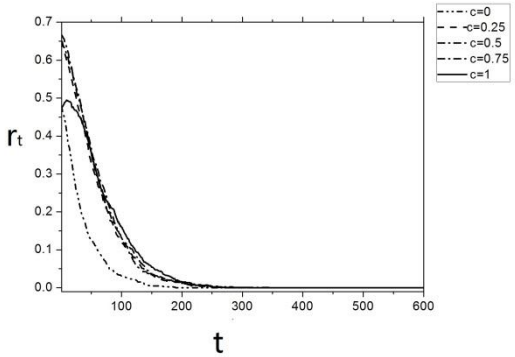


Figure 3. Income of Different Transform Degrees as Economic Growth Rate is -20%

5. Sensitivity Analyses

We test the sensitivity of our results in three aspects.

First, we adjust all the external variables of the model. We also test other probability distributions of external environment vector and enterprise capacity vector, such as using beta distribution and normal distribution other than uniform distribution.

Second, we test incidental turbulence of environment. Especially, we test all levels of turbulence happened in the first thirds of the total 600 simulation periods. We also test the turbulences happened in time widows of 50 simulation steps, including this window of turbulence happens in early and late stages of the whole process.

Third, we test the model with competitions between enterprises. For the 500 replications of the enterprise, we sift out a certain ratio of which the total asset is fallen behind by the others.

The model concludes in similar results with all these changes, showing robustness on these three aspects.

6. Conclusions

In different phases of business circle, how should enterprises adjust their transformation strategy to optimize their interests? In the real practice of management, many enterprises are trying to imply different degrees of transformations to cope with the changes of economic growth rate. In theory, the answer is still not clear. To answer this question, we try to investigate the mechanism of enterprise transformation from the angle of enterprise dynamic capability, establish a simulation model of enterprise transformation based on multi-armed bandit model to study and find out the optimal enterprise transformation degree according to different phases of economic circle.

The result of our model shows that when the economy is stable, the optimal transformation degree of the enterprise is of medium level. In the decline phase of the business circle, if the economy is declining slowly the optimal transformation degree of the enterprise is of high level, and if the economy is declining fast, the optimal transformation degree of the enterprise is of medium level. In the growth phase of the business circle, when the economy is growing slowly, the optimal transformation degree of the enterprise declines as the economic growing speed rises, and if the economy is growing very fast, the optimal transformation strategy of the enterprise is not to transform. Further analysis of the model shows the reason of these results. First, in the growth phase of business circle, enterprise's capacity to carry out its strategic option is more important to increase its asset; but in the decline phase of business circle, finding more prospective strategic option is more important to increase its asset. Second, enterprise's capacity declines as the degree of transformation rises. Third, economic growth and decline could erode enterprise's capacity, and this effect rises with the absolute value of growth rate.

The enterprise transformation model of this paper could be improved in many ways in further studies. For example, more detailed competition mechanisms could be introduced in the model. In our model, enterprise responds to the environmental change independently, so it is competing on its own performance. In reality, the competition is more complex. The probability of the failure of an enterprise is often related to other enterprises. On the other hand, the payment of an enterprise is also could be competitive income, which means the payment is decided by both the strategies of its own and other enterprises. These improvements make the enterprise much harder to recognize the consequence of its choice, so as to influence the evolving manner and consequences of their choices. Besides, some details of this model could also be adjusted, including different rates of capacity change; evolving transformation degrees; interdependence of possible options, and so on. These possibilities give chance to deeper study of enterprise transformation issue, but they will not change the conclusions of this paper.

Acknowledgment

This research was supported by National Social Science Foundation of China (No. 12&ZD207), National Natural Science Foundation of China (70973017, 71172044, 71273047), Doctoral Program Foundation of Education Ministry of China (20120092110039), Philosophy and Social Science Foundation of Jiangsu Province of China (2011ZDAXM009, 09JD018), Ministry of Education in Humanities and Social Sciences Planning Foundation (12YJC63003), Education Department of Henan Province Science and Technology Research Projects (NO.13A630114), Henan Province Government Decision-making Research Project (NO.2013B389), University Humanities and Social Science Planning Projects in Shandong Province (J11WG59).

References

1. Levy, A., Merry, U. (1986) *Organizational Transformation: Approaches Strategies, and Theories*, Praeger, New York.
2. Shaheen, G.t. (1994) Approach to transformation, *Chief Executive*, 3, 2-5.
3. Klein, M.M. (1996) Tip for Aspiring Reengineers, *Planning Review*, 24(1), 40-41.
4. Prahalad, C.K., Hamel, G. (1990) The core competence of the corporation, *Harvard Business Review*, 68(3), 79-91.
5. Teece, D.J., Pisano, G., Shuen, A. (1997) Dynamic capabilities and strategic management, *Strategic Management Journal*, 18(7), 509-533.
6. Robbins, H. (1985) Some aspects of the sequential design of experiments, *Herbert Robbins Selected Papers*, 58(5), 527- 535.
7. Gittins, C. (1979) Bandit processes and dynamic allocation indices, *Journal of the Royal Statistical Society*, Series B, 41(2), 148-177.
8. Berry, D.A., Fristed, T.B. (1985) *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall, London.
9. Holland, J. (1975) *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control & Artificial Intelligence*, University of Michigan Press, Ann Arbor.
10. Sutton, R.S., Barto, A.G. (1998) *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA.
11. Denrell, J., March, J. (2001) Adaptation as information restriction: The hot stove effect, *Organization Science*, 12(5), 523–538.
12. March, J.G. (1996) Learning to be risk averse, *Psychological Review*, 103(2), 309-319.
13. Posen, H.E., Levinthal, D.A. (2012) Chasing a moving target: exploitation and exploration in dynamic environments, *Management Science*, 58(3), 587-601.
14. Dess, G., Beard, D. (1984) Dimensions of organizational task environments, *Administrative Science Quarterly*, 29(1), 52-73.
15. Miller, D. (1987) The structural and environmental correlates of business strategy, *Strategic Management Journal*, 8(1), 55-76.
16. Wholey, D., Brittain, J. (1989) Characterizing environmental variation, *Academy of Management Journal*, 32(4), 867-882.
17. Baum, J., Wally, S. (2003) Strategic decision speed and firm performance, *Strategic Management Journal*, 24(11), 1107-1129.
18. Bush, R., Mosteller, F. (2001) *Stochastic Models for Learning*, John Wiley & Sons. New York.
19. Hannan, M., Freeman, J. (1984) Structural inertia and organizational change, *American sociological review*, 49(2), 149-164.

Received on the 17th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 296-300
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

PRIORITIZATION OF INTELLECTUAL CAPITAL STRATEGIES OF TAIWAN-FUNDED FABLESS INTEGRATED CIRCUIT DESIGN HOUSES IN CHINA

Meng-Chun Kao

Department of Business Administration, Yuanpei University of Science and Technology
No. 306, Yuanpei St., Hsin Chu 30015, Taiwan, R.O.C.
Phone: 886-3-538-1183. E-mail: ruth@mail.ypu.edu.tw

This study applies the fuzzy analytic hierarchy process (FAHP) to prioritize major construction of intellectual capital and development strategies of Taiwan-funded fabless integrated circuit design (FICD) houses in China. The empirical results indicate that human capital is recognized as the core construction of intellectual capital for FICD houses. In addition, these firms emphasize on building-up of structural capital. Accordingly, the most important development strategies of intellectual capital aim at human capital and structural capital. Finally, the findings and implications are discussed.

Keywords: intellectual capital, development strategy, fuzzy analytic hierarchy process, Taiwan-funded FICD houses

1. Introduction

With the advent of knowledge economy, intellectual capital has become the main source of competitiveness and the preeminent resource for wealth creation. This phenomenon is more obvious for knowledge-based enterprises [1, 2, 3]. Numerous studies have identified intellectual capital as the value driver for an enterprise [4, 5, 6, 7, 8, 10]. Intellectual capital thus can be considered as the core competence in business.

As the semiconductor industry becomes increasingly global, China has emerged as its new centre of gravity as integrated device manufacturers, fabless integrated circuit design houses (FICD), foundries, assembly and testing companies shift part of their operations to China. The semiconductor industry has become one of the blooming industries in China. Topology Research Institute (2008) points out Chinese foundry market is 590 billion in 2006, contributes to one third of the global market value. The FICD houses play an important role in the semiconductor industry. Because of the perceived market drive, local talent availability in China and policy incentives including preferential tax treatment and infrastructure arrangements, the majority of Taiwanese FICD houses have invested to China despite restrictive official regulations in Taiwan. Meanwhile, FICD is one of knowledge-intensive industries, for which intellectual capital is important on value creation. Future winners from competing in the global market of FICD would rely largely on their unique core competence including intellectual assets built up along with their development strategies. Therefore, it is the key issue how to localize intellectual capital development strategy of Taiwanese FICD houses in China.

Even though the concept of intellectual capital is easy to understand, there are bottlenecks in its practical application. Prior studies [11, 12, 13, 14, 15, 16, 17] have focused on identification, measurement of intellectual capital and the relationship between intellectual capital and firm value. Few prior empirical studies to my knowledge have addressed on intellectual capital development strategies taken by various companies, particularly for FICD industry in emerging economics. The process of developing strategies is complex and the empirical data concerning intellectual capital development strategies is not available or very hard to be extracted. The survey is conducted by the R&D top managers of Taiwanese FICD houses in China. A questionnaire, identified from previous literature and interviews with experts, is designed using Delphi Technique. In addition, Decision-makers prefer natural language expressions rather than explicit numerical values in assessing the factors of intellectual capital strategy. This study applies the fuzzy analytic hierarchy process to assess intellectual capital development strategies that would finally impact to and result in unique core competence in these companies. The advantage of fuzzy analytic hierarchy process can deal quantitatively with imprecision or uncertainty and at obtaining a fuzzy prioritization of intellectual capital development strategies from this point of view that will close this gap considerably. The remainder of this study is organized as follows. Section 2 then proposes methodology and the study setting. Next, section 3 presents empirical results. Finally, section 4 is conclusion and implications.

