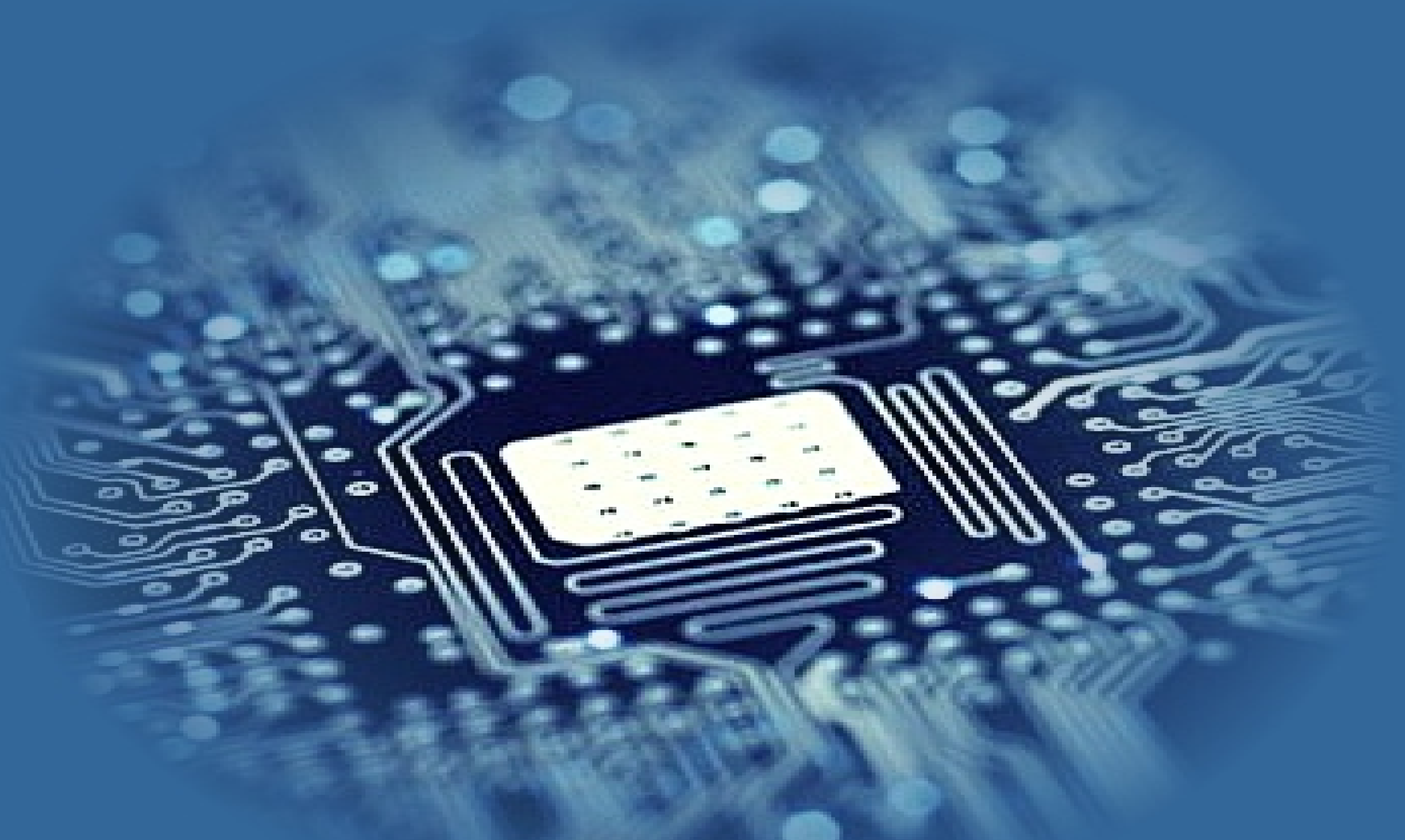
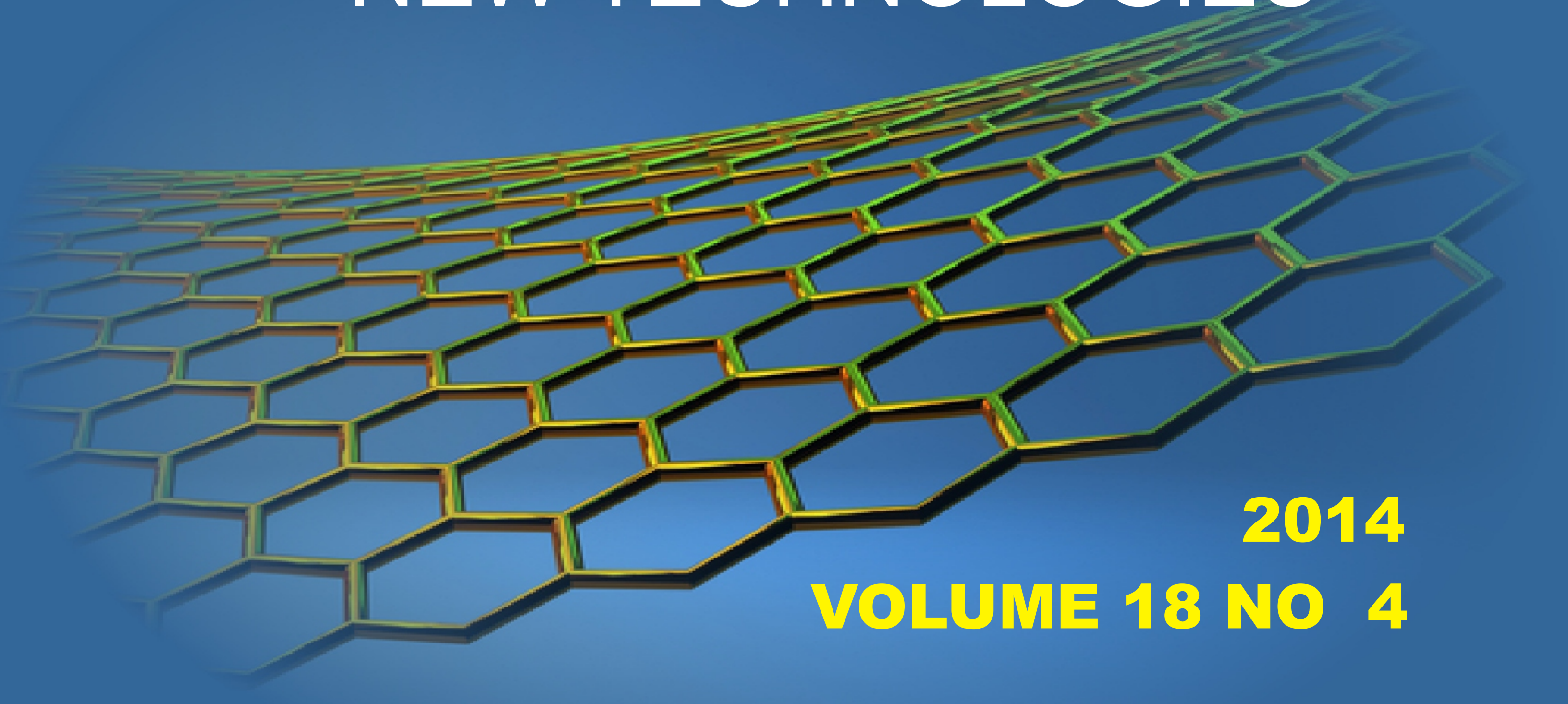


COMPUTER MODELLING AND NEW TECHNOLOGIES



2014

VOLUME 18 NO 4

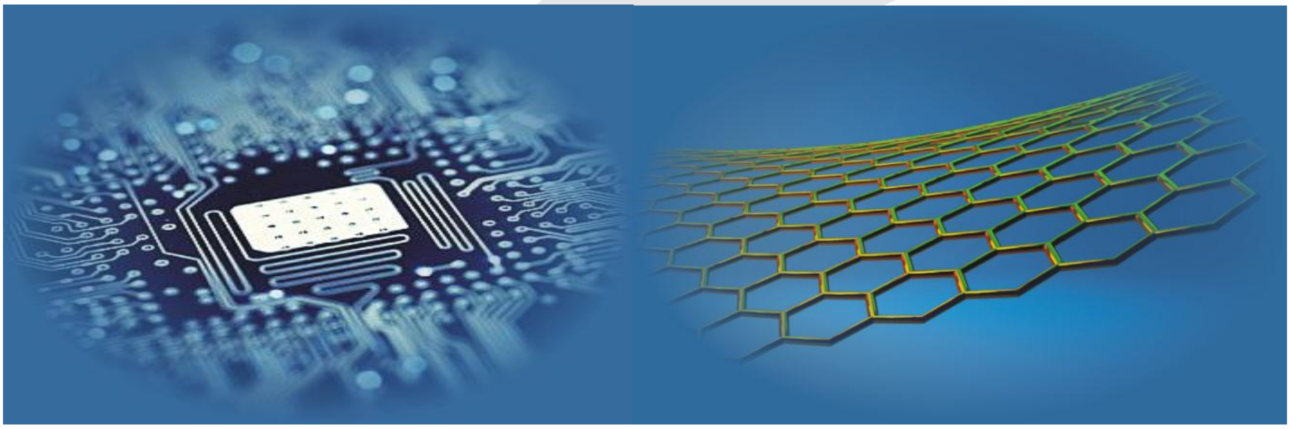
ISSN 1407-5806 ISSN 1407-5814 on-line

Transport and Telecommunication Institute
and
Latvian Transport Development and Education Association

Computer Modelling and New Technologies

2014 Volume 18 No 4

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



Riga – 2014

Journal topics:	**Host Organization**	**Supporting Organizations**
mathematical and computer modelling computer and information technologies natural and engineering sciences operation research and decision making nanoscience and nanotechnologies innovative education	Transport and Telecommunication Institute	Latvian Transport Development and Education Association Latvian Academy of Sciences Latvian Operations Research Society

Articles should be submitted in **English**. All articles are reviewed.

EDITORIAL CORRESPONDENCE	**COMPUTER MODELLING AND NEW TECHNOLOGIES, 2014, Vol. 18, No.4** **ISSN 1407-5806, ISSN 1407-5814 (on-line: www.tsi.lv)**
Transport and Telecommunication Institute	**Scientific and research journal** **The journal is being published since 1996**
1 Lomonosova, **Bld 4,** LV-1019, Riga, **Latvia** **Phone: (+371) 67100594** Fax: (+371) 67100535 E-mail: cm&nt@tsi.lv www.tsi.lv	The papers published in Journal 'Computer Modelling and New Technologies' are included in: **INSPEC (since 2010),** www.theiet.org/resources/inspec/ **VINITI (since 2011),** http://www2.viniti.ru/ **CAS Database** http://www.cas.org/ **EI Compendex**

Content

Mathematical and Computer Modelling

Information Technologies

Operation research and decision making

Nature Phenomena and Engineering

Editors' Remarks

****

Arithmetic on the Frontier

by Rudyard Kipling

A great and glorious thing it is
To learn, for seven years or so,
The Lord knows what of that and this,
Ere reckoned fit to face the foe -
The flying bullet down the Pass,
That whistles clear: "All flesh is grass."

Three hundred pounds per annum spent
On making brain and body meeter
For all the murderous intent
Comprised in "villanous saltpetre!"
And after -- ask the Yusufzaies
What comes of all our 'ologies.

A scrimmage in a Border Station -
A canter down some dark defile -
Two thousand pounds of education
Drops to a ten-rupee jezail -
The Crammer's boast, the Squadron's pride,
Shot like a rabbit in a ride!

No proposition Euclid wrote,
No formulae the text-books know,
Will turn the bullet from your coat,
Or ward the tulwar's downward blow
Strike hard who cares -shoot straight who can -
The odds are on the cheaper man.

One sword-knot stolen from the camp
Will pay for all the school expenses
Of any Kurrum Valley scamp
Who knows no word of moods and tenses,
But, being blessed with perfect sight,
Picks off our messmates left and right.

With home-bred hordes the hillsides teem,
The troop-ships bring us one by one,
At vast expense of time and steam,
To slay Afridis where they run.
The "captives of our bow and spear"
Are cheap - alas! as we are dear.

Rudyard Kipling (**1809-1849**) ♣

****

This 18th volume No.4 presents actual papers on main topics of Journal specialization, namely, **Mathematical and Computer Modelling, Information and Computer Technologies, Operation Research and Decision Making and Nature Phenomena and Innovative Technologies.**

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 Shunin **Yuri Shunin**

Igor Kabashkin

♣ **Joseph Rudyard Kipling** (30 December 1865 – 18 January 1936) was an English short-story writer, poet, and novelist. He is chiefly remembered for his tales and poems of British soldiers in India and his tales for children. He was born in Bombay, in the Bombay Presidency of British India, and was taken by his family to England when he was five years old. Kipling is best known for his works of fiction, including The Jungle Book (a collection of stories, which includes and his poems, including "Mandalay" (1890), "Gunga Din" (1890), "The Gods of the Copybook Headings" (1919), "The White Man's Burden" (1899), and "If—" (1910). He is regarded as a major "innovator in the art of the short story"; his children's books are enduring classics of children's literature; and his best works are said to exhibit "a versatile and luminous narrative gift".