2. Methodology and Research Design

Saaty [18] proposes AHP, which is a decision method for decomposing a complex problem into a multilevel hierarchical structure including objectives, criteria, sub-criteria and alternatives. AHP incorporates the assessments of all decision makers to yield a final decision via pair-wise comparisons of the alternatives. The decision making process depends on experience, imagination and knowledge to structure a problem and on logic and intuition to provide judgments for solving the problem. Judgments about the available data regarding a particular problem can be made at each level of the hierarchy by comparing each element in a pair-wise square matrix. Priorities are derived to determine the best decision regarding the overall goal. However, precise data is not available or hard to capture. In addition, decision makers could not express their preferences exactly. Beskese et al. [19] propose that fuzzy set theory is primarily concerned with vagueness in human thoughts and perceptions to deal quantitatively with imprecision or uncertainty. In this paper, fuzzy AHP will be performed in the prioritization of intellectual capital developmental strategies to generate more precise information about the preferences of decision makers.

In the procedure of questionnaire development, numerous depth interviews and Delphi survey are implied to obtain validity of research instrument. First, depth interviews are conducted with two senior top managers of Taiwanese FICD houses in China who are responsible for making corporate strategies and proficient at projects execution in their companies. The questionnaire is designed according to the related literature review of intellectual capital and the vitals of IC development process. After compiling the preliminary questionnaire, modifications are made based on discussions of the questionnaire content with top managers, and the Delphi survey is then taken. Convergence between the two rounds has been checked to see if consensus among five experts is achieved and the validity of final questionnaire could be obtained. Figure 1 shows the study framework.

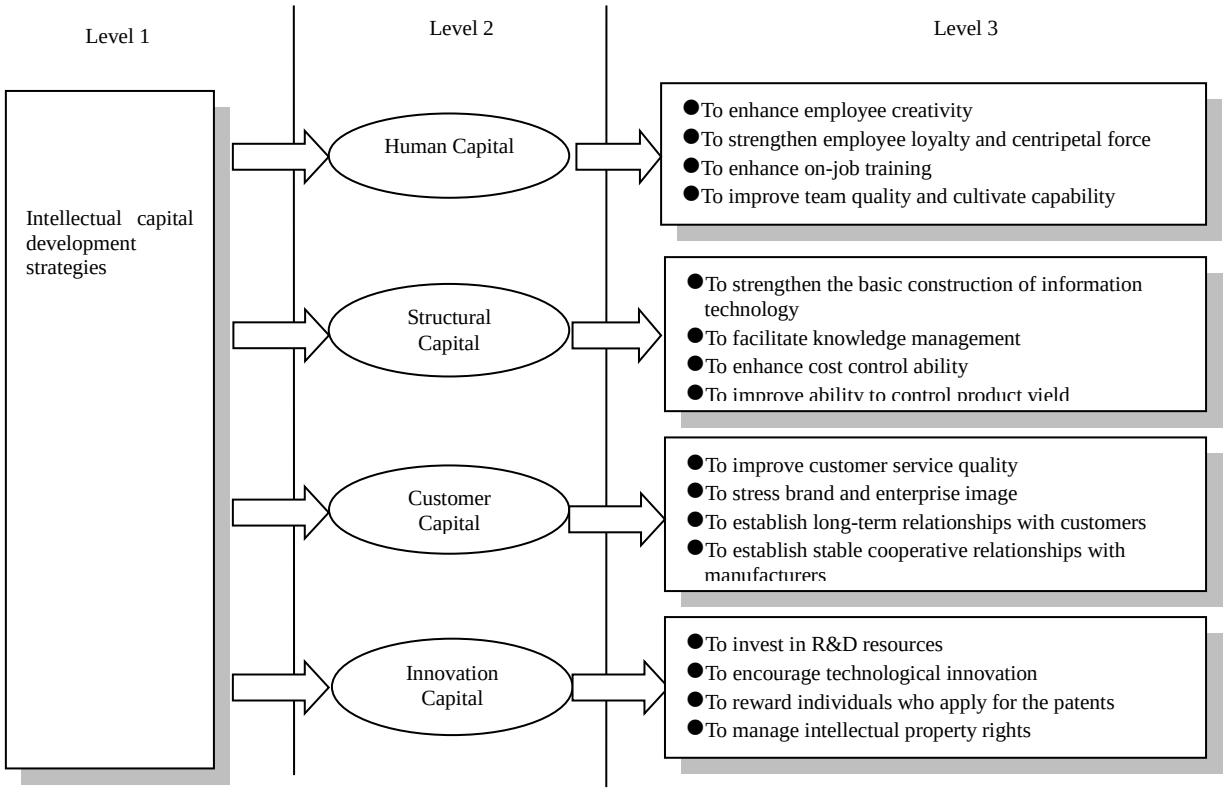


Figure 1. Framework of Questionnaire

3. Empirical Results

3.1. Sample Description

The sample includes top managers of Taiwan-funded FICD houses in China. Forty-two FICD houses with publicly-listed in Taiwan Stock Exchange have subsidiaries with direct or indirect investment in China. The rate of investing China is 70%, publicly. For this study, the questionnaire was mailed to 42 firms, and their top managers in China were invited to fill out the questionnaire. A total of 10 companies did not reply throughout the data collection period. Furthermore, there were seven incomplete questionnaires. No matter how hard to attempt, participants did not response to any of contacts. The removal of these questionnaires from the original sample resulted in a final sample size. Through the collection of questionnaires for three months, total 25 completed questionnaires were finally returned. The overall response rate is 47.62% that is higher than the average response rate (21%) of the industry surveys [20]. In addition, we use the expert questionnaire that can be empirically manipulated even though the sample size is small in this study. Therefore, the sample is representative.

All of participators are male and most of them have degree of graduate school. The majority of the respondents are 41-50 years old. In terms of work experience, 88 percent of the respondents have 10 years above of work experience. Most of them occupy the top managers for 4-6 years. Therefore, participators in our sample are experienced in operating of FICD houses. Detailed information about the survey of respondents is shown in table 1.

TABLE1. Sample Description

Variable	Sector type	Number	Percentage (%)
Gender	Male	25	100
	Female	0	0
Age	30-40 years old	2	8
	41-50 years old	17	68
	51-60 years old	6	24
Education degree	College/ University	2	8
	Master's degree	19	76
	Doctor's degree	4	16
Years of work experience	5-10 years	3	12
	11-15 years	11	44
	16-20 years	8	32
	20 years above	3	12
Years of top manager	3 years and under	5	20
	4-6 years	11	44
	7-9 years	6	24
	10 years and above	3	12

3.2. Priority of Intellectual Capital and Development Strategies

The analytical results show that human capital accounts for 40.84 %, structural capital 28.52%, customer capital 18.74% and innovation capital 11.91% in terms of the relative priority of each construction of intellectual capital in level 2. In response to the manpower demand that comes along with the industrial growth, the results also show that human capital is the most crucial competing factor for Taiwan-funded FICD houses in China, and the next is structural capital.

Regarding the human capital development strategies in level 3, the strategy with the highest weight is strengthening employee loyalty and centripetal force (43.91%). Professionals are the source of core competitiveness in FICD houses, and thus firms invest in creating workforces capable of making contributions that are valuable, unique and difficult for competitors to imitate. High turnover increases the costs associated with the loss of such employees. Therefore, strengthening employee loyalty and centripetal force is the most important task in developing human capital. Regarding the development of structural capital, the first task is improving the ability to control product yield (42.66%), followed by enhancing cost control ability (35.28%), and facilitating knowledge management (12.16%). Regarding customer capital development strategies, the highest weight strategy is establishing stable cooperative relationships with manufacturers (35.77%), followed by establishing long-term relationships with customers (34.15%), and improving customer service quality (20.84%). Regarding innovation capital development strategies, the first priority is managing intellectual property rights (44.99%), followed by rewarding individuals who apply for a patent (21.66%), and investing in research development resources

(21.58%). Among all the final weight of the development strategies of intellectual capital, the strategy with the highest weight is strengthening employee loyalty and centripetal force (17.94%), followed by advancing the ability to control product yield (12.16%), improving team quality and cultivate capability (11.11%), enhancing cost control ability (10.06%) and enhancing on-job training (8.25%). Most strategies focus on human capital and structural capital. The value of Consistency ratio (CR) is below 0.1, indicating consistency with expert judgment.

TABLE2. The Relative Weights Analysis of the Intellectual Capital Development Strategies FICD Houses in China

Citation and Strategy	Relative Weights of Intellectual Capital	Relative Weights of Each Strategy	Final Weights of Each Strategy	Rank
Human Capital (HC)	40.84%			
To enhance employee creativity (H1)		8.66%	3.54%	10
To strengthen employee loyalty and centripetal force (H2)		43.93%	17.94%	1
To enhance on-job training (H3)		20.21%	8.25%	5
To improve team quality and cultivate capability (H4)		27.20%	11.11%	3
Structural Capital (SC)	28.52%			
To strengthen the basic construction of information technology (S1)		9.90%	2.82%	12
To facilitate knowledge management (S2)		12.16%	3.47%	11
To enhance cost control ability (S3)		35.28%	10.06%	4
To improve ability to control product yield (S4)		42.66%	12.16%	2
Customer Capital (CC)	18.74%			
To improve customer service quality (C1)		20.84%	3.91%	9
To stress brand and enterprise image (C2)		9.23%	1.73%	15
To establish long-term relationships with customers (C3)		34.15%	6.40%	7
To establish stable cooperative relationships with manufacturers (C4)		35.77%	6.71%	6
Innovation Capital (IC)	11.91%			
To invest in R&D resources (I1)		21.58%	2.57%	14
To encourage technological innovation (I2)		11.77%	1.40%	16
To reward individuals who apply for the patents (I3)		21.66%	2.58%	13
To manage intellectual property rights (I4)		44.99%	5.36%	8
Total weights of all selection strategies			100%	

4. Conclusion

FICD industry is intelligence-intensive and therefore, development of intellectual capital is important to firm's competition ability and business success. This study applies the fuzzy analytic hierarchy process (FAHP) to prioritize intellectual capital development strategies in Taiwan-funded FICD houses in China. This result shows that human capital is regarded as extremely important and the next is structural capital. According, strategies on intellectual capital development in priority are strengthening employee loyalty and centripetal force (17.94%), advancing the ability to control product yield (12.16%), and improving team quality and cultivate capability (11.11%).

Since the FICD industry in China is younger, it suffers from a lack of high quality professional resources. Turnover and poaching talents from other corporations is a common practice that harms the fostering of professional technology, and adversely influence long-term development of Chinese FICD industry. Furthermore, there is an urgent need to respect and protect intellectual property rights, especially after China becoming one of members in WTO. For future development of FICD industry in China and coping with competition in the global market, at this stage, this study concludes that nothing is more important and critical than building-up R&D capability in each FICD houses. This study provides evidence of various intellectual capital development strategies and tactical actions taken by Taiwan-funded FICD houses in China. The implications of empirical results could be consistent with perspective of resource-based view that competitive advantage is sustained from firm-specific resources, which are rare and superior in use, relative to others and competitors cannot immediately imitate [21]. A firm has competitive advantage when they implement a value creating strategy based on their strategy resource. If intellectual capital is a tree, then human beings are the saps in some companies, the saps that make it grow [1, 5]. Findings from this study are thus consistent with that view of point. Taiwan-funded FICD houses in China put different priority in intellectual capital development strategies might be the reason that the characteristic of strategic resources.

Limitations of this study include the small sample size and the difficultly of precluding other determinant factors for the results. Since the sample is small, the applications of the results should be with

caution, even though small sample size is permitted in the analytic hierarchy process. Future studies could consider other settings to extend this study. For example, firms may alter intellectual capital development strategies for specific firms to achieve and sustain competitive advantage under rapid industry environment transition and considering other factors such as organizational culture.

References

1. Edvinsson, L., Malone, M. (1997 *Intellectual capital: The Proven Way to Establish Your Company's Real Value by Measuring its Hidden Brain power*, London: Piatkus.
2. Ethiraj, S.K., Kale, P., Krishnan, M.S., Singh, J.V. (2005) Where do capabilities come from and how do they matter? A study in the software services industry, *Strategic Management Journal*, 26(1), 25-45.
3. Haas, M., Hansen, M.T. (2005) When Using Knowledge Can Hurt Performance: An Empirical Test of Competitive Bidding in a Management Consulting Company**,** *Strategic Management Journal*, 26(1), 1-24.
4. Amir, E., Lev, B. (1996) Value-Relevance of Non-Financial Information: the Wireless Communications Industry, *Journal of Accounting and Economics*, 22(1), 3-30.
5. Stewart, T.A. (1997) *Intellectual Capital: The New Wealth of Organizations*, New York: Currency Doubleday.
6. Ittner, C.D., Larcker, D.F., Rajan, M.V. (1997) The Choice of Performance Measures in Annual Bonus Contracts, *The Accounting Review*, 72(2), 231-255.
7. Bontis, N. (1999) Managing Organization Knowledge by Diagnosing Intellectual Capital: Framing and Advancing the Sate of the Field, *International Journal of Technology Management*, 18(5-8), 33-462.
8. Sullivan, P.H. (2000) *Value- Driven Intellectual Capital*, New York: John Wiley Co.
9. Bontis, N., Fitz-Enz, J. (2002) Intellectual Capital ROI: A Causal Map of Human Capital Antecedents and Consequents, *Journal of Intellectual Capital*, 3(3), 223-247.
10. Wiklund, J., Shepherd, D. (2003) Knowledge-Based Resources, Entrepreneurial Orientation, and the Performance of Small and Medium-Sized Businesses, *Strategic Management Journal*, 24(13), 1307-1314.
11. Kaplan, R.S., Norton, D.P. (1996) *The Balanced Scorecard: Translating Strategy into Action*, Harvard Business School Press.
12. Johnson, W.H.A. (1999) An Integrative Taxonomy of Intellectual Capital: Measuring the Stock and Flow of Intellectual Capital Components in the Firm, *International Journal of Technology Management*, 18(5-8), 562-575.
13. Dzinkowski, R. (2000) The Measurement and Management of Intellectual Capital: An Introduction, *Management Accounting*, 78(2), 32-36.
14. Deeds, D.L. (2001) The Role of R&D Intensity, Technical Development and Absorptive Capacity in Creating Entrepreneurial Wealth in High Technology Start-ups, *Journal of Engineering and Technology Management*, 18(1), 29-47.
15. Carmeli, A., Tishler, A. (2004) The Relationships between Intangible Organizational Elements and Organizational Performance, *Strategic Management Journal*, 25(13), 1257-1278.
16. Swart, J. (2006) Intellectual Capital: Disentangling an Enigmatic Concept, *Journal of Intellectual Capital*, 7(2), 136-159.
17. Bozbura, F.T., Beskese, A., Kahraman, C. (2007) Prioritization of Human Capital Measurement Indicators Using Fuzzy AHP, *Expert Systems with Applications*, 32(4), 1100-1112.
18. Saaty, T.L. (1980) *The Analytic Hierarchy Process*—*Planning, Priority, Setting, Resource Allocation*, New York: McGraw-Hall.
19. Beskese, A., Kahraman, C., Irani, Z. (2004) Quantification of Flexibility in Advanced Manufacturing Systems Using Fuzzy Concept, *International Journal of Production Economics*, 89(1), 45–56.
20. Dillman, D.A. (2000) *Mail and Internet Surveys: The Tailored Design Method*; New York: John Wiley Co.
21. Barney, J. (1991) Firm Resources and Sustained Competitive Advantage, *Journal of Management*, 17(1), 99-120.

Received on the 17th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 301-307
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

MULTI-CRITERIA ABC METHOD USING HYBRID HEURISTIC MODELS

Bo Li[1], Lianghao Li[1] and Ping Chen[1]

[1]College of Management and Economics, Tianjin University,
Weijin road, Nankai str., Tianjin, 300072, China
E-mail: bingling.0815@163.com

Inventory classification using ABC method is the most widely techniques in enterprises. A hybrid model of rough sets (RS) and Support Vector Machines (SVMs) for ABC classification of stock keeping units (SKU) has been presented in this paper. The whole process involves two stages: Firstly, the attribute reduction of RS has been applied as pre-processor so that we can delete the redundant attributes and conflicting objects from decision-making table but without loss of the efficient information. Then, we realize the classification modelling and forecasting test based on SVMs. By the hybrid heuristic method, we can greatly reduce the dimension of data and highly decrease the complexity in the process of SVMs classification but obtain the good classification performance. Finally, a numerical simulation shows the effectiveness of the proposed hybrid method.

Keywords: Multi-criteria classification, rough sets, support vector machine, decision rules

1. Introduction

ABC classification is a classical scheme to be used to inventory classification. In this scheme, items are ordered in descending order based on some criteria, such as annual dollar usage values, and then divided stock keeping units into three classes: 20% of the first items belong to Class A, which control 80% of the total annual dollar usage of all items; 50%-60% of the last items constitute Class C, which control a relatively small portion of the total annual dollar usage; Items between the two classes are Class B (about 20%-30%). The division of these classes is designed artificially according to the control policy of companies. One company may have the different control policy to the different classes. Generally, Class A has a tight management control and Class C receives a loose one, Class B should have an intermediate control between them. The criterion used in the classical ABC classification is the annual dollar usage value, however, there are many other criteria which can be affect the inventory classification management, such as, the lead time, the rate of obsolescence, the order size requirement etc..

To deal with the multi-criteria ABC classification problems, a joint criteria matrix for two criteria has been proposed by [1], but it cannot be extended to more than two criteria and its applications are limited. The Analytic Hierarchy Process (AHP) is an alternative option for the multi-criteria ABC classification [2]. The approach can incorporate many relevant qualitative and quantitative criteria, but when constructing the pair wise comparison matrix, the preferences of the decision maker have involved in the matrix and the quality of outputs depends directly on this matrix. However, different decision makers may have different views. Therefore, it is necessary to introduce the AI-based methods to reduce the amount of human involvement in the classification process. Guvenir and Erel [3] proposed a method using Genetic Algorithm to multi-criteria classification, and Partovi, et. al. [4] showed that an artificial neural network with BP algorithm could achieve the higher predictive accuracy as a multi-criteria inventory classifier. Wan [5] gave a simple model for multiple criteria inventory classification model which converted all criteria measures of an inventory item into a scalar score. The classification based on the calculated scores using ABC principle is then applied. Further, Ozan and Mustafa [6] integrated fuzzy concepts with real inventory data and designed a decision support system assisting a sensible multi-criteria inventory classifications, they applied it to a small electrical appliances company to validate the design of the proposed multi-criteria inventory classification system. However, their methods have some common drawbacks that may cause difficulties in practice. One is the number of variables which can be input into these models are limited because of the great amount of calculation. However, to satisfy the various demands of markets, companies have an increasing trend of carrying a great number of different items in inventory. The other is that the models cannot directly deal with the qualitative variables, such as, the manager's experience, knowledge and judgment, and it is difficult to explain the relationships between criteria and classification for extracting simple rules. Recently, Yu [7] compared artificial intelligence (AI)-based classification techniques with traditional multiple discriminate analysis (MDA).

The results show that AI-based techniques demonstrate superior accuracy to MDA. Statistical analysis reveals that Support Vector Machines (SVMs) enables more accurate classification than other AI-based techniques. Referring to this result, we adopt SVMs to inventory classification in this paper.

The rough sets theory introduced by Pawlak [8] has proved to be an effective tool for analysis of a vague description of objects, especially studying how to express, learn and induct incomplete data and imprecise knowledge following from information granulation. It has been applied in several fields such as machine learning, data mining, decision and prediction, pattern recognition and fuzzy control [9]. For example, using rough sets, Angel and Alfonso [10] presented a real application where the information system is reduced so as to get a minimum subset of attributes without loss of quality. Chen et al. [11] proposed multiple criteria ABC analysis method that employed a dominance-based rough set approach (DRSA) to generate linguistic rules to represent a decision maker's preferences based on the classification of a test data set. These linguistic rules have been applied to classify other SKUs. Because of the properties of these linguistic rules represented by rough sets, we investigate such an extension to the inventory classification problem based on the rough sets [12].

In this paper, a hybrid approach based on rough sets and SVMs to inventory classification is proposed. On one hand, rough set theory is applied so that the redundant criteria can be ruled out in multi-criteria information system with the qualitative representation without any information loss. On the other hand, classification modelling and forecasting test based on SVMs will further be realized. The dimension of data and the complexity can be greatly reduced in the process of SVMs classification but obtain the good classification performance. Finally, numerical simulations show the effectiveness of the proposed hybrid method.

2. Fundamentals

2.1. Attributes Reduction Based on Rough Set

Formally, a decision system, DS (or an approximation space) can be seen as a system $< \mathrm{U}, \mathrm{C} \cup \mathrm{D}, \mathrm{V}, \mathrm{f} >$, where U is the universe (a finite set of objects, $U = \{x_1, x_2, \cdots, x_m\}$), C is the finite sets of the conditional attribute and D is the finite sets of decision attribute. Each attribute $a \in (C \cup D)$ (attribute a belonging to the considered set of attributes A) defines an information function $f_a : U \to V_a$, where V_a is the set of values of a, called the domain of attribute a. DS is written as a form of decision table where rows represent objectives and columns represent attribute. In decision table, one of the core contents is attribute reduction that reflects the natural information of a decision table. In most cases, attributes in an approximate space are not equally important and especially some ones are redundant. Redundant attributes not only waste many resources (need large storage space) but also have effect on precise and concise decision. Therefore, attribute reduction is very indispensable.