A SoPC design of a real-time high-definition stereo matching algorithm based on SAD

Xiang Zhang*, Huaixiang Zhang, Yifan Wu

School of Computer, Hangzhou Dianzi University, Hangzhou 310018, China

Received 1 January 2014, www.tsi.lv

Abstract

The System-on-Programmable-Chip (SoPC) architecture to implement a stereo matching algorithm based on the sum of absolute differences (SAD) in a FPGA chip is proposed. The hardware implementation involves a 32-bit Nios II microprocessor, memory interfaces and stereo matching algorithm circuit module. The Nios II microprocessor is a configurable soft IP core in charge of managing the buffer of the stereo images and users' configuration data. The system can process any different sizes of stereo pair images through a configuration interface. The maximum horizon resolution of stereo images is 2048. When the algorithm core works under 60MHz, the 1396×1110 disparity map can be achieved at 30 fps speed.

Keywords: Stereo matching, System-on-programmable-chip, FPGA, Disparity map, SAD

1 Introduction

The Stereo vision has been one of the most active research topics in computer vision and widely used in many application areas including intelligent robots, automated guided vehicle, human-computer interface, and so on [1- 3]. Stereo matching algorithms have played an important role in stereo vision. They can be classified into either local or global methods of correspondence. Local methods match one window region centred at a pixel of interest in one image with a similar window region in the other image by searching along epipolar lines. The performance of local stereo matching algorithms depends to a large extent on what similarity metric is selected. Typical similarity metrics are cross-correlation (CC), the sum of absolute differences (SAD), the sum of squared differences (SSD), *etc.* SSD and SAD find correspondences by minimizing the sum of squared or that of absolute differences in W×W windows. The computational complexity for N×N resolution image pair, W×W window size and D disparity level is O(N^2W^2D). It can be decreased to O(N^2D) by some kind of optimization tips [4]. Therefore, the stereo vision has limitations for real-time applications due to its computational expense.

Many researchers have proposed their FPGA implementations of stereo vision algorithms in literature. The circuit [5] is a stereovision system based on a Xilinx Virtex II using the SAD algorithm. The system can process images with a size of 270 × 270 at a frame rate of 30 fps. Paper [6] presents a FPGA-based stereo matching system that operates on 512 × 512 stereo images with a maximum disparity of 255 and achieves a frame rate of 25.6 fps running under a frequency of 286 MHz. In [7], a development system based on four Xilinx XCV2000E chips is used to implement a dense, phase correlation-based stereo system that runs at a frame rate of 30 fps for 256 × 360 pixels stereo pairs. Gardel *et al* introduce in [8] their design, which can obtain 30,000 depth points from images of 2 Mpix at a frame rate of 50 frames per second under a 100 MHz working frequency. A real-time fuzzy hardware module based on a colour SAD window-based technique is proposed in [9]. This module can theoretically provide accurate disparity map computation at a rate of nearly 440 frames per second without considering the memory delay and other factors of time consumption, thus giving a stereo image pair with a disparity range of 80 pixels and 640 × 480 pixels resolution. The design in [10] is a 7 × 7 binary adaptive SAD based real-time stereo vision architecture with a depth range of 80, which is implemented on the Altera Cyclone II EP2C70 FPGA chip based on 800 × 600 colour images and operates in real-time at a frame of 56 Hz. The architecture captures the 90 Megapixel/sec 12 bit signals of two cameras in real-time and does not require memories external to the FPGA. However, these designs rarely attain the target of producing above 720P resolution disparity map at real-time speed.

In this paper, we propose the SoPC architecture to implement a stereo matching algorithm, which can process HD stereo, images in real-time by using the SAD stereo matching algorithm. The stereo matching process, including cost calculation, cost aggregation and L-R checking, are designed and parallelized within a pipelined architecture. Based on efficient hardware-oriented optimizations, our design achieves 30 frames per second when it matches 1396 × 1110 high-definition stereo images under 60MHz working frequency.

* ***Corresponding author*** e-mail: gavin@hdu.edu.cn

2 SoPC architecture for SAD matching algorithm

The SoPC architecture proposed herein is divided into the following main modules as shown in Figure 1:

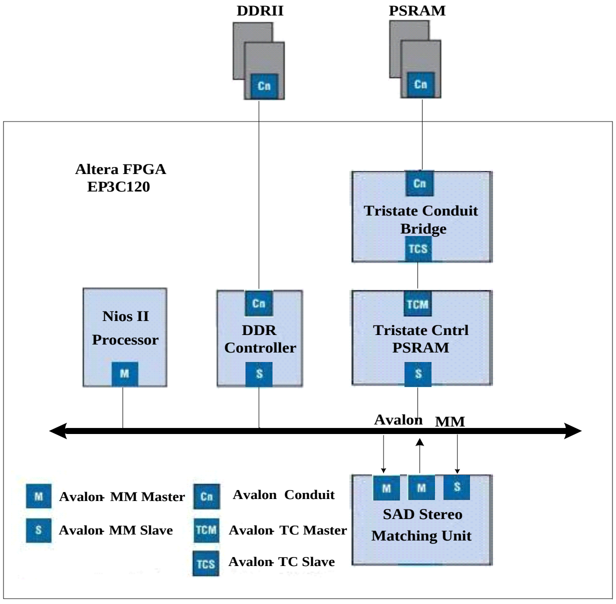


FIGURE 1 Proposed SoPC architecture

(a) Nios II processor system: It consists of a 32-bit Nios II processor core, a set of on-chip peripherals, on-chip memory, and interfaces to off-chip memory.

(b) SAD Stereo Matching Unit (SSMU): This unit computes sum of the absolute difference as similarity metrics to seek disparities from 64 candidates 5×5 windows. The disparities for right and left image are computed concurrently and sent to L-R checking module to take valid detection. The unit has three Avalon-MM interfaces. One is slave interface to communicate with the Nios II CPU. The other two are read and write master interfaces. The read master is in charge of reading raw data of stereo images from the off-chip PSRAM acting as frame buffer in system. The write master takes charge of write final disparities to the DDRII.

(c) DDR Controller and PSRAM Controller: These two IPs are provided by Altera. They enable the system to access the DDRII and PSRAM memory out of the FPGA chip.

The modules listed above are all synthesized in one EP3C120 FPGA chip produced by Altera Company. They connect with each other by Avalon bus interface.

3 Implementation of the SAD stereo matching unit

3.1 THE SAD STEREO MATCHING ALGORITHM

The SAD algorithm has the advantage of computational efficiency. The SAD equation used for 5 × 5 windows with a maximum disparity of 64 can be seen below:

$$\mathrm{SAD}(i,\ j, disp)$$
$$=\sum_{h=-2}^{2}\sum_{k=-2}^{2}\left|P_R(i+h,\ j+k)-P_L(i+h,\ j+k+disp)\right|, \qquad (1)$$

where ***disp*** is the disparity value ranging from 0 to 63, P_R (i, j) serves as the reference pixel in the right image and $P_L(i, j+disp)$ as the currently analysed candidate pixel in the left image. The reference 5 × 5 window centred at $P_R(i, j)$ is compared to 64 possible candidate windows in left image to calculate 64 SAD values. There are 25 bytes of data in the right image and 340 bytes of data in the left image involved in the calculation of disparity(i, j), where disparity(i, j) means the disparity value of the pixel(i, j).

3.2 IMPLEMENTATION OF THE SAD STEREO MATCHING UNIT

The layout plan of the SSMU is showed in Figure 2. At the first stage of the SSMU is a custom DMA engine. It is in charge of transferring all raw image data from the PSRAM to the dual-clock FIFOs.

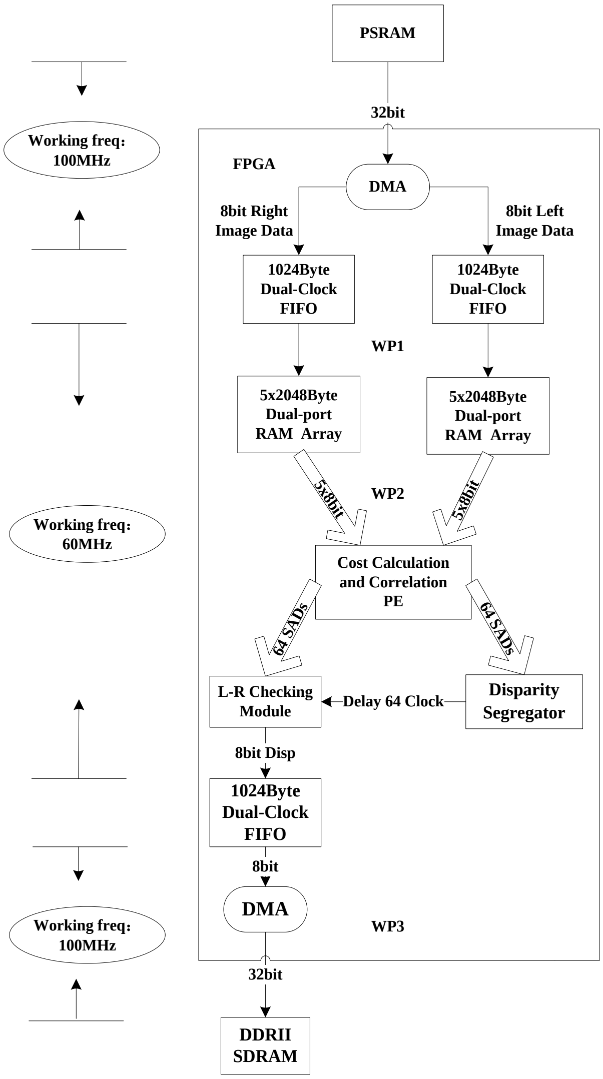


FIGURE 2 Layout plan of the SSMU module

Every time the DMA engine invokes 8 times of word size pipelined Avalon-MM interface reading to get 32 bytes data from the right or left image in turn and it almost consumes 27 clocks. Therefore, under the working frequency of 100MHz, the DMA engine can offer about 113MB per second data bandwidth. It is enough for 2×1396×1110@30fps needs.

The two dual-clock FIFOs (DCFIFO), followed by the DMA engine, has two functions. One is a temporary storage for the raw pixel data. The other is separating the workspace into two regions, which work under different working frequency. At the writing and reading side of the FIFO, the working frequency is 100MHz and 60MHz respectively.

Beside the read side of the DCFIFO is a dual-port RAM (DPRAM) array, which is constituted by 5 DPRAMs, each with 2048 byte of memory space. They are used as line buffers for cost calculation and correlation processing element (CCCPE). Every DPRAM has data bus connected with CCCPE; therefore, every DPRAM array can output 5 bytes of image data to the CCCPE in every clock cycle.

The CCCPE is the most complex hardware circuits in the SSMU. As shown in Figure 3, there are 64 SAD computers and 2 shift-tap devices in the CCCPE.

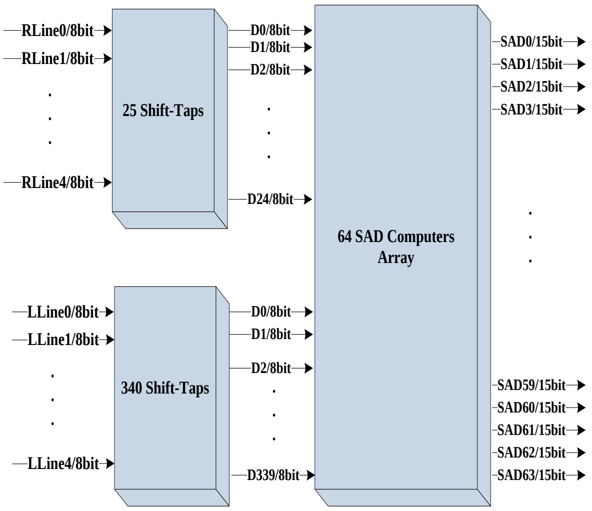


FIGURE 3 Block diagram of the CCCPE

The 2 shift-tap devices have 25 and 340 shift registers respectively to store the pixel data participated in computation of the SADs. One SAD computer can sum 25 absolute differences up produced by two 5 × 5 matching windows in a clock-period. Therefore, the CCCPE module can achieve 64 SADs from a template window in the right census image compared with 64 candidate windows in the left census image in a single clock. Figure 4 is the block diagram of the SAD computers. The SAD computer is constituted with 25 absolute difference calculators (AD) and a parallel adder with 25 inputs. The parallel adder has 4 pipelined stages architecture for improving the f_{max}.

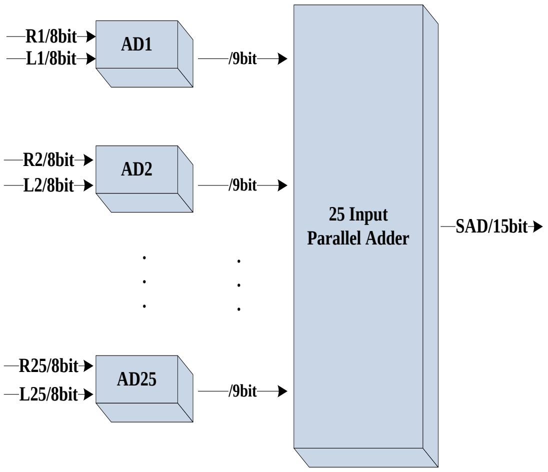


FIGURE 4 Block diagram of the SAD computer

The 64 SADs produced by CCCPE are sent to the disparity segregator (DS) and L-R checking module simultaneously. The DS calculates the minimum SAD using parallel comparators from 64 SADs and outputs the index number as the disparity of the right image, and the cost time is one clock period. Figure 5 is the block diagram of the disparity segregator. The L-R checking module produces the disparity of the left image and checks the right and left disparity to get the final valid one.

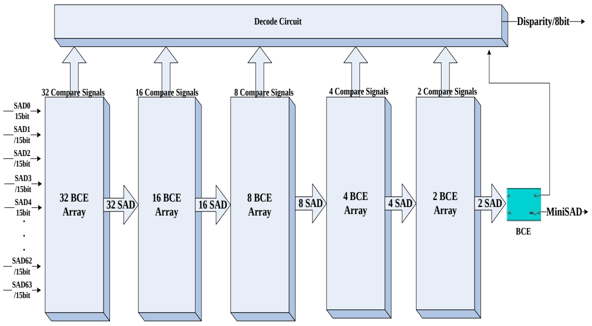


FIGURE 5 Block diagram of the disparity segregator

The final disparities are pushed into the DCFIFO connected with the L-R checking module. The DCFIFO's function in here is similar with the one mentioned in previous. A custom DMA engine followed with read port of the DCFIFO is responsible for writing the final disparities to the disparity table stored in the off-chip DDRII SDRAM.

The work procedure of L-R checking module can be separated into two parts. The one is calculating the disparity map, which takes the left image as reference, and the other is comparing the right and left disparity map with the metric listed in (2):

$$\mathrm{d}_{p'}(x-\mathrm{d}_p(x,y),y)=\mathrm{d}_p(x,y)\,, \qquad (2)$$

where $\mathrm{d}_p(x,y)$ and $\mathrm{d}_{p'}(x,y)$ are the disparities of the p and p', which are pixels in the left and right image separately. If the $\mathrm{d}_p(x,y)$ and $\mathrm{d}_{p'}(x,y)$ make the equation (2) true, the $\mathrm{d}_p(x,y)$ is a valid disparity, otherwise we should replace it with the valid disparity near to it.

The principle of producing the left image disparity in L-R checking module is showed in Figure 6. The number pair (L/R) displayed in the block in Figure 6 means that the window L in the left image matches with the window R in the right image at the same horizon line. We can get a conclusion that the 64 blocks at the vertical direction of Figure 6 are the results produced by matching the right image as the reference with the left image. From the every column, we can get a disparity of the right image. Along the diagonal of Figure 6, we can get the disparity of the left image. Therefore, we can get both right and left disparities by establishing only one stereo matching algorithm circuit.

63/0	64/1	65/2	······	126/63
62/0	63/1	64/2	······	125/63
61/0	62/1	63/2	······	124/63
⋮	⋮	⋮	······	⋮
0/0	1/1	2/2	······	63/63

FIGURE 6 The principle of producing the left image disparity

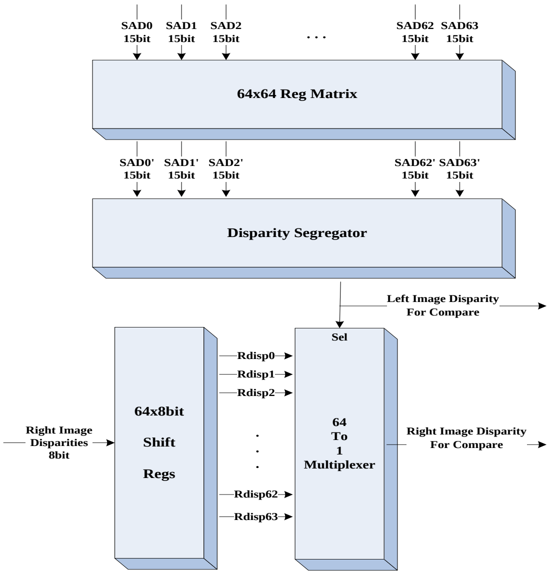


FIGURE 7 Diagram of the L-R checking module

The hardware diagram of the L-R checking module is showed in Figure 7. At the beginning of the L-R checking module is a 64 × 64 register matrix in unit of byte. It consists of 64 shift-taps with 64 taps each, and accepts 64 SADs produced by CCCPE in every clock. The register matrix is fully initiated with 4096 SADs, and then outputs buffered SADs along the diagonal of the matrix to the disparity segregator to calculate the left image disparity. So when first disparity of the left image is calculated, there were 64 right image disparities have been calculated, and they are buffered into a shift-tap device with 64 taps. The shift-tap device outputs the all buffered right image disparities to a 64 to 1 multiplexer, which takes the left image disparity as the select signal. The output of the multiplexer is the right image disparity, which will compare with the left image disparity as listed in (2).

The disparities produced after L-R valid checking are the last results we need. They are pushed into a dual-clock FIFO and a DMA engine is responsible for writing them to the disparity table stored in the off-chip DDRII SDRAM. The dual-clock FIFO's function in here is similar with the one mentioned in CTU module.

3.3 CONTROL PRINCIPLES OF THE SSMU

The main management work of the SSMU module is listed below:

(a) Initialization and updating data of the DPRAM array;
(b) The CCCPE work process control;
(c) Getting the finial disparities and writing them back to the buffer memory.

There are three Finite State Machines, which are in charge of the process management of the SSMU. They take effect at the work positions "WP1", "WP2" and "WP3" which are marked in Figure 2. The FSMs are custom IPs and designed in Verilog HDL. The hardware modules modelled by DSP Builder have interfaces for controlled by the FSMs to work together. Figure 8 shows the FSM for data updating management of the DPRAM array.

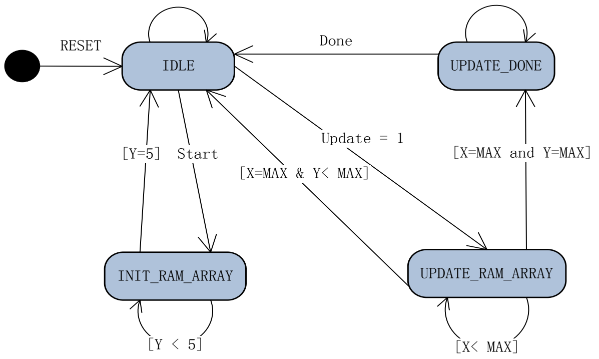


FIGURE 8 Finite state machine of the DPRAM array updating management

Figure 9 indicates the FSM for the computation control of the CCCPE.

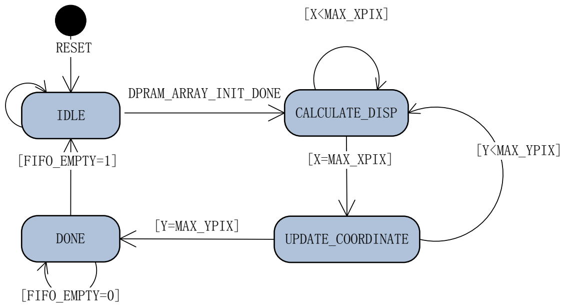


FIGURE 9 Finite state machine of the CCCPE control

The data updating in DPRAM array is activated by the start signal of the system. After the set action of the start signal, the FSM begins to read the raw data from the dual-clock FIFO if it is not empty, and send them to the DPRAM array according to the pixels' coordinate order. This state is not finished until 5 DPRAMs are fully initialized. Then the FSM comes into idle state. An update signal activates the FSM into the UPDATE_RAM_ARRAY state. The FSM continuously updates data using a unit in bytes to the DPRAM array. The FSM will not change into idle state until the X coordinate reaches maximum. The detail description about state machine of Figure 8 is listed as follows:

(a) After system reset, the FSM enters into the IDLE state automatically. In this state, the FSM waits for two signals to invoke the state transition. The one is a "Start" signal sent by the program executed in Nios II microprocessor to make the state change to INIT_RAM_ARRAY. The other is a "Update" signal sent from the FSM at "WP2" to make the state transit to UPDATE_RAM_ARRAY state.

(b) The task of the INIT_RAM_ARRY state is to initiate the DPRAM array with first 5 lines data of the stereo images, then return to the IDLE state to wait for update signal.

(c) Every time the FSM enters into the UPDATE_RAM_ARRAY state, a line of new pixel data will be transferred into the DPRAM array. After finishing the update work, the Y coordinate will be increased by one and checked if it reaches maximum. If Y equals to maximum, the state transits to UPDATE_DONE state, otherwise the state moves to IDLE state to wait for the next update signal.

(d) In UPDATE_DONE state, the FSM checks the "Done" signal of the system. If the signal is set, the FSM backs to IDLE state to wait for a new "Start" signal.

The work process control of the CCCPE is performed by the FSM working at "WP2" which is described in Figure 9. In the following list, the details of the tasks performed in each of the states are described:

(a) After system reset, the FSM enters into the IDLE state automatically. In this state, all variables are initialized.

(b) In IDLE state, the FSM waits for a DPRAM_ARRAY_INIT _DONE signal to transit into CALCULATE_DISP state. This signal is sent by the FSM at "WP1".

(c) In CALCULATEDISP state, the FSM reads the data from the DPRAM array to the CCCPE uninterruptedly till the pixel's horizontal coordinate of current line reaches maximum. An updated signal is also sent to notify the FSM at "WP1" to start updating a new line data to the DPRAM array. The update signal of the right DPRAM array is sent after 68 pixels of the left image were sent to the CCCPE for avoiding collision. The update signal of the left DPRAM array is sent after 5 pixels of the left image were sent to the CCCPE for avoiding collision. When entire 5 lines of data are sent, the state comes into the UPDATE_COORDINATE state.

(d) In the UPDATE_COORDINATE state, the pixel's vertical coordinate is updated. There are two exits: (1) if the vertical coordinate is smaller than the maximum, the FSM transits to CALCULATE_DISP state; (2) if the vertical coordinate is equal to the maximum, the DONE state is the next state.

(e) In DONE state, the FSM scans the disparity FIFO's empty signal. If all disparities are written back to the DDRII SDRAM, the FSM becomes idle again.

4 Results and discussion

The stereo matching circuit has been realized in Altera Cyclone III EP3C120f789 FPGA board shown in Figure 10. Table 1 and Table 2 show the resources required from the FPGA device in order to implement the designs presented in this paper.

FIGURE 10 The Cyclone III development board

TABLE 1 Resources needed for the implementation of the algorithm (with L-R checking module)

Device :Altera EP3C120F780C7N (Cyclone III device family)		
Resource		
Total logic elements	98598 / 119088	(83%)
Total combinational functions	89145 / 119088	(75%)
Dedicated logic registers	51099 / 119088	(43%)
Total registers	51328	
Total pins	151 / 532	(28%)
Total memory bits	530216 / 3981312	(3.3%)
Embedded Multiplier 9-bit elements	12 / 576	(2%)
Total PLLs	2 / 4	(50%)

TABLE 2 Resources needed for the implementation of the algorithm (without L-R checking module)

Device :Altera EP3C120F780C7N (Cyclone III device family)		
Resource		
Total logic elements	93242 / 119088	(78%)
Total combinational functions	86012/ 119088	(72%)
Dedicated logic registers	46765 / 119088	(39%)
Total registers	46994	
Total pins	151 / 532	(28%)
Total memory bits	448296 / 3981312	(12%)
Embedded Multiplier 9-bit elements	12 / 576	(2%)
Total PLLs	2 / 4	(50%)

The quality measures proposed by [11] are based on known ground truth data $d_T(x,y)$ offered by Middlebury Stereo datasets were used for evaluation. The percentage of bad matching pixels is computed with respect to some error tolerance δ_d :

$$M = \frac{1}{N}\sum_{x,y}(|\,d_c(x,y) - d_T(x,y)\,| > \delta_d)\ , \qquad (3)$$

where $d_c(x,y)$ is a disparity map produced by the proposed hardware .