Dependence between attributes determines whether attributes can be reduced and importance of attributes in terms of attributes dependence is often important information of attribute reduction. For a set $B \subset C$, approximate classification quality is defined as:

$$r(B,D) = \frac{|pos_B(D)|}{|U|}$$

Where $|\cdot|$ represents the number of elements in limited sets and $pos_B(D)$ is defined as positive region based on a set of B. If D is all decision attributes and B is some conditional attribute, $r(B,D)$ is defined as probability that can be accurately classified by D after B is used as the partition. Meanwhile $r(B,D)$ represents the capacity of conditional attribute B describing decision attribute D. Based on the dependence between attributes attribute reduction may be defined. Consider $r(C,D) = r(B,D)$ and there is not $B' \subset B$ satisfying $r(B',D) = r(B,D)$, B is called as a reduction of C, where conditional attribute $B \subset D$. There are more than one reduction of attribute sets and $Core(C)$ called as the core of attribute C is defined as the intersection of all reductions of C. The importance of attribute a is defined as $sig(a,B,D) = r(B+\{a\},D) - r(B,D)$. In short, the more important the attribute a

contributing to decision attribute is, the bigger effect the attribute a has on the decision partition.

The fundamental idea of attribute reduction is to use some kind of standard to measure the importance of attributes as heuristic information and to find the optimal reduction from information system or decision system. Optimization means the minimal attributes extraction or being able to cover the largest universe space. The selection of all the standards is determined according to concrete problems. This paper uses the forward attribute reduction based on attribute approximate classification quality. The detailed procedure of which can be seen from the literature [7].

2.2. Classification Mechanism of SVMs

The main idea of SVM is the model of learning from examples, which can be described as: there are n random independent identically distributed examples

$$\{(x_1, y_1),(x_2, y_2),...,(x_n, y_n)\},\ \ y_i \in \{-1,+1\},\ i = 1,2,...,n$$

where (x_i, y_i) is the i^{th} input sample, $y_i \in \{-1,+1\}$ is the expected output of $x_i \in R^n$. The optimal separating hyper plane is determined by giving the largest margin of separation between different classes. The optimal hyper plane bisects the shortest line between the convex hulls of the two classes. The points on the margin of the two classes are called Support Vectors (SV). A set of SV can uniquely determine a hyper plane. SVM can be tracked back on the optimal separating hyper plane on the condition of linear separability. The hyper plane of SVM is written as $w \bullet x + b = 0$.

This paper uses the nonlinear SVM, which satisfies the following optimization problem:

$$\min f(X) = \frac{1}{2}\|W\|_2^2 + C\sum_{i=1}^{n}\xi_i \tag{1}$$

$$\text{s.t. } y_i(W \bullet \Phi(x_i) + b) \geq 1 - \xi_i \tag{2}$$

$$\xi_i \geq 0, i = 1,...,n$$

Where $\bullet$ denotes the inner product, C are used to weight the penalizing variables ξ_i and $\Phi(\cdot)$ is a nonlinear function which maps the input space into a higher dimensional space.

Minimizing the first term $\frac{1}{2}\|W\|_2^2$ in the goal function plays the role of controlling the capacity of the learning machine and avoiding the over fitting of the machine. While minimizing the second term is to minimizing the empirical risk (i.e. error rate). Here, a Lagrange method is used to solve the above problem. Then Eq. (1) can be written as

$$\max W(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j K(x_i, x_j) \tag{3}$$

$$\text{s.t. } \sum_{i=1}^{n} a_i y_i = 0, a_i \in [0, C],\ i = 1,...,n \tag{4}$$

And then the decision function can be written as:

$$f(x) = \sum_{SV} a_i y_i K(x_i, x) + b \tag{5}$$

Where $K(x, y) = \Phi(x) \bullet \Phi(y)$ is a Kernel function. Sigmoid kernel function written as

$K(x, x_i) = \exp\{-\frac{\| x - x_i \|_2^2}{\sigma^2}\}$ is used in this paper because the feature space in accordance with the Kernel function is infinite and finite samples in the space is bound to be separable.

The classification mechanism of Binary-class problem based on SVM is discussed. For multi-class classification, three methods are employed as follows:

1) one-against-rest. For different k classes, k classifiers are constructed and among them the i^{th} classifier is separating the i^{th} class from the rest, i.e. The i^{th} classifier is labelled as 1 and the rest are labeled as –1. k quadric programming is running in the overall process in order to separate each sample. The shortcoming of this method is that multi-class points and none-class points are produced easily.

2) One-against-one. For any two classes, a classifier is constructed to recognize the two. $k(k-1)/2$ classifiers are required to be constructed in the overall process. Each sample is determined to belong to which class according to the result of each classifier. Its drawback is that multi-class points may be produced.

3) hierarchy classification. The method approves the shortcoming of one-against-one. The mechanism is combining all k classes into two classes and then separating each big class into two classes respectively until the basic k leaf nodes are produced. Thus, many different hierarchies are produced and SVM is used for classification in each hierarchy.

3. The Hybrid Classification Method Based on RS and SVM

Fig.1 illustrates the whole modelling procedure based on the hybrid classification model. The process is composed of two steps that are pre-processing based on RS and classification modelling and forecasting test based on SVMs. Firstly, the attribute reduction of RS has been applied as pre-processor so that we can delete redundant attributes from decision table but efficient information lossless. To deal with data more effectively, this paper employs 2-dimension data-pre-processing algorithm based on attribute reduction, which need implement not only reduction lengthways, i.e. attribute reduction but also conflict check breadth, i.e. noise data bearing the same conditional attributes but different decision attributes because the results using SVM training conflict objectives may be different. In general, the number of "wrong" output value on objectives is more than the number of objectives on "right" output value, wrong classification will be gained on the basis of SVM; otherwise right classification will be achieved, but the result will lose its sense if the number between right and wrong is the same. Therefore, conflict objectives must be deleted and this is traverse reduction process.

Redundant attributes and conflict objectives will be deleted from a decision table through reduction. Then the algorithm will transform the front into the back, i.e. transform attributes reduction on RS into the training and test on SVM. As SVM mainly deals with binary-class classification, multi-class classification can be implemented on the foundation of *k*-class algorithms such as one-against-one, one-against-rest and two-prong tree.

RS is used for the fore processor of SVM. Input numbers into SVM are decreased to some extent by way of attributes reduction and conflict reduction based on RS. Thus the computation in the process of SVM classification is diminished and much training time is saved largely. Moreover, the over-fitting is avoided in the way of the hybrid algorithm and classification performance is not decreased.

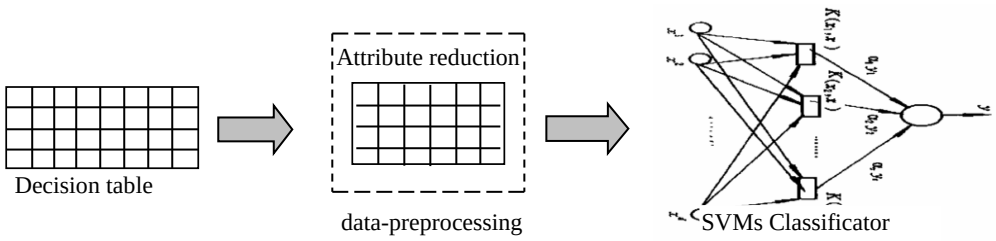


Figure 1. The classification procedure using the RS and SVMs

4. Empirical Comparisons

To demonstrate the application of rough sets and SVMs for inventory classification, we have performed the simulation experiments using the real world data obtained from the reference [9]. A small

sample of 95 data sets from a large pharmaceutical company located in Northeastern United States was used to realize the classification tasks.

4.1. Representing Inventory Items

The classification items with regard to four criteria are represented and the four types of information are unit price ($/unit from 0.1 to 7132), ordering cost ($/order from 0.05 to 300), demand range (units/year from 1-75 units) and lead time (days from 1-70days). Based on the four criteria, the inventory managers have used ad hoc techniques to integrate the above criteria for ABC classifications. The detail materials can refer to the reference [4].

In our experiments, the classification table of inventory items with regard to the above four criteria forms a decision system, where unit price, ordering cost, demand range and lead time are the conditional attributes and the ABC classifications constitute the decision attribute. The system concerns 95 items described by means of this four attributes. Then the decision system *S* can be expressed:

U={1,2, …, 95},

C={unit-price, ordering-cost, demand-range, lead-time},

D={class},

$V_1 = [0.1, 7132], V_2 = [0.05, 300],$

$V_3 = [1, 75], V_4 = [1, 70]$

The information function

$f(x, c_j) \in V_j, x \in U, c_j \in C$

and $f(x, d) | D = \{ClassA \quad or \quad ClassB \quad or \quad ClassC\}$.

We must note that rough sets are concerned with discrete values, it is necessary to transform quantitative attributes (or continuous attributes) into qualitative representations based on some rules, experience or prior knowledge. Therefore, rough sets method is fit for inventory classification problem, because there are many qualitative elements involved in the inventory classification, and in the real application the decision maker always make some judgment according to the vague value, such as, the unit price of some item is high, medium or low etc.. In our experiments, as two criteria of unit price and ordering cost are given the quantitative values, rough sets only can deal with the discrete values, so we need perform some transformations from the quantitative forms into the qualitative values. The simplest rule is the equal interval method, but if the distribution of the criterion values in their value domain is not uniform, the results obtained by this method are not good. Here we use a clustering partition algorithm to realize the discrete process of the quantitative criterion values. The main idea is: first, cluster the classes according to the same values or similar values of the criterion of all inventory items, and then divide all items into several classes; finally, mark the interval labels to represent a qualitative value of this criterion. The number of clustering partition can affect the results of decision rules. In real application, decision maker can determine the suitable interval size based on their experiments, judgment and knowledge. Generally, the important criterion should be partitioned into the more intervals. We give the transformed rules shown in the table 1 after a series of tests.

TABLE 1. The transformation rules

Attributes	Qualitative Values
Unit-price	[0.1,1] or [1,10]; [10,12] or [12,20]; [20,30] or [30,48]; [48,65] or [65,116]; [116,150] or [150,161]; [161, + ∞]
Ordering-cost	[0,2]; [2,8]; [8,150];[150,300]; [300, + ∞]
Demand-range	[1,2];[3,10]; [11,20]; [21,40]; [41,75]
Lead-time	[1,2]; [3,5]; [6,14]; [15,35]; [36,70]

To produce unbiased estimates for the probability of misclassification, 2-fold cross validation for re-sampling method is used. The experiment data is split into two sets, 48 items are used for training and 47 items are used for testing.

Running the criterion reduction algorithm over the inventory classification dataset has resulted in

the three attributes being selected for ABC classification problem. The reduction experiments have been performed for two repetitions using 2-fold cross validation. Table 2 lists the reduction attributes and the quality of approximation at every step and gives the final reduction results. They are unit-price, demand-range and lead-time, and the decision table consisting of 65 objects has been reduced to that of 57 objects.

TABLE 2. The results after reduction of criteria

Exp. No.	1	2
Quality in step 1	{U} 0.0625	{U} 0.4583
Quality in step 2	{U, L} 0.625	{U,L} 0.8333
Quality in step 3	{U, L, D} 0.8750	{U, L, D} 0.8750
Reduction results	{U, L, D}	{U, L, D}

4.2. The Classification and testing by SVMs

The 65 samples are used to train and the rest 30 samples are used to test. The experiments have been performed by SVMs and SVMs+RS. The parameters of SVMs are adjusted accordingly, the results are shown as Table 3.

In the experiment, the first row results are trained and tested by SVM individually. The second row results are trained and tested by RS and SVM. And the three models, One-against-one, One-against -rest and hierarchy, have been applied to prove the proposed method. As shown in table 3, RS is used for feature selection from conditional and decision attributes; the results of reduction are trained and tested by SVMs. By this method, the generalization performance classified by SVMs+RS can be increased compared with SVMs alone.

TABLE 3. The comparisons of the three classification precisions

Param.	One-against-one				One-against-rest				Hierarchy			
	SVM		SVM+RS		SVM		SVM+RS		SVM		SVM+RS	
	T	P	T	P	T	P	T	P	T	P	T	P
$\sigma^2 = 0.1$ C =300	0.77	0.67	0.88	0.63	0.77	0.67	0.77	0.53	0.77	0.67	0.79	0.63
$\sigma^2 = 1$ C=300	0.71	0.60	0.79	0.54	0.79	0.60	0.86	0.60	0.80	0.67	0.88	0.60
$\sigma^2 = 5$ C=300	0.71	0.60	0.77	0.53	0.78	0.57	0.65	0.57	0.80	0.67	0.89	0.63
$\sigma^2 = 10$ C=1000	0.71	0.60	0.77	0.53	0.78	0.57	0.84	0.50	0.80	0.63	0.86	0.53

Where “T” and “P” represents the training and the testing phase respectively.

5. Conclusion

A hybrid heuristic model for multi-criteria classification of inventory items is presented in this paper. The whole procedure includes two steps. The first one is to apply the advantages of rough sets (RS) to handle the qualitative criteria and realize the reduction of multi-criterion attributes without information loss. The second one is to further classify the items of SKU based on SVMs. SVMs have been proven to successfully solve many real world applications. SVMs can find maximal margin boundaries between the classes, minimize the structural risk, and therefore endow the classifier with the excellent generalization performance. Finally, the comparable simulation experiments for the real inventory classification problem have been performed using the method proposed in this paper. It is validated that the use of the hybrid model is a promising analytical tool in the real decision- making systems.

References

1. Flores, B.E., Whybark, D.C. (1987) Implementing multiple criteria ABC analysis. Journal of Operations Management, 7(1), 79-84.
2. Gajpal, P.P., Ganesh, L.S., Rajendran, C. (1994) Criticality analysis of spare parts using the analytic hierarchy process. International Journal of Production Economics, 35(1-3), 293-297.
3. Guvenir, H.A., Erel, E. (1998) Multicriteria inventory classification using a genetic algorithm. European Journal of Operational Research, 105, 29-37.
4. Partori, F.Y., Anandarajan, M. (2002) Classifying inventory using an artificial neural network approach. Computer & Industrial Engineering, 41, 389-404.
5. Wan, L.N. (2007) A simple classifier for multiple criteria ABC analysis. European Journal of Operational Research, 1(16), 344-353.
6. Ozan, C., Mustafa, S.C. (2008) A web-based decision support system for multi-criteria inventory classification using fuzzy AHP methodology. Expert Systems with Applications, 35(3), 1367–1378.
7. Yu, M.C. (2011) Multi-criteria ABC analysis using artificial intelligence-based classification techniques. Expert Systems with Applications, 38(4), 3416–3421.
8. Pawlak, Z. (1982) Rough sets. International Journal of Information and Computer Sciences, 11, 341-356.
9. Greco, S., Matarazzo, B., Slowinski, R. (2001) Rough sets theory for multi-criteria decision analysis. European Journal of Operational Research, 129, 1-47.
10. Angel, M.G., Alfonso, R. (2003) Rough sets and maintenance in a production line. Expert Systems, 20(5), 271-279.
11. Chen, Y., Kevin, W.L., Jason, L., et al. (2008) A rough set approach to multiple criteria ABC analysis. Lecture Notes in Computer Science, 5084, 35-52.
12. Li, B., Zhao, Z.Y., Duan, T.Y. (2004) Hybrid predictive method for multi-criteria inventory classification (in Chinese). Computer Integrated Manufacturing System, 10(5), 594-599.

Received on the 17th of December 2013

Computer Modelling and New Technologies, 2013, Vol.17, No. 4, 308-312
Transport and Telecommunication Institute, Lomonosov 1, LV-1019, Riga, Latvia

ROBUST OPTIMAL PORTFOLIO MODEL

Xing Yu

Department of Mathematics & Applied Mathematics
Humanities & Science and Technology Institute of Hunan Loudi, 417000, P.R. China
Phone: (+86)13723804684. E-mail: hnyuxing@163.com

In this paper, we consider the robust optimal portfolio selection problem where the return mean and covariance are supposed to be uncertain comparing to Markowitz's model. The return mean is uncertain changing in intervals, and we introduce the - cuts of fuzzy number to build the uncertain intervals for covariance. We report on empirical tests in which we compare the robust model with the classical mean-variance model. It shows that the portfolio's return is lower than MV model, but the minimum risk in our model is higher than it. It is benefit for invertors to invest with caution.

Keywords: Robust optimal; mean-variance model; α -cuts set

1. Introduction

Markowitz [1] is the first to formulate the model for maximizing the expected return and minimizing its risk. The risk of his model is measured by the variance of the portfolio return. Markowitz's model is also called mean-variance (M-V) optimization. Although, the M-V model has triggered a tremendous amount of research activities in the field of finance, there have been many practitioners who criticized the model because in M-V model, the important inputs, such as the means and covariance of asset returns have to be estimated from history data, ignoring estimation error, simply treat the estimates as true parameters and plugs them into the optimal portfolio optimization model. Nevertheless, investment is performed for future case. In fact, the two inputs practically are unknown and need to be estimated. Many papers (see for Chopra 1993 [2-3], show that a small perturbation of inputs many lead to a large change in the optimal portfolio. Therefore, it is necessary to deal with the problem of uncertain parameters. Robust optimization is a good choice. Soyster [4] is the first to introduce the idea of robust optimization. Simply speaking, robust optimization aims to find solutions to a given optimization problems with uncertain parameters which would achieve good objective values for all or most of realizations of uncertain parameters [5]. After that, there are considered attention paid to uncertain inputs [6-8] in investment portfolio models.

Another important problem adopted in the model is the method of building the uncertain sets for mean and covariance. Following Ben's [9] approach we suppose that the real expected return mean are within particular ranges or confidence regions constructed, with a prescribed confidence level around their estimated value (Goldfarb and Iyengar, 2003) [10]. That is, we denote an interval to describe the uncertain set for expected return mean. In some literatures, it is assumed that the estimate covariance is accurate such that there is no uncertainty about it. According to [11], which show that the error in mean estimation by far out weights the estimation error in covariance. Nevertheless, in our model, we still consider the uncertainty for covariance. The interval set for uncertain input is similar to the α-cuts of fuzzy number, which is a closed interval.

The rest of this paper is organized as follows. In section 2, we quickly review Markowitz's model. In section 3, we build the uncertain sets for mean and covariance. In section 4, the robust optimal model is proposed. In section 5, we report on empirical tests in which we compare the robust model with the classical mean-variance model. Conclusions are drawn in section 6.

2. Mean-Variance Model

A rational investor does not aim solely at maximizing the expected return of an investment, but also at minimizing its risk. The MV optimization problem was formulated as follows (M1):

$$\text{(M1)} \ \min_{X} X \sum X'$$

subject to

$$\begin{cases} \sum_{i=1}^{n} x_i = 1 & (1) \\ \sum_{i=1}^{n} E(r_i)x_i = r_p & (2) \\ 0 \le x_i \le 1 & (3) \end{cases}$$

where $X = (x_1, x_2 ... x_n)$, x_i are portfolio weights, r_i is the rate of return of instrument i, and $\sum$ is the covariance. The second constraint requires portfolio's expected return to be equal to a prescribed value r_p.

This model is under certain and exact environment, but in real market, the inputs are changing, history cannot replace future. So we should consider the inputs $\sum$ and r_i as an uncertain parameters.

3. Uncertain Sets

In this section, we will build the uncertain sets for mean and covariance. Firstly, we define r' is the estimation of real value r, the uncertainly set I as $I = \{r : r_i' - p_i \le r_i \le r_i' + p_i\}$ for mean. According to Anna [12], the robust counterpart:

$$\min \sum_{i=1}^{n} r_i x_i \ge r_p$$

can be transferred to the following form:

$$\begin{cases} \sum_{i=1}^{n} r_i' x_i - \sum_{i=1}^{n} p_i m_i \ge r_p \\ m_i \ge x_i \\ p_i \ge 0 \end{cases}$$

Then, we turn to the object function $X \sum X' = \sum\sum \sigma_{ik} x_i x_k$, where σ_{ik} is also uncertain. There is rare literature associated to estimate covariance. However, in my view, there is a relation between uncertain set and the α - cuts of fuzzy number, which are expressed as closed intervals.

In [13] Papadopoulos and Sfiris proved the next proposition.