The ground truth image has a disparity range from 0 to 59. The disparity range of our design is 0 to 63. So a disparity error tolerance δ_d = 1 to 4 is used. The measures are computed over the whole disparity map, excluding image borders, where part of the image is totally occluded.

Several tests have been performed. In the Table 3, the disparity maps are produced without L-R checking module. In the Table 4, the examples of disparity maps obtained with the L-R Checking module. Disparities are all encoded using a scale factor of 4 for grey levels 0 to 252.

TABLE 3 Evaluation result of the proposed system using Middlebury stereo images (without L-R checking module)

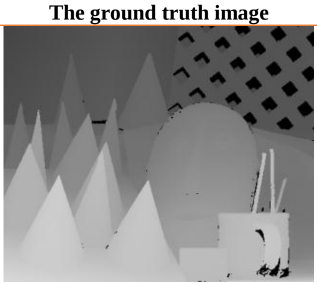
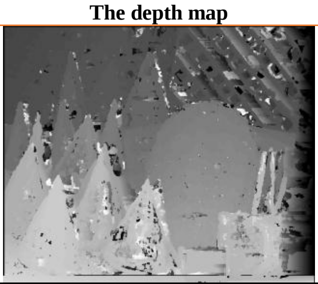
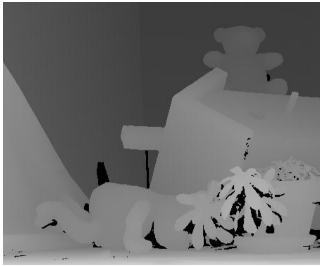
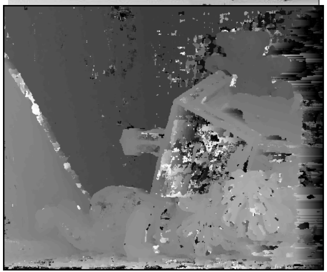
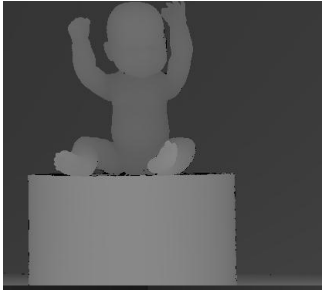
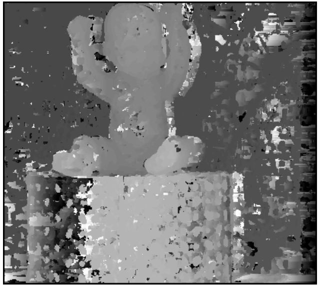
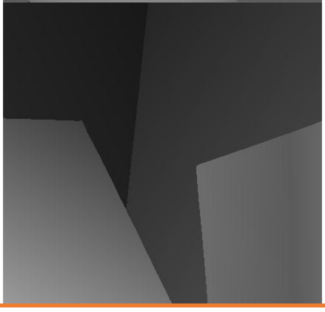
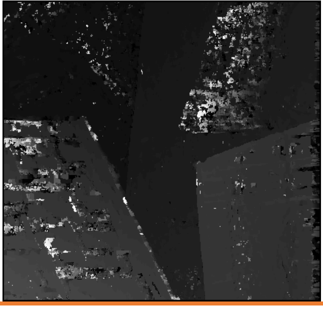
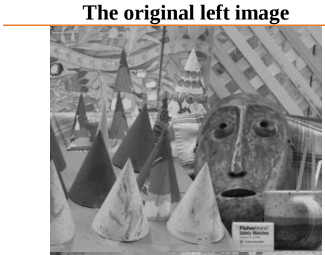
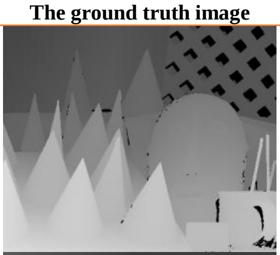
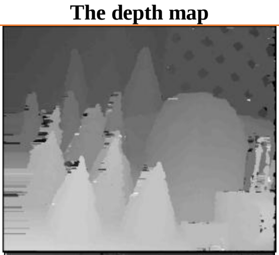
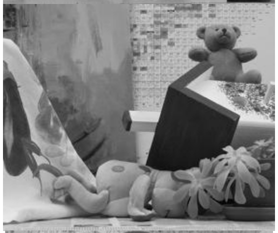
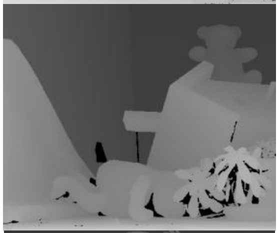
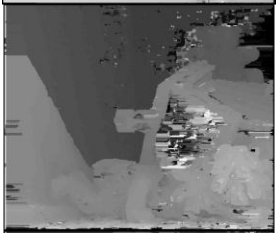
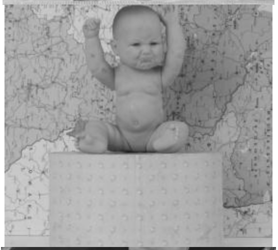
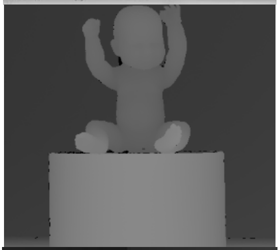
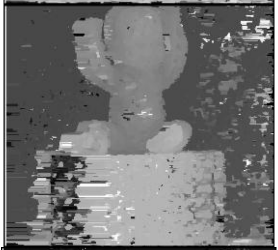
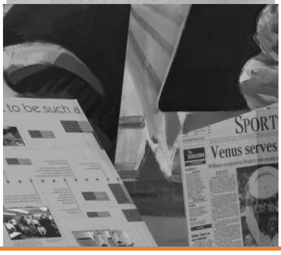
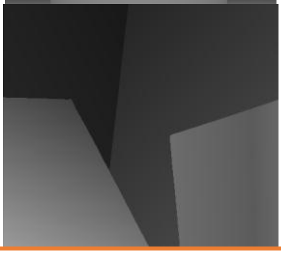
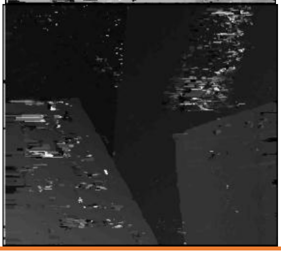
The original right image	The ground truth image	The depth map	Bad pixels
			25.3% (δ_d = 1) 22.15% (δ_d = 2) 19.33% (δ_d = 3) 17.09% (δ_d = 4)
			23.92% (δ_d = 1) 20.98% (δ_d = 2) 18.93% (δ_d = 3) 17.44% (δ_d = 4)
			31.39% (δ_d = 1) 28.25% (δ_d = 2) 26.44% (δ_d = 3) 25% (δ_d = 4)
			16.34% (δ_d = 1) 14.07% (δ_d = 2) 12.31% (δ_d = 3) 10.76% (δ_d = 4)

TABLE 4 Evaluation result of the proposed system using Middlebury stereo images (with L-R checking module)

The original left image	The ground truth image	The depth map	Bad pixels
			22.46% (δ_d = 1) 19.64% (δ_d = 2) 16.74% (δ_d = 3) 14.52% (δ_d = 4)
			23.39% (δ_d = 1) 20.6% (δ_d = 2) 18.75% (δ_d = 3) 17.40% (δ_d = 4)
			28.43% (δ_d = 1) 24.96% (δ_d = 2) 23.33% (δ_d = 3) 22.16% (δ_d = 4)
			14.31% (δ_d = 1) 12.56% (δ_d = 2) 11.28% (δ_d = 3) 9.78% (δ_d = 4)

In Table 5, the proposed system is compared to the existing approaches in terms of speed.

TABLE 5 Speed comparison of the stereo matching implementations

Authors	Frame Rate	Image Size	Max. Disp	Algorithm	Window Size	Platform
Motten *et al.* [10]	56 fps	800 × 600	80	SAD	7 × 7	1FPGA
Proposed impl.	30 fps	1396 × 1110	64	SAD	5 × 5	1 FPGA
Software impl. [12]	2.55 fps	320 × 240	100	SAD	3 × 3	PC
Kalomiros *et al.* [13]	162 fps	640 × 480	64	SAD	3 × 3	1FPGA + PC
Niitsuma *et al.* [14]	30 fps	640 × 480	27	SAD	7 × 7	1 FPGA
Miyajima *et al.* [15]	18.9 fps	640 × 480	80	SAD	7 × 7	1FPGA + PC

5 Conclusions

An efficient hardware implementation of a real-time stereo matching algorithm is proposed for the calculation of disparity maps. It takes full advantage of the convenience of IP reuse based on SoPC architecture. The frame rate could enable real time performance at the resolution of 1396×1110. The architecture of our system is very promising and may get better in the future. The system has been implemented on static image input from C code in the Nios II processor. We plan to incorporate live stereo video streams and combine the algorithm with pre-stage to make it more suitable for the operation in robot auto- navigation and visual servo applications.

Acknowledgments

This work has been cooperatively supported by the projects as follows:
(a) National Nature Science Foundation of China (No. 61202093);
(b) National Key Technology R&D Program of China (No. 2014BAF07B01);
(c) 3D Innovation team of Zhejiang provincial.

References

[1] Bertozzi M, Broggi A 1998 *IEEE Transactions on Image Process* **7**(1) 62-81

[2] Uchida N, Shibahara T, Aoki T, Nakajima H, Kobayashi K 2005 3-D face recognition using passive stereo vision *Proceedings of IEEE International Conference on Image Process* 950-3

[3] Bensrhair A, Bertozzi M, Broggi A, Fascioli A, Mousset A, Toulminet G 2002 Stereo vision-based feature extraction for vehicle detection *IEEE Intelligent Vehicle Symposium 2* 465-70

[4] Brown Z M, Burschka D, Hager D G 2003 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(8) 993-1008

[5] Yi J, Kim J, Li L, Morris J, Lee G, Leclercq P 2004 *Lecture Notes in Computer Science* **3189** 309–20

[6] Perri S, Colonna D, Zicari P, Corsonello P 2006 SAD-Based Stereo Matching Circuit for FPGAs *Proceedings of the 13th IEEE International Conference on Electronics, Circuits and Systems* 846–9

[7] Darabiha A, Rose J, MacLean W J 2003 Video-Rate Stereo Depth Measurement on Programmable Hardware *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 203–10

[8] Gardel A, Montejo P, Garcia J, Bravo I, Lázaro J 2012 *Sensors* **12** 1863–84

[9] Georgoulas C, Andreadis I 2011 *J. Real Time Image Process* **6** 257–73

[10] Motten A, Claesen L 2011 Low-Cost Real-Time Stereo Vision Hardware with Binary Confidence Metric and Disparity Refinement *Proceedings of the 2011 International Conference on Multimedia Technology* 3559–62

[11] Scharstein D, Szeliski R 2002 *International Journal of Computer Vision* **47**(1-3) 7-42

[12] Ambrosch K, Humenberger M, Kubinger W, Steininger A 2007 Hardware Implement of an SAD Based Stereo Vision Algorithm *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1–6

[13] Kalomiros J, Lygouras J 2009 Comparative study of local SAD and dynamic programming for stereo processing using dedicated hardware *EURASIP J. Adv. Signal Process* doi:10.1155/2009/914186

[14] Niitsuma H, Maruyama T 2004 *Lecture Notes in Computer Science* **3203** 1155–7

[15] Miyajima Y, Maruyama T 2003 *Lecture Notes in Computer Science* **2778** 448–57

Authors

Xiang Zhang, born in 1979

Current position and grades: MSc, a lecturer at Hangzhou Dianzi University in China.
University studies: BSc and degrees in computer science, Hangzhou Dianzi University, Hangzhou City, China, in 2001 and 2004, respectively. He is currently working toward a Ph.D. degrees with the State Key Laboratory of Fluid Power Transmission and Control, Zhejiang University.
Research interests: computer vision, image parallel processing, embedded systems, and real-time applications

Huaixiang Zhang, born in 1978

Current position and grades: PhD, Department of Computer Science, Hangzhou Dianzi University, since March 2007
University studies: BSc in Mechanical Engineering from North China University of Technology in 2000, MSc in Automatic Control from Beijing Institute of Technology in 2003, and PhD in Automatic Control from Institute of Automation, Chinese Academy of Science in 2007
Research interests: motion control theory in mobile robot, non-linear and adaptive control, electrical motors and artificial intelligence

Yifan Wu

Current position and grades: Assistant Professor at Hangzhou Dianzi University in China
University studies: BSc and BSc degrees in Control Science and Engineering from Zhejiang University in 2003 and 2006, respectively, and the Ph.D. degree in Computer Engineering from Scuola Superiore Sant'Anna of Pisa in January 2010, a visiting student at the Department of Automatic Control, Lund University, 2009.
Research interests: real-time control systems, scheduling algorithms, and resource management

Study on improved LLE algorithm based on a sample set of well-distributed and weights matrix

Song Fei[*], Cui Zhe

Chengdu Computer Institute of Chinese Academy of Science

Received 1 March 2014, www.tsi.lv

Abstract

There are large amounts of data has accumulated along with technology of computer, information and network developed. How can we using these data and mining out the valuable information are hot topics in information processing field. There are some distress and difficulties caused by the high-dimensional data on data modelling and data analysis. In this paper, a local linear embedding algorithm based on the improved uniform sample set and the weight value matrix is proposed. The test shows that the improved dimensionality reduction algorithm accuracy is significantly higher than the original LLE algorithm.