Proposition. Let $X_1, X_2 ... X_n$ be a random sample and let $x_1, x_2 ... x_n$ be sample values assumed by the sample. Let also $\beta \in [0,1)$. If the sample size is large enough, then

$$M(x) = \begin{cases} \dfrac{2-\beta}{1-\beta} - \dfrac{2}{1-\beta}\Phi\left(\sqrt{\dfrac{n-1}{2}}\left(\dfrac{s^2}{x} - 1\right)\right), if & \dfrac{s^2}{1+\Phi^{-1}\left(1-\dfrac{\beta}{2}\right)\sqrt{\dfrac{2}{n-2}}} \le x \le s^2 \\ \dfrac{2-\beta}{1-\beta} - \dfrac{2}{1-\beta}\Phi\left(\sqrt{\dfrac{n-1}{2}}\left(1 - \dfrac{s^2}{x}\right)\right), if & s^2 \le x \le \dfrac{s^2}{1-\Phi^{-1}\left(1-\dfrac{\beta}{2}\right)\sqrt{\dfrac{2}{n-1}}} \end{cases}$$

is a fuzzy number, the base of which is exactly the $(1-\beta)$ confidence interval for s^2 and the α - cuts of this fuzzy number are the closed intervals:

$$^{\alpha}[M]=\left[\frac{s^2}{1+z_{g(a)}\sqrt{\frac{2}{n-1}}},\frac{s^2}{1-z_{g(a)}\sqrt{\frac{2}{n-1}}}\right] \tag{4}$$

where $g(a)=\left(\frac{1}{2}-\frac{\beta}{2}\right)\alpha+\frac{\beta}{2}$ *and* $z_{g(a)}=\Phi^{-1}(1-g(a))$

4. The Optimal Portfolio Model

As showed in section 3, the optimal portfolio model can be formalized as (M2):

$$\text{(M2) } \min X\sum X'$$

subject to

$$\begin{cases} \sum_{i=1}^{n} x_i = 1, 0 \le x_i \le 1 \\ \sum_{i=1}^{n} r_i' x_i - \sum_{i=1}^{n} p_i m_i \ge r_p \\ m_i \ge x_i \\ p_i \ge 0 \end{cases}$$

where r_i' is the estimation of real value r_i, which could be replaced by the rate of return. And $\sum$, for simply, we suppose there are only two instrument, is

$$\Sigma=\begin{pmatrix} \sigma_1^{\,2} & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^{\,2} \end{pmatrix}$$

where σ_1^2,σ_2^2 are respectively the variance of instruments, and ρ is the relationship between the two instruments, $0\le\rho\le1$.

The uncertain sets for σ_1^2,σ_2^2 are as form as (4). That is, the elements of $\sum$ are all uncertain and described as intervals.

5. Empirical Tests

In this section, we will choose two stocks as an example to illustrate the efficient of this model, and compare the results to the MV model. The Yunnan Baiyao Group Limited by Share Ltd stocks NO 000538 and MSFT stocks for a portfolio investment are selected from Chinese market. The data are respectively from 2010-12-15 to 2012-10-15 $(n_1=462)$ and 2008-1-2 to 2011-8-26 $(n_2=921)$. The statistics of the two return are respectively

$$r_1'=0.000179,\ r_2'=0.000384,\ s_1=0.055459,\ s_2=0.0162493$$

And we suppose that the investor expected return is $r_p=0.0002$.

So the MV model could be described as

$$\min X\Sigma X'$$

subject to

$$\begin{cases} \sum_{i=1}^{n} x_i = 1, 0 \le x_i \le 1 \\ \sum_{i=1}^{n} r_i' x_i - \sum_{i=1}^{n} p_i m_i \ge r_p \\ m_i \ge x_i \\ p_i \ge 0 \end{cases}$$

where $X = (x_1, x_2 ... x_n)$ is the variable to be determined, $\sum$ is a robust matrix

$$\Sigma = \begin{pmatrix} \sigma_1^{\ 2} & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^{\ 2} \end{pmatrix}$$

The intervals for σ_1^2, σ_2^2 following (4) are

[0.0028314, 0.0033661], [0.00024884, 0.00028122]

and let ρ change between $[0,1]$. p_i, m_i are dummy variables.

By solving MV model, we obtain $x_1 = 0.44$, $x_2 = 0.56$, and the minimum risk is 0.015833, the return is 0.0002938. In addition, in our model, the minimum risk is 0.024276, the return is 0.0002089. Comparing the result, we find that the portfolio's return is lower than MV model, but the minimum risk in our model is higher than it. It is benefit for investors to invest with caution.

6. Conclusion

This paper extents the general portfolio model in two aspects. The first is considering the uncertain inputs in the model with robust optimization. The second is establishing the relation of fuzzy number and robust intervals for variance. In addition, empirical study comparing the general mean-variance to our robust model shows that although our portfolio return is less low, the minimum risk in our model is higher than the former one, which is important for investors to be scrupulous.

Acknowledgment

This research is supported by Department of Education Topics of Hunan Province. (NO 12C0749).

References

1. Markowitz, H.M. (1952) Portfolio selection. *Journal of Finance*, 7, 77-91.
2. Chopra,V.K. (1993) Mean-variance revisited: Near-optimal portfolios and sensitivity to input variations. *Journal of Investing*, 2(1), pp.51-59.
3. Chopra, V.K. (1993) The effects of errors in means, variances, and covariance on optimal portfolio choice. *Journal of Portfolio Management*, 19(2), 6-11.
4. Soyster A.L. (1973) Convex programming eith set-inclusive constraints and applications to inexact linear programming. *Operations Research*, 21, 1154-1157.
5. Sadiadi, S.J., Mohsen, G. Ehram, S. (2012) Robust optimization framework for cardinality constrained portfolio problem. *Applied soft computing*, 12, 91-99.
6. Chopra, V.K., Ziemba, W.T. (1993) The effect of errors in means, variances and covariance's on optimal portfolio choice, *Journal of Portfolio Management*, 19(2), 6–11.
7. Rockafellar, R.T., Uryasev, S.P. (2000) Optimization of conditional value at risk. *Journal of Risk*, 2, 21–41.

8. El Ghaoui L., Oks, M., Oustry, F. (2003) Worst-case value-at-risk and robust portfolio optimization: Aconic programming approach. *Operations Research*, 51, 543–56.
9. Ben-Tal, A., Nemirovski, A. (2000) Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, 88, 411–24.
10. Goldfarb, D., Lyengar, G. (2003) Robust portfolio selection problems. *Mathematics of Operations Research*, 28, 1-38.
11. Chopra, V.K., Ziemba, W.T. (1993) The effect of errors in means, variances and covariance's on optimal portfolio choice, *Journal of Portfolio Management*, 19(2), 6–11.
12. Anna, G.Q., Alberto, Z. (2008) Robust optimization of conditional value at risk and portfolio selection. *Journal of Banking & Finance*, 32, 2046–2056.
13. Papadopoulos, B.K., Sfiris, D.S., (2012) Non-Asymptotic Fuzzy Estimators Based on confidence intervals. *Computers & Security*, 31(6), 782-790.

Received on the 17th of June 2013

Computer Modelling & New Technologies, 2013, Volume 17, No. 4
Transport and Telecommunication Institute, Lomonosov Str.1, Riga, LV-1019, Latvia

Authors' index

Yuri N. Shunin (born in Riga, March 6, 1951)

- Vice-Rector on Academic Issues (Information Systems Management Institute), professor, Dr.Sc.Habil., Member of International Academy of Refrigeration
- Director of Professional Study Programme Information Systems (Information Systems Management Institute)
- Director of Master Study Programme Computer systems (Information Systems Management Institute)
- **University studies**: Moscow physical and technical institute (1968–1974).
- Ph.D. (physics & mathematics) on solid state physics (1982, Physics Institute of Latvian Academy of Sciences), Dr.Sc.Habil (physics & mathematics) on solid state physics (1992, Ioffe Physical Institute of Russian Academy of Sciences)
- **Publications:** 430 publications, 1 patent
- **Scientific activities**: solid state physics, physics of disordered condensed media, amorphous semiconductors and glassy metals, semiconductor technologies, heavy ion induced excitations in solids, mathematical and computer modelling, system analysis, nanotechnologies, nanodevices, nanoeducation, nanorisks

Igor V. Kabashkin (born in Riga, August 6, 1954)

- Vice-rector for Research and Development Affairs of Transport and Telecommunication Institute, Professor, Director of Telematics and Logistics Institute
- PhD in Aviation (1981, Moscow Institute of Civil Aviation Engineering) Dr.Sc.Habil. in Aviation (1992, Riga Aviation University), Member of the International Telecommunication Academy, Member of IEEE, Corresponding Member of Latvian Academy of Sciences (1998)
- **Publications:** 485 scientific papers and 67 patents
- **Research activities:** information technology applications, operations research, electronics and telecommunication, analysis and modelling of complex systems, transport telematics and logistics

H. Yue

- Ph.D. student of Guangdong University of Technology; Lector of Zhongkai
- University of Agriculture and Engineering (China); Member of the
- biomedical engineering society
- **Publications**: 1 patent, 15 scientific papers, 2 textbooks
- **Scientific activities**: computer vision, artificial intelligence, computer modelling, programming languages

Peng Yang

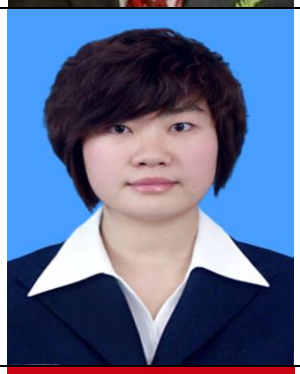
- Master degree candidate, Department of Computer Science Kunming University of Science and Technology

Xiaoling Ren

- Master graduate student of Shandong Normal University
- **Research interests:** Biocomputing, Data mining, and computational intelligence.

Xiyu Liu

- Professor and supervisor doctoral students of Shandong Normal University, Taishan scholars.
- Research interests: information management and e-commerce, computational intelligence etc. Host of National Natural Science Foundation of China, Hong Kong Polytechnic University international cooperation projects, science and technology projects of Shandong province and other several, 2001-2003 study at the Hong Kong Polytechnic University.
- Publications: 120 publications, winner of the China Youth Science and Technology Award in 2006, the second prize-winner of the Shandong Province Natural Science Award in 2005, in 2006 and 2003 respectively won the second prize of Shandong province science and technology progress prize.



Xiuting Li

- Master graduate student of Shandong Normal University
- **Research interests:** Biocomputing, optimization calculation, security and privacy.

Lijun Zhang

- University of Petroleum, East China; Associate professor
- Professional interests: Precision forming technology
- Publications: 15 articles, 2 Patents

Aiming Liu

- Wuhan University of Technology (China), lecturer
- **Professional interest :** Eecohydraulics and fluid mechanics
- **Research interests:** Flow control, drag reduction and Eecohydraulics

Xiaobo Xiong

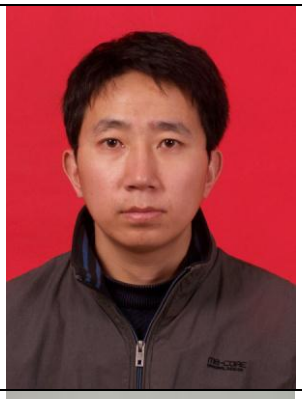
- Doctor of Civil Engineering, Geotechnical Engineering, graduated of Tongji University in Shanghai.
- He participated in the work of higher education in October of 2003, and held a post of Master Instructor in geotechnical engineering in 2004, and the academic leaders of the Geotechnical Engineering from 2006.
- A member of ISRM; A member of IAEG; A member of Chinese Society of Rock Mechanics and Engineering.
- Publications: 43 articles, 4 textbooks.

Jun Liu (born in China, May 8, 1980)

- The director of the Teaching and Research Section, Construction Engineering, Southwest Petroleum University.
- University studies: Chongqing University (1999-2003), bachelor's degree, on engineering mechanics; Chongqing University (2006-2011), Ph.D, on engineering mechanics.
- Scientific activities: Computational mechanics and its application in engineering.

Zairan Li

- Wenzhou Vocational & Technical College; Lecture, Senior Shoe Designer
- **Professional interest:** Shoe Industry and Design
- **Research interest:** Modeling and simulation on shoe industry and optimization design, digital art and industry design methods in shoe industry.
- **University studies:** Huazhong University of Science and Technology (MD) 2007-2010.
- **Publications:** 6 Patent, 7 scientific papers, 3 methodical papers



Chengjun Chen

- Doctor of Mechanical Engineering, Associate professor of Qingdao Technological University
- **University Stu**dies: School of Mechanical Engineering in Shandong University (1999-2008), postdoctoral research in School of Mechanical Engineering of Xi'an Jiaotong University.
- **Research Interests**: Virtual reality and Augmented reality applications in Mechanical Engineering
- Publications: 3 Patents, 30 Papers

Jun Hong

- Doctor of Mechanical Engineering, professor and doctoral supervisor in Xi'an Jiaotong University
- University Studies: School of Mechanical Engineering in Xi'an Jiaotong University (1991-2001),
- Research Interests: Digital design and manufacturing
- Publications: 15 Patents, 83 Papers

Long Yan

- Doctor of Mechanical and Electronic Engineering. Lecture of Shandong Institute of Business and Technology.
- University Studies: Mechanical and Electronic Engineering in Shandong University (1999-2008)
- Research Interests: Computer Vision and Reverse Engineering
- Publications: 30 papers.

Wenhui Li

- Assistant Dean of College of Mechanical Engineering (Taiyuan University of Technology), Associate Professor, Doctor of Mechanical Manufacture and Automation.
- University studies: Bachelor & Master on mechanical manufacture (1992-1999, Taiyuan University of Technology); Doctor on mechanical manufacture and automation (2004-2009, Taiyuan University of Technology).
- Publications: 5 Patents, 50 scientific papers, 1 monographs, winner of provincial natural science foundation of Shanxi (2011021021-4).

Shengqiang Yang

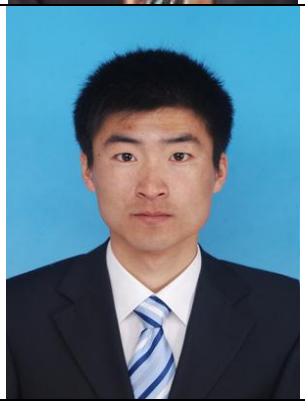
- Professor, Doctor of Mechanical and Electrical Engineering.
- University studies: Bachelor on IC Engine design and manufacture (1981-1985, Taiyuan University of Technology); Master on mechanical manufacture (1994-1997, Taiyuan University of Technology); Doctor on Mechanical and Electrical engineering (2004-2009, Taiyuan University of Technology).
- Publications: 5 Patents, 70 scientific papers, 1 monographs, 5 textbooks, winner of the national natural science foundation of China(51175365).

Xiuhong Li

- Associate department head of Mechanical Foundations (Taiyuan University of Technology), Associate professor, Doctor of Mechanical Manufacture and Automation.
- University studies: Bachelor on mechanical manufacture (1990-1994, Taiyuan University of Technology); Master on mechanical design and theory (1999-2002, Taiyuan University of Technology); Doctor on mechanical manufacture and automation (2005-2010, Taiyuan University of Technology).
- Publications: 3 Patents, 15 scientific papers, 1 monographs, 2 textbooks.

ZhiBin Li

- Doctor of computational mathematics; research supervisor in DaLian JiaoTong University.
- publications: The existence and uniqueness problem of the solution of the inverse problem of 1-D wave equation, The Third international Conference on Nonlinear Mechanics, Shanghai, Aug., 1998



LiJia Ding

- grad student of DaLian JiaoTong University

Lan Liu

- Student of Jiangsu Normal University named after Z. Jia

Shude Li

- Student of Jiangsu Normal University named after Z. Jia

Hong Zhang

- Student of Jiangsu Normal University named after Z. Jia

Chi Zhang

- Student of Jiangsu Normal University named after Z. Jia

Zhigang Jia

- Professor of Jiangsu Normal University; Doctor of Mathematics.
- University studies: PHD on Numerical Mathematics, East China Normal University named after M. Wei, Shanghai, China, Sep. 12006–Jun. 2009.
- Publications: 30 articles
- Scientific activities: Numerical Linear Algebra (especially in least squares problems, structured matrix polynomials and quadratic inverse eigenvalue problems), Numerical Algorithm, Image Processing
- Research projects: National Science Foundation of China (11201193): Real-valued Spectral Decomposition of Structured Matrix Polynomials with Applications in Model Updating. Principal Investigator, 2013.1—2015.12; National Science Foundation of China (11171289): Highly Efficient Algorithms for Large Scale Linear Equations and Applications in Image Processing. Investigator, 2012.1—2015.12. National Science Foundation of China (10771073): Some Problems of Rigid System and Matrix Polynomials. Investigator, 2008.1—2011.12.

Lifen Yun (born in Heilongjiang Province, P.R. China, July 17, 1984)

- Ph. D. candidate of Beijing Jiaotong University
- visiting scholar of Mississippi State University
- **Research interests:** traffic planning and management, network optimization of the comprehensive passenger transport hub service

Changxu Ji (born in Jiangsu Province, P.R. China, April 10, 1960)

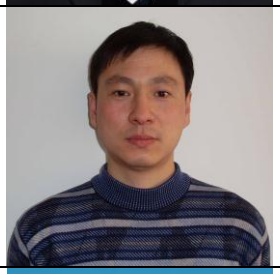
- Professor, Beijing Jiaotong University
- Ph.D. degree in Jilin university of technology, 1996
- **Research interests:** train safety, train fault diagnoses, traffic planning and management, comprehensive passenger transport hub service optimal

Hongqiang Fan (born in Heibei Province, P.R. China, October 25, 1985)

- Ph. D. candidate of Beijing Jiaotong University
- **Research interests:** traffic flow, theory, cellular automaton, model and urban intersection

Xi Li

- Subway Operation Technology Centre, Mass Transit Railway Operation Corporation LTD, Engineer
- State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, post-doctoral
- **Research interests**: traffic safety analysis and early-warning, urban traffic design and management, traffic vehicle structure analysis



Dawei Yin

- Ph.D. student of College of Instrumentation & Electrical Engineering, Jilin University, China

Xianyong MENG

- School of Resources and Environment Science, Xinjiang University, Urumqi, 830046, China;
- Key Laboratory of Oasis Ecology Ministry of Education, Xinjiang University, Urumqi, 830046, China;
- Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, 830011, China;

Yuchuan Guo

- School of Resources and Environment Science, Xinjiang University, Urumqi, 830046, China

Axin Xi

- School of Resources and Environment Science, Xinjiang University, Urumqi, 830046, China

Xiaonan JI

- Jiang An Campus of Sichuan University, College of Architecture and Environment, Chengdu, 610207, China

Qiudong Zhao

- Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou,730000, China

Yang Liu

- School of Resources and Environment Science, Xinjiang University, Urumqi, 830046, China

Sufang Yue

- Lecturer
- School of Mathematics and Computing Science, Anqing Normal University,Anqing, 246011, China
- Scientific activities: numerical solution of partial differential equations

Guixian Xu

- Associate professor, supervisor of master
- Department of Computer, College of Information Engineering, Minzu University of China, Beijing, China
- Scientific activities: Data Mining, Tibetan Information Processing

Lirong Qiu

- Associate professor, supervisor of master
- Department of Computer, College of Information Engineering, Minzu University of China, Beijing, China
- Scientific activities: Artificail Intelligence, Semantic Ontology

Qichao He

- Department of Computer, College of Information Engineering, Minzu University of China, Beijing, China
- Scientific activities: Data Mining

Zhibin Zeng

- Associate professor
- Engineering Research Centre of Digital Audio & Video, Ministry of Education, Communication University of China, Beijing, China,
- Research interests: digital signal processing and digital broadcasting

Qian Ma

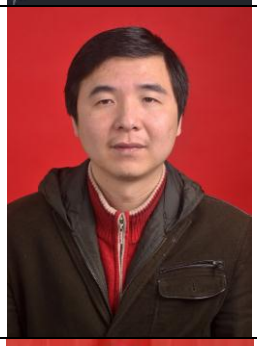
- Engineer
- Beijing Institute of Information Technology, Beijing, China

Yang Sheng

- Physics Specialty, Bachelor of science, associate professor
- **University studies:** Major in Physics in BeiJing Normal University from 1982 to 1986. Major in Computer Technology, Master of Engineering, in Fuzhou University from 2004 to 2006.
- **Publication:** having issued more than ten papers.

Yu WenSen

- Computer software and theory Specialty, Master of Engineering, Associate professor
- University studies: Major in Computer application technology in Fuzhou University from 1998 to 2001; Major in Computer software and theory, Master of Engineering, in Jiangxi Normal University from 2001 to 2004.
- Publication: having issued more than ten academic dissertations.

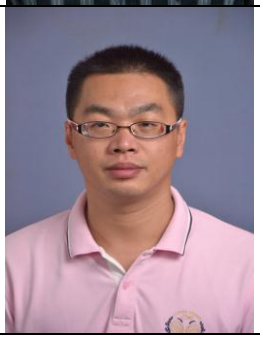


Liu ChangYong

- Computer Specialty, Bachelor of Engineering, Lecturer
- University studies: Major in Computer and Its Application in Fujian Forestry Institute from 1993 to 1996; Major in Computer Technology, Master of Engineering, in Fuzhou University from 2004 to 2006.
- Publication: having issued more than ten academic dissertations.