Keywords: LLE, Well-distributed, Weights Matrix

1 Introduction

Big data cause a lot of inconvenience on applications because of the diversity and complexity of the data. These rich data resources bring convenience to people but have also brought a lot of problems at the same time. Such as Information overload, Data is difficult to choose and useful information submerged in the massive amounts of data, etc. Faced with these difficult to deal with data, we cannot mining out the effective information implied in big data and cannot speculate on future trends if we have no effective means for analysis and processing[1].

Data dimensionality reduction techniques [1-2] can be used to solve the above-mentioned problems. It can explore the internal structure and association of the original data, and eliminated redundant data, improved efficiency of computation. It can also improve the understand ability of the data to improve the accuracy of data [2].

2 Dimensionality reduction principle

Depending on the type of the data to be processed, existing dimensionality reduction algorithm [3-4] is divided into two categories: linear dimension reduction technique and nonlinear dimensionality reduction technique.

Define: Let $\mathbf{X}=(x_1,x_2,...,x_D)^{\mathrm{T}}$ is a vector in high-dimensional space, by the following formula:

$$\mathbf{F}(\mathbf{X})=\begin{pmatrix}F_1(\mathbf{X})\\F_2(\mathbf{X})\\\cdots\\F_d(\mathbf{X})\end{pmatrix}=\begin{pmatrix}F_1(x_1,x_2,...,x_D)\\F_2(x_1,x_2,...,x_D)\\\cdots\\F_d(x_1,x_2,...,x_D)\end{pmatrix}.$$

We can get a vector $\mathbf{Y}=(y_1,y_2,...,y_D)^{\mathrm{T}}$ in low-dimensional space. If each component F_i of $\mathbf{F}$ is a linear function, then $\mathbf{F}$ is a linear dimensionality reduction. Otherwise, $\mathbf{F}$ is nonlinear dimensionality reduction.

3 Locally linear embedding algorithm LLE analyses

Locally linear embedding algorithm that proposed by Roweis and Saul [5] in 2000. It is a method for nonlinear dimensionality reduction. Its core idea is using local linear approaching global nonlinear, keep the geometry structure of local sample points unchanged and using local neighbourhood data that overlapped with each other to provide global information. So as to maintain the overall geometric properties of whole sample points [3].

Assuming that the sample set consists of N D-dimensional vector X_i and the entire sample points in a $d(d \le D)$ -dimensional manifold. When we have a sufficient number of sample points, we can think approximately that each sample point and its adjacent sample points in a local linear manifold. In this case, each sample point can be represented with a linear combination of its adjacent sample points that possessed weighting coefficient [4]. These weighting coefficients reflect the local geometry information of a small area. We can use this information to seek out a low-dimensional embedding space that keeps the geometrical characteristics of the original high-dimensional space. We can get the overall information of the original high-

[*] ***Corresponding author*** e-mail: asfei@aliyun.com

dimensional space by the overlapping local neighbourhood and get a global coordinate system [5]. The greatest advantage of this method is that the data does not produce a large offset. That because of the curvature of the manifold is very small so it can be approximated as flat.

LLE using matrix $X_{D,N}$ as it input and matrix $Y_{d,N}$ as it output. Here $X_{D,N}$ and $Y_{d,N}$ are composed by *N* d-dimensional vector (*d*<<*D*) and the *K*th column of matrix Y corresponding to K-th column of matrix *X*.

The algorithm is divided into three steps.

1) Looking for a point and its adjacent point, constitute a piece of local adjacent area. For a sample point X_i (i=1,2,…,N) that in the high-dimensional space, we can calculate the distance between X_i and the other N-1 sample point. According to their distances, we can find a close neighbour points of X_i. We usually measure the distance between two points using the Euclidean distance, that is $d_{ij} = \|X_i - X_j\|$. There are two ways for choosing nearest neighbour points: Choosing the K points that have minimum distance from X_i as the adjacent points of X_i. All the points that X_i as the centre and all the points within the sphere of radius ε are the adjacent points of X_i. We generally use the first method to determine the adjacent points. Each point in the space have the same number of neighbour points due to .the number of K is defined. Thus, calculate getting more simple and convenient.

2) Calculate the weight of X_i and its nearest neighbour points. Weight values describe the degree of approximation between the two points. When we defined X_i and found its K adjacent points, we need to calculate the weights between this point and each of its adjacent point. Assume that X_j is a close neighbours point of the of X_i, then the weight between them is $W_{ij} = e^{\frac{|x_i - x_j|}{2\sigma^2}}$, wherein σ is a parameter. The calculated weight W_{ij} that between X_i and each of its adjacent points, and let the error minimum when X_i is reconstructed by this *K* points, then: $\min \varepsilon(W) = \sum_{i=1}^{N} \left\| X_i - \sum_{j=1}^{N} w_{ij} X_j \right\|$.
In order to ensure translational invariance, make $\sum_{j=1}^{N} w_{ij} = 1$. If we want to make $\varepsilon(W)$ minimum, then: $\varepsilon(W_i) = \left\| X_i - \sum_{j-1}^{N} w_{ij} X_j \right\|^2 = \sum w_{ij} w_{ik} c_{jk} = W_i' C W_i'^T$, where $W_i' = (w_{i1}', w_{i2}', w_{i3}' \cdots w_{iK}')$ are *K* components that value are not 0 of W_i. Now the loss function can be rewritten as: $\min \varepsilon(Y) = \sum_{i=1}^{N} \sum_{j=1}^{N} M_{ij} y_i^T y_j$.
Construct a local covariance matrix C: $c_{jk} = (X_i - X_j)^T (X_i - X_k)$.
Set $W_i' l = 1$, $l = \underbrace{(1,1,1 \cdots 1)}_{K}^T$, by the Lagrange multiplier method to obtain: $\min L = W_i' C W_i'^T + \lambda (W_i' l - 1)$,
$\frac{\partial L}{\partial W_i'} = 2W_i' C + \lambda l^T = 0 \quad ===> W_i' = \frac{l^T C^{-1}}{l^T C^{-1} l}$.
When *K*>*D*, *C* is a singular matrix. Therefore, we need a renormalization operation for it. We sum regular numbers on the diagonal of the singular matrix: $C_{jk} \leftarrow C_{jk} + \delta_{jk} (\frac{\Delta^2}{K}) Tr(C)$, where $Tr(C)$ is the trace of *C*, $\Delta^2 << 1$.
$C_{jk} \leftarrow C_{jk} + \delta_{jk} (\frac{\Delta^2}{K}) Tr(C)$, where $Tr(C)$ is the trace of *C*, $\Delta^2 << 1$.

3) Calculated points Y_i in low-dimensional space. The final step of LLE is a calculating of the value of low-dimensional embedding space according to samples X_i in the high-dimensional space and the weight values W_{ij}. W_{ij} is the weight value between X_i and its adjacent point X_j. To make the low-dimensional space as much as possible be consistent with the partial linear structure in the high-dimensional space, a local information W_{ij} should be fixed. We should minimize a loss function $\phi(Y) = \sum_{i=1}^{N} \left\| Y_i - \sum_{i=1}^{N} w_{ij} Y_j \right\|^2$, taking into account requirements for $\phi(Y)$ without deformation, translation, rotation and scaling transformation. Thus, $\sum_{i=1}^{N} Y_i = 0$ and $\frac{1}{N} \sum_{i=1}^{N} Y_i Y_j^T = 1$. So, $\phi(Y) = \sum_{i=1}^{N} \left\| Y_i - \sum_{i=1}^{N} w_{ij} Y_j \right\|^2$,
$= tr((Y - WY)^T (Y - WY)) = tr(Y^T M Y)$, wherein M is an $N * N$ stacked matrix: $M = (1 - W)^T (1 - W)$.

At this moment, the minimization solution of a loss function is eigenvectors matrix that consisted by several minimum eigenvalues of matrix M. We take the non-feature vector that corresponding m eigenvalues of matrix M according to the order of small to large. Because the smallest eigenvalue is infinitely close to 0, therefore we discarded this eigenvector so as to satisfy the condition

$\sum_{i=1}^{N} Y_i = 0$. The rest of 2 to m +1 feature vectors will compose a matrix. This matrix is the samples in the low-dimensional space. LLE algorithm processes are shown in Figure 1.

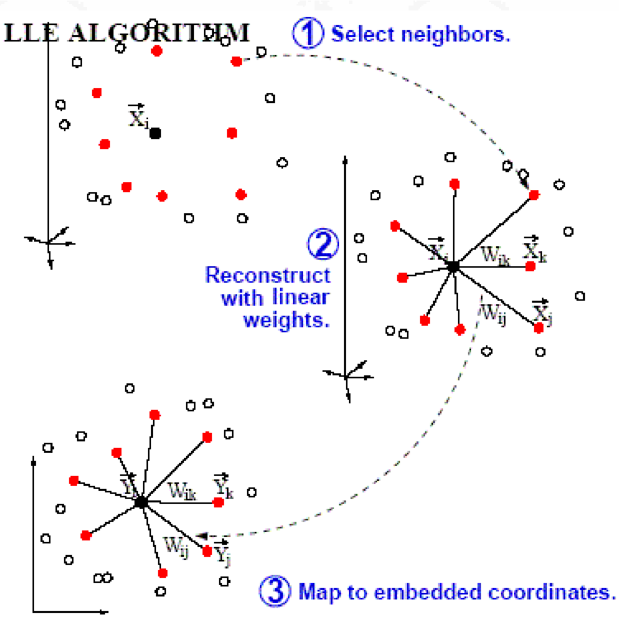


FIGURE 1 LLE algorithm processes

4 The improved LLE algorithm - DLLE

The data that correspond to LLE algorithm is static. That is every time it will enter the entire sample set and then mapped sample set to a low-dimensional embedding space and get the corresponding samples from the embedding space finally. When a new sample point (the new data) is added, the new sample point and the original sample must be merged into a new sample set. The new sample set will be re-entered into LLE algorithm and be running.

K value of LLE algorithm indicates the quantity of selected adjacent sample points. LLE is very sensitive on *K*. The larger the value *K*, the greater the difference of the geometric characteristics of high-dimensional space manifold. This will be lead to the range of data mining conclusions greater and lead to reduced accuracy. Eventually. But, if the value of *K* is not big enough (i.e, the number of adjacent points is not enough), then this may cause the continuous manifold in high-dimensional space split into disconnected submanifold. Obtained conclusions by data mining can be completely unrelated with expected ones. References [6-8] have analysed how to select the range of *K* value. For the large number of data, the ranges of *K* value are between 5 and 20 generally.

In the really big data mining process, we found that the different component values have different effects on mining conclusions. This means that the weights values of each component importance are not the same and the different values will have a huge impact on the actual results. We propose an improved algorithm to solve the considered problems.

In order to reduce the sensitivity of the LLE algorithm on the value of K, we design a new method for definition of distance, where the average of the distance between a central point and its *K* neighbouring points is represented.

This new approach allows to obtain the distance between the samples in the sample-point-intensive areas (relatively increased) and the distance between the samples in the sample point’s sparse area (relatively narrow). Thus, the distribution of the sample set leads to homogenizing and reducing of the impact on the K value for LLE calculation result [6].

For solution of the problem of weights values between components, we present an idea of embedded importance weights into LLE algorithm.

Indicating the original sample set, the paper contains N samples, setting an importance weight vectors. Each component is a positive number and the sum of all the components is equal to 1. Calculating a new eigenvector using LLE algorithm of original sample set, we can get the weight matrix and the sample, which in embedding space [7].

Making the neighbours of all the sample points in the original sample set non-changed, we can get the weight matrix of the samples space. Finally, by step 3 (section 3), we obtain the sample set of low-dimensional space, realizing data dimensionality reduction.