Cheng Rengui

- Master of Computer Software and Theory, Associate professor
- University studies: 1990-1994 Fujian Normal University, studying physical education, 1998-2001 Northwestern University Master of Computer Software and Theory.
- Publication: Published more than 10 papers



Guo Lei

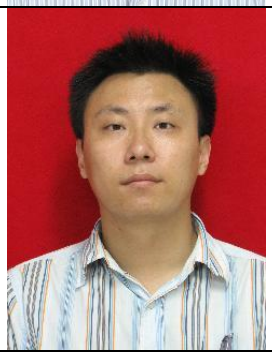
- Computer Science Specialty ,Bachelor of science, associate professor
- **University studies:** Major in computer science in Fujian Normal University from 1997 to 2001. Major in Computer Technology, Master of Engineering, in Fuzhou University from 2004 to 2006.
- **Publication:** having issued more than ten papers.

Meng Shimin

- CEO of UlinkM.com, Human-computer interaction and cognitive science.
- **University studies:** 1984-1987 Fujian Province Jianyang secondary normal. 1994-1996 USTC Department of Earth and Space Sciences, Computer Science.
- **Publication:** more than six papers on cognitive coupling states and cognitive digital manifold.

Weilong Li

- Ph.D. student of Zhejiang University, lecture of Guangxi university of Science and Technology
- Fields of research: geometric modelling, reverse engineering and computer graphics, 3D Garment CAD, Shoe CAD, Sketch-based CAD of toys and cartoons.
- Publications: 5 scientific papers



Jin Wang

- Associate Professor of the Institute of Engineering and Computer Graphics.
- University study: Ph.D in applied mathematics(2008, Zhejiang University)
- Fields of research: design automation and optimization, geometric modelling, reverse engineering and computer graphics, 3D Garment CAD, Shoe CAD, Sketch-based CAD of toys and cartoons.
- Publications: More than 20 scientific papers and 3 teaching books.

Guodong Lu

- Professor of the Institute of Engineering and Computer Graphics, Deputy Dean of Undergraduate School, Director of the Teaching & Researching Division, and the Deputy Dean of the honours college of Chu Kochen.
- University study: Ph.D in applied mathematics (2000, Zhejiang University)
- Fields of research: free-form curves and surfaces modelling and simulation including 3D garment CAD, 3D carton toy CAD, 3D last CAD and knitting CAD, Numerical control machine, Digital machine prototype, virtual reality, web3D,Enterprise informationization, e-government.
- Publications: More than 40 scientific papers and 6 teaching books.

Xuechun Wang (born in Luoyang City, Henan Province, China, January 28 1982)

- School of Information Engineering, Huanghe Science and Technology College (China). Lecturer.
- **University studies:** Daqing Petroleum Institute. 2004-2007, master.
- **Publications**: 1 Patent, 15 scientific papers.

Zhang Na

- Doctor of College of Information Science and Technology of Donghua University; Lector of Department of Computer Science and Engineering of Henan University of Urban Construction.
- **Professional interests:** Computer and Professional English Language
- **Research interests:** machine intelligence; pattern recognition; semantic retrieval.

JingXiaLiu, female, born in September, 1971

- graduate from Inner Mongolia industrial university in July 1996, the Han nationality,
- Associate professor, Inner Mongolia University of science and technology information engineering college of automation.
- Research direction: fault analysis for electric power system and intelligent electric network.

XueDong Xu

- Changchun institute of Technology;
- lecturer
- **Professional interests:** Advanced manufacturing, Control Technology
- **Research interests:** Computer Application, Security and Encryption

Yi Wang

- The Physics and Electronics Science College (Guizhou Normal University), Professor
- **University studies:** Ph.D. on electronic engineering (2009, Guizhou University)
- **Research interests:** Electronic Design, Circuit & System

Liren He

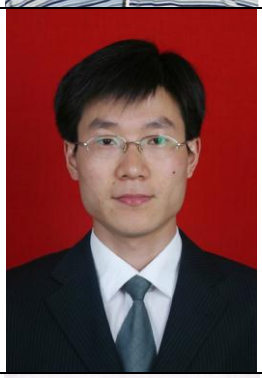
- The Science College (Guizhou University), Postgraduate
- **Research interests:** Electronic Design and Circuit & System.

Cuiye Song

- Doctoral candidate of Northwestern Polytechnical University(China) named after Q. Wang
- **Research interests:** Technique of simulation and verification based on model of cyber physical system

Chenglie Du

- Born in Xi'an city, China, October 25,1971
- Doctor of Navigation, Guidance and Control, Northwestern Polytechnical University since 1999. Assistant dean and professor of Department of Computer, NWPU
- **Research interests:** Design and verification of complex embedded systems and cyber-physical systems
- **Publications:** 20 publications



Gang Li

- Doctor of Computer Science and Technology, NWPU, China since 2011. Lecture of North China Electric Power University, China
- **Research interests:** Reliability of complex systems and cyber physical power system
- **Publications:** 10 publications

Fan Zhang

- Ph.D. student of China University of Geoscience (Wuhan)
- **Research interests:** GPS solution, GIS, Indoor Location

Meng Wu (born in Hancheng, Shaanxi, November 15, 1979)

- Ph.D. student of Xi'an University of Architecture and Technology; Lector
- **Professional interests:** Image inpainting and image fusion.
- **Research interests:** The virtual revelation research of murals in tang dynasty.

Huiqin Wang

- Xi'an University of Architecture and Technology; professor.
- **Professional interests:** Image processing and pattern recognition
- **Research interests:** The digitalization of antique and museum.

Zhenlong Zhang

- Xi'an University of Architecture and Technology;
- Lector

Chen He

- Ph.D student at Shandong University, P.R.China.
- Research interest: Computer graphics, Geometry processing and Data compression

Die Wang

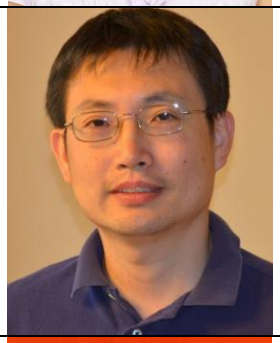
- Master student at Shandong University, P.R.China.
- Research interest: Computer graphics, and Compression of dynamic meshes

Xueqing Li

- Professor, Supervisor of Doctoral Candidate, Shandong University, P.R.China.
- Research interest: Computer graphics, Virtual reality and Digital image processing.

Yan Zhang

- Ph.D. student in multi-media of Shandong University, Lecturer at Shandong Jianzhu University
- **Research interests**: image processing, image retargeting



Jonathan Z. Sun

- Ph.D in computer science, associate professor at The University of Southern Mississippi and adjunct professor at Shandong University of Technology.
- **Research interests:**:algorithms and data structures, with applications in security, biology, information visualization, etc.

Jingliang Peng

- Professor, School of Computer Science and Technology, Shandong University, China
- **University studies:** Computer Science Department, Peking University (1992-1997), B.S.; Computer Science Department, Peking University (1997-2000), M.S.; Ming Hsieh Department of Electrical Engineering, University of Southern California (2001-2006), Ph.D.
- **Publications:** about 40 technical papers.
- **Research interests:** digital geometry processing, content-based multi-media data retrieval, and image analysis & understanding

Shi yuan

- Da Lian JiaoTong University doctor
- Da Lian institute of Science and Technology Lector

Meng Chun Kao

- Department of Business Administration at Yuanpei University of Science and Technology, Assistant Professor
- **University studies**: Ph.D. in Accounting (2003-2008, Graduate School of Management at Yuan-Ze University)
- **Research interests**: accounting, management and performance evaluation

Bo Li

- Tianjin University; Professor
- **University studies**: Tianjin University (1996–2000), Ph.D., Major: Management Science and Engineering. Nankai University (1989-1992), Master, Major: Computer Control. Nankai University (1985-1989), Bachelor, Major: mathematics.
- **Publications:** 70 articles, 2 textbooks.
- **Research interests:** Logistics and Supply chain management; Intelligent Computation.

Lianghao Li

- Tianjin University; Doctor
- **Research interests:** Computation network; Intelligent Computation.

Ping Chen

- Tianjin University; Master
- **University studies**: Tianjin University (2011-2014), Master, major in Logistics and Engineering.
- **Research interests:** Logistics and Supply chain management.

Xing Yu

- Hunan university of humanities, Science and technology (China); Lector
- **Professional interests**: Applicate mathematics; Finance model
- **Research interests:** the optimal portfolio model; price discovery model; data mining

Cheng Wang

- Math Dept. China JiLiang University, Hangzhou, China
- Lecture
- **Research interest:** probability and statistics, mathematical finance.

Lei Hua

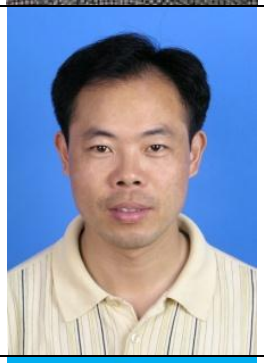
- Lecturer
- School of Economics and Management, Southeast University, Nanjing, China
- School of Business, Linyi Normal University, Linyi, China

Chunjie Yu

- Associate professor
- School of Business, Linyi Normal University, Linyi, China

Wenping Wang

- Professor
- Ph.D. supervisor
- School of Economics and Management, Southeast University, Nanjing, China



Zhang Jilin

- Associate Professor, Phd.
- **Research interest:** Financial Information and data processing, Quantity Economy, financial econometrics

Xu Wei

- Ph.D. student, associate chief physician, School of Management, Harbin Institute of Technology, Harbin, China.
- **Scientific activities:** medical administration management and health policy.

Ding Yunlong

- Ph.D.supervisor,School of Management, Harbin Institute of Technology, Harbin, China.
- **Scientific activities:** Science and technology policy and management

Liu Yang

- Ph.D. student, School of Management, Harbin Institute of Technology, Harbin, China.
- **Scientific activities:** medical administration management and health policy

Rong Li (born in Xinzhou, China, December, 1974)

- Master of School of Computer and Information Technology from Shanxi University in Taiyuan, China, associate professor in the Department of Computer at Xinzhou Teachers' University in Xinzhou, China.
- Research interests: intelligent information processing and machine learning.
- Publications: 25 scientific papers, 2 monographs, 4 textbooks.