5 Experiments and conclusions

Randomly select a UCI database as experimental data to compare the effect of the improved DLLE algorithm and traditional LLE algorithm. This database contains a training set and a test set. 2000 samples from the training set to do the training object and take 1000 samples from the test set to do the test were taken. The principle of nearest neighbour to look for the K-nearest neighbours of the training set for each test sample was used. The range of dimension d in embedding space is from 2 to 10 and the range of K is from 2 to 10. Two experiments use the same training samples and the same test samples. The result of the experiment is as follows tables.

Based on the data in Table 1 and Table 2 can be seen that the error rate of improved LLE algorithm lower than the error rate of original LLE algorithm and the accuracy of dimensionality reduction higher than the original LLE algorithm obviously.

TABLE 1 The average error rate of the original LLE algorithm

	2	**3**	**4**	**5**	**6**	**7**
10	0.192	0.124	0.118	0.092	0.082	0.076
11	0.212	0.136	0.112	0.084	0.072	0.066
12	0.256	0.182	0.158	0.108	0.098	0.092
13	0.328	0.174	0.144	0.120	0.116	0.110
14	0.282	0.168	0.138	0.104	0.108	0.098
20	0.386	0.364	0.332	0.324	0.302	0.284

TABLE 2 The average error rate of the improved LLE algorithm

	2	**3**	**4**	**5**	**6**	**7**
10	0.142	0.118	0.084	0.080	0.078	0.062
11	0.164	0.130	0.096	0.088	0.074	0.068
12	0.208	0.166	0.104	0.094	0.082	0.076
13	0.220	0.158	0.118	0.106	0.098	0.104
14	0.234	0.144	0.126	0.104	0.102	0.086
15	0.246	0.182	0.154	0.126	0.116	0.104
16	0.262	0.262	0.186	0.148	0.124	0.108
17	0.266	0.294	0.242	0.186	0.158	0.132
18	0.274	0.308	0.278	0.230	0.212	0.188
19	0.280	0.314	0.306	0.264	0.232	0.206
20	0.298	0.344	0.312	0.272	0.238	0.220

References

[1] Yan S, Bouaziz S, Lee D, Barlow J 2012 *Neurocomputing* **76**(1) 114-24

[2] Huang W, Yin H 2012 *Image and Vision Computing* **30**(4) 355-66

[3] Cho J H, Kurup P U 2011 *Sensors and Actuators* B: *Chemical* **160**(1) 542-8

[4] Gu H, Deng S P, Wang X, Shi Jin-Qin, Jin Jian-Qiu, Tian Shi-Yi 2012 *Sensors and Actuators* B: *Chemical* **163**(1) 281-9

[5] Roweis S T, Saul L K 2000 *Science* **290**(5500) 2323-6

[6] Xing E P, Ng A Y, Jordan M I, Russell S 2003 Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems* **15** 505-12 MA: MIT Press

[7] Tsai F S 2012 *Expert Systems with Applications* **39**(2) 1747-52

[8] Raymer M L, Punch W F, Goodman E D, Kuhn L A, Jain A K 2000 *IEEE Transactions on Evolutionary Computation* **4**(2) 164-71

Authors

Fei Song, born on October 31, 1982, Chengdu, Sichuan, China

Current position, grades : PhD student in Chengdu Institute of Computer Application, Chi Doctor
University studies: PhD studies Chengdu Computer Institute of Chinese Academy of Science
Scientific interest: Data Disaster Recovery, Reliability Engineering and Network coding
Publications : 4
Experience: Academy of Sciences, China. He has published more than four articles in reputed international journals and International Conferences
Research interests: Data Disaster Recovery, Reliability Engineering and Network coding

Zhe Cui, born on September 20, 1997, Chengdu, Sichuan, China

Current position, grades: Professor in Chengdu Institute of Computer Application, Chinese Academy of Sciences, China.
University studies: Chengdu Computer Institute of Chinese Academy of Science
Scientific interest: Trusted Computing, Embedded Systems and Reliability Engineering
Research interests: Trusted Computing, Embedded Systems and Reliability Engineering

The nonlinear vibration analysis of the fluid conveying pipe based on finite element method

Gongfa Li*, Jia Liu, Guozhang Jiang, Jianyi Kong, Liangxi Xie, Wentao Xiao, Yikun Zhang, Fuwei Cheng

College of Machinery and Automation, Wuhan University of Science and Technology, Wuhan, 430081, China

Received 1 March 2014, www.tsi.lv

Abstract

A Coupling between the fluid and the structure existed almost in all industrial areas the vibration of fluid solid coupling for fluid conveying pipe was called the "dynamics of typical"[1], Because of the physical model and mathematical description for the fluid conveying pipe was simple, especially it was easy to design and manufacturing, according to the characteristics of fluid conveying pipe, transformed the transverse vibration of the fluid conveying pipe to the beam element model of two nodes. Using Lagrangian interpolation function, the first order Hermite interpolation function and the Ritz method to obtain the element standard equation, and then integrated a global matrix equation. Used the mode decomposition method, obtained the vibration modal of the fluid conveying pipe with Matlab programming. The vibration modal of the fluid conveying pipe in four kinds of boundary conditions was analysed. The characteristics of pipes conveying fluid was obtained which the pipeline system parameters under different boundary constraints. To provide the theoretical support for the research of vibration attenuation of fluid conveying pipes.

Keywords: Fluid solid coupling, Nonlinear vibration, Modal analysis, Interpolation, The finite element algorithm

1 Introduction

A system of conveying fluid pipe was widely used in the city water supply and drainage, water power, chemical machinery, aerospace, marine engineering and the nuclear industry and other fields, it was play an important role for improving the living standards of the nation and the national economic strength. However, according to statistics, in the industrial production, the damage of the water hammer in pipeline interface and pipeline rupture accounted for over 75% in the total system failure rates, seriously affected the normal production and operation, resulting in huge economic losses. Coupling between the fluid and the structure is almost exist in all industrial areas, the fluid and solid coupling vibration of pipeline flow is called the typical "dynamic" [2], because of its simple physical model and mathematical description, especially the pipeline is easy to design and manufacture, which provides convenience to the coordinated development of the theoretical and experimental research. But the pipeline, as the application extremely widely, of the coupling vibration of pressure flow is the most representative in this field, which has a broad background in engineering applications, and a very high theoretical research value and practical significance, but also has many challenges [3].

2 The establish of mathematical model for the output response of the fluid conveying pipe

Taking into account the pipe ratio for length to diameter is relatively large, the deformation of radial is the same, just only exists a certain angle difference, which can be regarded the pipe as plane beam element to consider, using two node element, as shown in Figure 1, the node number of I and J. The conveying fluid pipe is only affected by the lateral force, no axial force, so analysis with the two node element, the nodal displacement model can be defined as [4]:

$$y(t) = [y_i, \theta_i, y_j, \theta_j]^T \,. \tag{1}$$

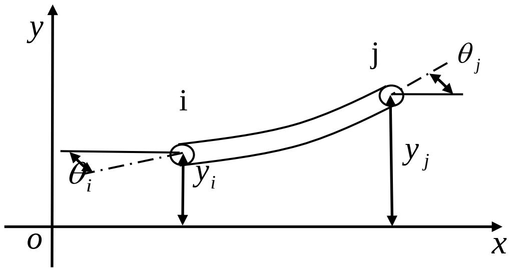


FIGURE 1 The deformation of two node element

[*] ***Corresponding author-*** E-mail: ligongfa@aliyun.com

In the node parameter of unit, in addition to the node value of field function, also contains node value for a derivative $\partial y/\partial x$ of the field function. In order to maintain the continuity of field function derivative between the public node element, and in the end nodes to keep the derivative order is first for the field function, so the first-order Hermite interpolation polynomial is used [5]:

$$N(\xi)=[N_1(\xi),N_2(\xi),N_3(\xi),N_4(\xi)], \tag{2}$$

where

$$N_1(\xi)=1-3\xi^2+2\xi^3,$$
$$N_2(\xi)=\xi-2\xi^2+\xi^3,$$
$$N_3(\xi)=3\xi^2-2\xi^3,$$
$$N_4(\xi)=\xi^3-\xi^2.$$

When the local dimensionless coordinate is taken, the ξ is get $0\le\xi\le1$.

Using the Ritz method interpolation function to establish standard unit equation of the approximate solution of the lateral vibration we find the interpolation function:

$$[M^e]\{\ddot{y}_j\}+[C^e]\{\dot{y}_j\}+[K^e]\{y_j\}=[Q^e], \tag{3}$$

where

$$[M^e]=[M^e_{ij}],[C^e]=[C1^e_{ij}]+[C2^e_{ij}],$$
$$[K^e]=[K1^e_{ij}]+[K2^e_{ij}]+[K3^e_{ij}],$$
$$[Q^e]=[Q1^e_{ij}]+[Q2^e_{ij}]+[Q3^e_{ij}],$$
$$[M^e_{ij}]=\int_l N_i(m_p+m_f)N_j dx,$$
$$[C1^e_{ij}]=\int_l N_i(2m_f v_f)\frac{\partial N_j}{\partial x}dx,$$
$$[C2^e_{ij}]=\int_l N_i(\frac{A_f}{c^2}\frac{\partial P}{\partial t})N_j dx,$$
$$[K1^e_{ij}]=\int_l \frac{\partial^2 N_i}{\partial^2 x}(EI)\frac{\partial^2 N_j}{\partial^2 x}dx,$$
$$[K2^e_{ij}]=-\int_l \frac{\partial N_i}{\partial x}(m_f v_f^2+(1-2\gamma)A_f P)\frac{\partial N_j}{\partial x}dx,$$
$$[Q1^e_{ij}]=\int_l(-m_f\frac{\partial v_f}{\partial t})N_j dx,$$
$$[Q2^e_{ij}]=\int_l(-\frac{v_f A_f}{c^2}\frac{\partial P}{\partial t})N_j dx,$$
$$[Q3^e_{ij}]=\int_l(m_p g+m_f g)N_j dx.$$

2.1 THE ENTIRETY MATRIX

There are some matrix must be appropriately expanded rewrite when the unit matrix integrated to the entirety matrix so that the matrix of all elements with uniform format, then according to the superposition to assembly [6].

The usually study boundary constraint conditions include fix to hinge and fix to fix constraints, the mathematical expression of its boundary is given below, respectively: (I), (II), (III), (IV):

$$\begin{cases} y(0,t)=0,\left.\frac{\partial y}{\partial x}\right|_{x=0}=0 \\ y(L,t)=0 \end{cases}, \tag{I}$$

$$y(0,t)=0,\left.\frac{\partial y}{\partial x}\right|_{x=0}=0, \tag{II}$$

$$y(0,t)=0,\left.\frac{\partial y}{\partial x}\right|_{x=0}=0, \tag{III}$$

$$\begin{cases} y(0,t)=0,\left.\frac{\partial y}{\partial x}\right|_{x=0}=0 \\ y(L,t)=0,\left.\frac{\partial y}{\partial x}\right|_{x=L}=0 \end{cases}. \tag{IV}$$

The four kinds of boundary conditions of above given all belong to the first class constraint conditions, for this kind of constraint conditions can usually use "row row column method" and "multiplied with bigger number method". The "multiplied with bigger number method" is make the main diagonal element about the specified node displacement in the overall stiffness matrix with multiply by the large number λ, at the same time, give the specified value of node displacement to the corresponding element of load matrix [7], then multiply by the same number as well as the main diagonal elements. Using the “multiplied with bigger number method ‘to deal with the boundary constraint condition by’, finally forms the whole matrix:

$$[M]\{\ddot{y}\}+[C]\{\dot{y}\}+[K]\{y\}=[Q]. \tag{4}$$

2.2 THE MODAL SOLUTION OF FLUID CONVEYING PIPES BASED ON MODE-SUPERPOSITION METHOD

Solving the modal of the system when the movement of the global matrix is obtained, we select the mode-superposition method for the model. Solving the inherent frequency and vibration type of fluid conveying pipes it is usually divided into two cases about damped and non-damped. The non-damping case is solved in the real domain, and the damping case is solved in the complex domain.

1) The free vibration equation without consider the damping:

$$M\ddot{y}(t)+Ky(t)=0\,, \tag{5}$$

using a solution:

$$y=\phi\sin\omega(t-t_0)\,, \tag{6}$$

where ϕ is the n order vector, ω is the vibration ϕ frequency vector, t is the time variable, t_0 is the time constant that determined by the initial conditions.

Using (6) and (5) we can get the generalized eigenvalue, i.e.:

$$K\phi-\omega^2 M\phi=0\,. \tag{7}$$

According to the general solution of eigenvalues and eigenvectors to the identified ϕ and ω, the results can be obtained as n characteristic solutions $(\omega_1^2\,,\varphi_1)$, $(\omega_2^2\,,\varphi_2)$,…, $(\omega_n^2\,,\varphi_n)$, which the characteristic values $\omega_1\,,\omega_2\ldots\omega_n$ for the natural frequency of the conveying fluid pipes system take place. There is also the ordering $0\le\omega_1<\omega_2<\cdots<\omega_n$ and the eigenvectors $\phi_1,\phi_2,\cdots,\phi_n$ for N inherent vibration feature vector, which corresponds to the inherent natural frequency.

2) The free vibration equation with a damping:

$$M\ddot{y}(t)+C\dot{y}(t)+Ky(t)=0\,. \tag{8}$$

A solution of the natural frequency and vibration type equations in the damped, early obtained the natural frequency and vibration model, which corresponds to the free damped vibration equation, gives:

$$\Phi=[\phi_1,\phi_2,\cdots,\phi_n]\,, \tag{9}$$

$$\Omega=diag[\omega_1,\omega_2,\cdots,\omega_n]\,, \tag{10}$$

where the Φ is the vector based on the n natural modes without damping, the Ω is the vector based on the n natural frequency without damping. Then, using the method of truncated model for the ratios of model damping, we obtain: $\xi_1,\xi_2,\cdots,\xi_n$,

$$\Xi=diag[\xi_1,\xi_2,\cdots,\xi_n]\,. \tag{11}$$

Using the vibration equation of the generalized coordinates and the equation (8) we obtain:

$$\begin{aligned}&[\Phi]^T[M][\Phi][\Phi]^{-1}[\ddot{y}(t)]+[\Phi]^T[C][\Phi][\Phi]^{-1}[\dot{y}(t)]\\&+[\Phi]^T[K][\Phi][\Phi]^{-1}[y(t)]=0.\end{aligned} \tag{12}$$

The generalized mass matrix, the generalized damping matrix and generalized stiffness matrix, respectively are: $[M]_z=[\Phi]^T[M][\Phi]$, $[C]_z=[\Phi]^T[C][\Phi]$, $[K]_z=[\Phi]^T[K][\Phi]$.

After the generalized coordinate transformation $\{q\}=[\Phi]^{-1}\{y\}$ we obtain:

$$\ddot{q}+2\Xi\Omega\dot{q}+\Omega^2 q=0\,. \tag{13}$$

The equation (13) is a group of uncoupled equations in component form:

$$\ddot{q}_i+2\xi_i\omega_i\dot{q}_i+\omega_i^2 q_i=0\,,\ (\text{i}=1,2,\ldots\text{n}) \tag{14}$$

It is easy to obtain the complex eigenvalue and complex eigenvectors using the method of the generalized eigenvalues and then to obtain the natural vibration modes and natural frequencies in the damping case.

3 The effects of the system parameters on the conveying fluid pipe

The developed equation of transverse motion for conveying fluid pipes with the Matlab simulation software and all of the parameters in the simulation process are given in Table 1. During the simulation process, we make two end points of pipe as the supporting points assuming a rigid constrain with a room temperature water of the fluid and rolling copper as pipe material.

3.1 FOUR KINDS OF BOUNDARY CONDITIONS OF THE FIRST FOUR ORDER VIBRATION MODE

The first four order mode of vibration of the constraint of fix to hinge as shown in Figure 2. From Figure 2, the relative position of all nodes in the vibration pipes that reflect the inherent form vibration of the conveying fluid pipe in the constraint of fix to hinge. The first four-order vibration mode that corresponding to the first four order natural frequency of the conveying fluid pipe, the natural frequency of the first order vibration mode is the minimum and two order mode, three order mode, four order mode is increasing [8]. The constraints of fix to suspension extension and fix to hinge does not belong to the same class, so the vibration mode is also different, as shown in Figure 3. For the same boundary, damped and no damped vibration mode is basically the same.

3.2 FOUR KINDS OF BOUNDARY OF VIBRATION MODE OF DEFLECTION

For different boundary constraints, a vibration mode of deflection for conveying fluid pipe is not identical [9]. The constraint of fix to hinge is shown in Figure 4, the constraint of hinge to hinge is shown in Figure 5, the constraint of fix to fix is shown in Figure 6, the constraint of fix to suspension extension is shown in Figure 7.

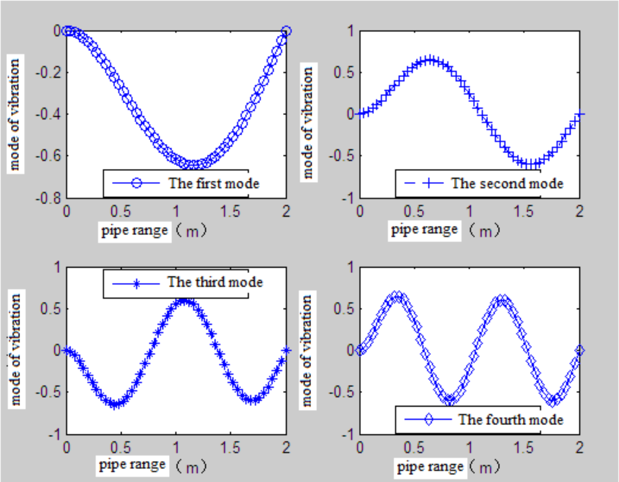


FIGURE 2 The first four order mode of vibration of the constraint of fix to hinge

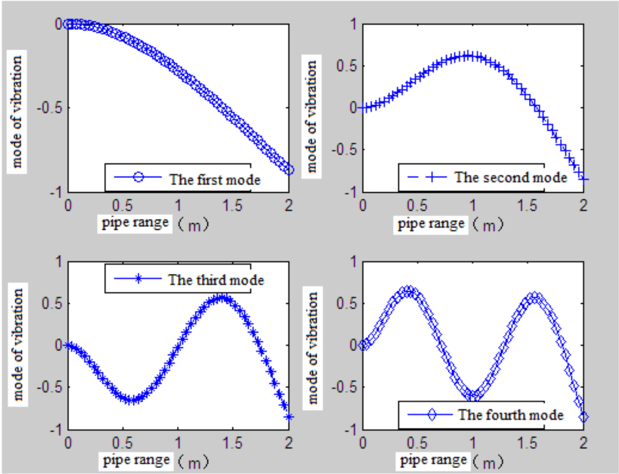


FIGURE 3 The first four order mode of vibration of the constraint of fix to suspension extension

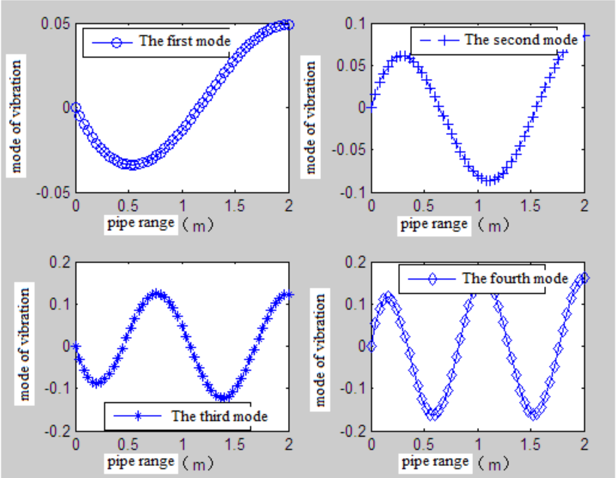


FIGURE 4 The first four order mode of deflection vibration mode for the constraint of fix to hinge

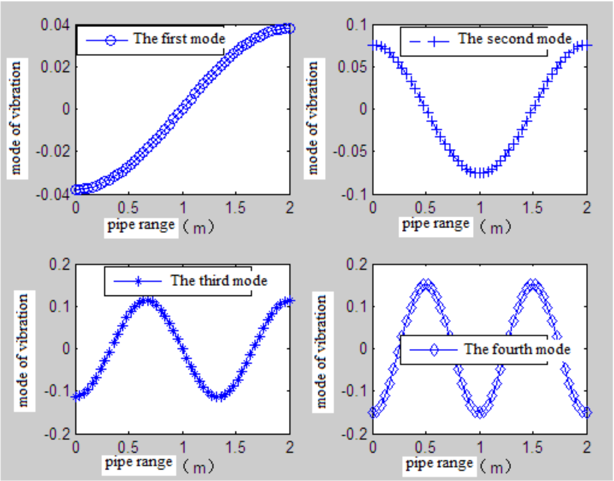


FIGURE 5 The first four order mode of deflection vibration mode for the constraint of hinge to hinge

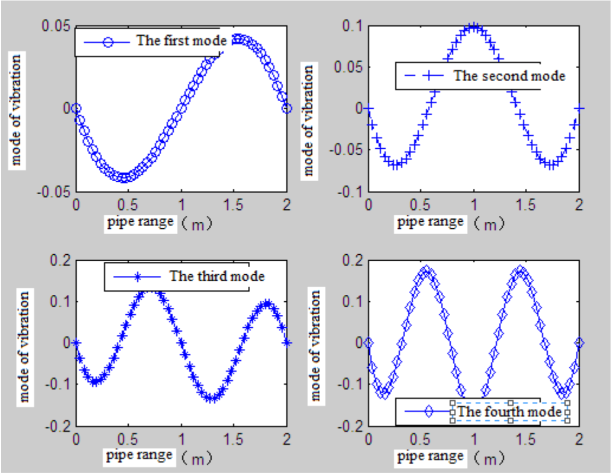


FIGURE 6 The first four order mode of deflection vibration mode for the constraint of fix to fix

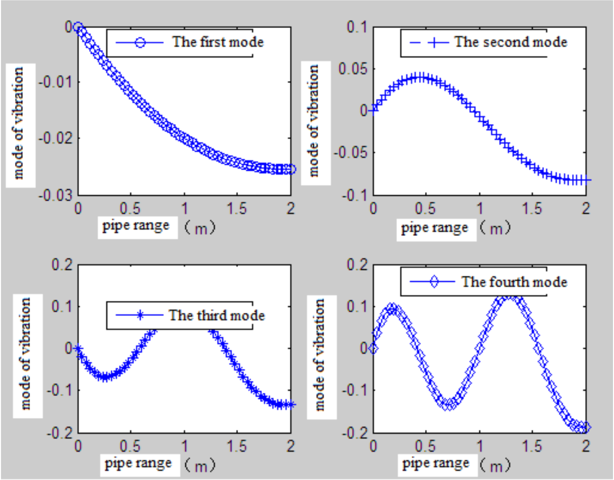


FIGURE 7 The first four order mode of deflection vibration mode for the constraint of fix to suspension extension

From the simulation results, the boundary constraints have great influence on the vibration characteristics of the conveying fluid pipe. Compare with vibration characteristics under the various boundary conditions, it can be concluded that the constraint of fix to hinge, hinge to hinge, fix to fix, fix to suspension extension:

a) The first-order natural frequency arrange as follows: the constraint of fix to suspension extension, hinge to hinge, fix to hinge, fix to fix;

b) Node vibration period arrange as follows: the constraint of fix to fix, fix to hinge, hinge to hinge, fix to suspension extension;

c) Node vibration amplitude arrange as follows: the constraint of fix to fix, fix to hinge, hinge to hinge, fix to suspension extension.

TABLE 1 The vibration characteristics in four kinds of boundary conditions

Name		**Boundary conditions**			
		Fix to hinge	Hinge to hinge	Fix to fix	Fix to suspension extension
Natural frequency (Hz)	First order	23.3534	13.9358	34.8789	1.8728
	Second order	78.5668	61.4928	97.5413	32.5998
	Third order	165.5251	140.6239	192.3795	96.666
	Fourth order	284.1623	251.3922	318.8951	191.7252

4 Conclusions

This paper transformed the fluid conveying pipe to the beam element model for two nodes, and with the Matlab to simulate the transverse motion equation of the conveying fluid pipe, summarized the simulation results and analyzed. In four kinds of boundary conditions, analysis the affect factors of pipeline system parameters for the vibration modal of the fluid conveying pipe, verified the correctness of the established vibration model of the fluid conveying pipeline. The characteristics of pipes conveying fluid is obtained which the pipeline system parameters under different boundary constraints. To provide the theoretical support for the research of vibration attenuation of fluid conveying pipes.

Acknowledgments

This research reported in the paper is supported by National Natural Science Foundation of China (71271160).

References

[1] Chen Guo, He Li dong, Han Wan Fu, Pei Zheng Wu 2013 Vibration analysis of butylenes centrifugal pump and plunger pump pipeline and research of vibration-damping technology *Journal of Mechanical & Electrical Engineering* **30**(2)167-70 *(in Chinese)*

[2] Paidoussis M P, Li G X 1993 Pipes Conveying Fluid: A Model Dynamical Problem *Journal of Fluids and Structure* **7(**2) 137-204 *(in Chinese)*

[3] Li Hailiang, Yan Jinli, Wang Xufeng 2013 Fluid structure coupling analysis of turbine runner based on ANSYS *Journal of Mechanical & Electrical Engineering* **30**(9) 1093-6 *(in Chinese)*

[4] He Guo, Li dong, Han Wan Fu 2013 Vibration analysis of butylenes centrifugal pump and plunger pump pipeline and research of vibration-damping technology *Journal of Mechanical & Electrical Engineering* **30**(2) 167-70 *(in Chinese)*

[5] Che Yong-qiang, Xu Jing-xia, Qian Xiao-dong 2012 Coupled bending and torsional vibrations of geared rotor system *Journal of Mechanical & Electrical Engineering* **29**(6) 632-5, 649 *(in Chinese)*

[6] Ouyang Jie, Wang Zai-Fu, Zhu Qing-peng, Hu Lin-qiang 2013 Model analysis' parametric design of planetary gear based on VB and ANSYS *Journal of Mechanical & Electrical Engineering* **30**(10) 28-30 *(In Chinese)*

[7] Zhang L, Xiang H 2013 *Computer Modelling and New Technologies* **17**(4) 74-82

[8] Liu J, He X, Wei X, Huang C 2013 *Computer modelling and New Technologies* **17**(4) 102-11

[9] Guo H, He J 2013 *Computer Modelling and New Technologies,* **17**(3) 63–8

Authors

Gongfa Li, born on October 7, 1979, Honghu, China

Current position, grades: Associate professor. College of Machinery and Automation, Wuhan University of Science and Technology
Scientific interest: computer aided engineering, mechanical CAD/CAE, Modelling and optimal control of complex industrial process.
Publications number or main:58
Experience: Dr. Gongfa Li received the Ph.D. degree in mechanical design and theory from Wuhan University of Science and Technology in China. Currently, he is an associate professor at Wuhan University of Science and Technology, China. His major research interests include modelling and optimal control of complex industrial process. He is invited as a reviewer by the editors of some international journals, such as *Environmental Engineering and Management Journal, International Journal of Engineering and Technology, International Journal of Physical Sciences, International Journal of Water Resources and Environmental Engineering,* etc. He has published nearly twenty papers in related journals.

Jia Liu, born in 1990, Shanxi, China

Current position, grades: Currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology
Scientific interest: mechanical CAD/CAE, signal analysis and processing.
Experience: Jia Liu was born in Shanxi province, P. R. China, in 1990. He received B.S. degree in mechanical engineering and automation from Wuchang institute of Technology, Wuhan, China, in 2012. He is currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology. His current research interests include mechanical CAD/CAE, signal analysis and processing.

Guozhang Jiang, born on December 15, 1965, Tianmen, China

Current position, grades: Professor of Industrial Engineering, and the Assistant Dean of the college of machinery and automation, Wuhan University of Science and Technology.
University studies: He received the B.S. degree in Chang'an University, China, in 1986, and M.S. degree in Wuhan University of Technology, China, in 1992. He received the Ph.D. degree in mechanical design and theory from Wuhan University of Science and Technology, China, in 2007.
Scientific interest: computer aided engineering, mechanical CAD/CAE and industrial engineering and management system.
Publications:100
Experience: Guozhang Jiang was born in Hubei province, P. R. China, in 1965. He is a Professor of Industrial Engineering, and the Assistant Dean of the college of machinery and automation, Wuhan University of Science and Technology. Currently, his research interests are computer aided engineering, mechanical CAD/CAE and industrial engineering and management system.

Jianyi Kong, born on February 19, 1961, Jiangxi, China

Current position, grades: The president of Wuhan University of Science and Technology, China
University studies: Jianyi Kong received the Ph.D. degree in mechanical design from Universität der Bundeswehr Hamburg, Germany, in 1995.
Scientific interest: intelligent machine and controlled mechanism, dynamic design and fault diagnosis in electromechanical systems, mechanical CAD/CAE, intelligent design and control, etc.
Publications :200
Experience: He was awarded as a professor of Wuhan University of Science and Technology in 1998. Currently, he is the president of Wuhan University of Science and Technology, China. He services on the editorial boards of the Chinese journal of equipment manufacturing technology. He is a director of the Chinese society for metals, etc. His research interests focus on intelligent machine and controlled mechanism, dynamic design and fault diagnosis in electromechanical systems, mechanical CAD/CAE, intelligent design and control, etc.

Liangxi Xie, born on October 27, 1971, Honghu, China

Current position, grades: Associate professor, College of Machinery and Automation, Wuhan University of Science and Technology
Scientific interest: rotorary vane steering gear (RVSG) and vane seals
Publications :35
Experience: He is a major in mechanical design and theory and focus on the research of rotorary vane steering gear (RVSG) and vane seals. He has published more than ten papers in related journals.

Xiao Wentao, born in 1989, Hubei, China

Current position, grades: Currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology
Scientific interest: mechanical CAD/CAE, signal analysis and processing.
Experience: Wentao Xiao was born in Hubei province, P. R. China, in 1989. He received B.S. degree in mechanical engineering and automation from City College of Wuhan University of Science and Technology, Wuhan, China, in 2013. He is currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology. His current research interests include mechanical CAD/CAE, signal analysis and processing.

Yikun Zhang, born in 1990, Hubei China

Current position, grades: Currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology
Scientific interest: mechanical CAD/CAE, signal analysis and processing.
Experience: Yikun Zhang was born in Hubei province, P. R. China, in 1990. He received B.S. degree in mechanical engineering and automation from Hu Bei University of Arts and Science, Xiangyang, China, in 2013. He is currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology. His current research interests include mechanical CAD/CAE, signal analysis and processing.

Fuwei Cheng, born in 1988, Hubei, China

Current position, grades: Currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology
Scientific interest: mechanical CAD/CAE, signal analysis and processing.
Experience: Fuwei Cheng was born in Hubei province, P.R,China, in 1988.He received B.S. degree in mechanical engineering and automation from Donghu college of Wuhan University, Wuhan, China, in 2012.He is currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology. His current research interests include mechanical CAD/CAE, signal analysis and processing.

Dynamic coupling analysis of rocket propelled sled using multibody-finite element method

Jianhua Zhang[*]

School of Science, Chang'An University, Xi'an 710064, China

Received 1 March 2014, www.tsi.lv

Abstract

Rocket propelled sled is a most important testing tool in aerospace and aviation industries flying along the rails on the ground. It is very difficult to simulate the operating conditions in the computer using numerical analysis method. In consideration of this fact, the dynamics analysis and simulation of the rocket propelled sled were done based on Multibody System Dynamics and Finite Element Analysis theory in this paper. The most difficult work during the analysis was establishing the boundary conditions of the rocket propelled sled. This paper made this kind of attempt. Then the relevant post processing figures and data were obtained, thereby providing the designer and manufacturer with detailed and reliable data. The conclusion is the combination of finite element analysis and multisystem theory is more effective than those before and the boundary conditions are correct and acceptable. The results of it can be important references of structure designers.

Keywords: Rocket propelled sled, Finite Element Analysis, Multibody Dynamic Analysis (MDA), Multibody-Finite-Element Method, Rail Irregularity

1 Introduction

As the increase of R&D projects of large civil aircraft, weapon system and aviation life-saving equipments, the demand for deferent types of rocket-propelled sled is growing quickly. Rocket propelled sled [1-4], as its name implies, is a specifically made sled flying along specially made tracks, which propelled by one or more rocket engines. Rocket propelled sled is a most important testing tool flying along the rails on the ground. It is extensively applied in the design of many fields such as the testing of advanced weapons especially the long-range missiles, the detection of aeronautic and astronautic devices and the examination of equipments in aircraft's escape systems etc. Meanwhile such aeronautical ground facilities must be upgraded to match the increased size and performance of future aircraft, so the importance of designing of rocket propelled sled with hypervelocity and high reliability becomes apparent. It can be equipped with the test specimen and fly in extra high speed, often several times the speed of sound. Due to the high cost of the test equipments, it is very difficult to design. Therefore, how to secure the security of the vehicle and the expensive equipments is very difficult. But the test failure have occurred from time to time and the loss is very huge, so how to secure the vehicle's structural design as reasonable as possible, in addition to experiences, computer simulation and optimization is also very important.

Rocket propelled sled's flying environment is very abominable, such as high velocity, high acceleration, strong vibration etc. Traditionally, in order to ensure the recycling of rocket propelled sled and the valuable equipment on it, the designers placed much emphasis on previous practical experience and chose a large safety factor. The design like this makes rocket propelled sled cumbersome and poorly stabilized. In addition, the heavy weight of rocket propelled sled not only increases the cost of test (need more propellant, more number of rocket engines), but also increases the uncertainty of the flight. Although the departments concerned are now able to do some numerical simulations, they are only restricted to the simple finite element analysis of a single part in rocket propelled sled testing system. The new theories and methods, especially the way of building boundary conditions must be developed for rocket-propelled sled's high acceleration, extremely adverse testing environment. Most important of all, the rocket-propelled sled must be researched as a whole system. Therefore, this paper uses the method of coupling mechanical system simulation (ADAMS) and finite element analysis (ANSYS) technologies [5, 6] to model and simulate the flight of rocket propelled sled. And the rocket propelled sled testing system's response for the coupled case and the function of the external force are offered. This methods based on theories of dynamics of multibody system and finite element method [7, 8]. At last, the integrated and accurate data are obtained for the designers of rocket propelled sled.

[*] ***Corresponding author*** e-mail: zhangjianh@ chd.edu.cn

2 The algorithm of multibody system mechanics and Finite Element Method

To derive the basic equation of the multibody finite element theory [9], we assume that the continuum is linear elastic and small internal deformation. Then we mesh these continuums respectively, a transient kinetic equation at time will be derived from the Lagrange Dynamics Equation [10] like the below:

$$\sum_{i=1}^{N}(M_i\ddot{u}_{t+\tau}+C_i\dot{u}_{t+\tau}+K_iu_{t+\tau})=\sum_{i=1}^{N}(F^i_{t+\tau}+G^i_{t+\tau}\lambda^i_{t+\tau}),(1)$$

where, M_i , C_i , K_i are Mass matrix, Damping Matrix, stiffness Matrix respectively. $F^i_{t+\tau}$ is a vector of a given load. $G^i_{t+\tau}$ is a matrix of coordinate conversion, $\lambda^i_{t+\tau}$ is a coupling unknown load vector.

A simple form of this equation is:

$$M\ddot{u}_{t+\tau}+C\dot{u}_{t+\tau}+Ku_{t+\tau}=F_{t+\tau}+G\lambda_{t+\tau}\,. \quad (2)$$

We use equation (3) like blow to get the value of $\ddot{u}_{t+\tau}, \dot{u}_{t+\tau}, u_{t+\tau}, \lambda_{t+\tau}$, which is a expression including the known variables $\ddot{u}_t$, $\dot{u}_t$ and u_t.

$$\begin{cases}\dot{u}_{t+\tau}=u_{t+\tau}[(1-a)\ddot{u}_t+a\ddot{u}_{t+\tau}]\\ u_{t+\tau}=u_t+\tau\dot{u}_t+\frac{1}{2}\tau^2[(1-b)\ddot{u}_t+b\ddot{u}_{t+\tau}]\end{cases}, \quad (3)$$

where a, b, τ are the formatting parameters. Substituting equation (3) into equation (2), we will get:

$$\bar{K}_{t+\tau}u_{u+t}=\bar{F}_{t+\tau}+G_{t+\tau}\lambda_{t+\tau}, \quad (4)$$

where, $\bar{K}$ is effective stiffness matrix, $\bar{F}$ is equivalent load vector. As a rule, we also give the simplified form like the below:

$$\bar{K}u=\bar{F}+G\lambda\,. \quad (5)$$

Then, consulting equation (5), we can construct a functional equation as is equation (6) shown:

$$J(u,\lambda)=\frac{1}{2}\lambda^T\bar{K}u-u^T\bar{F}-u^TG\lambda\,. \quad (6)$$

The another form of equation (5) is as equation (7) describes:

$$u=\bar{K}^{-1}(\bar{F}+G\lambda)\,. \quad (7)$$

Substituting equation (7) into equation (6):

$$J(\lambda)=-\frac{1}{2}\lambda^TG^T\bar{K}^{-1}G^{-1}\lambda-\lambda^TG^TK^{-1}F-\frac{1}{2}\bar{F}^T\bar{K}^{-1}\bar{F}\,. \quad (8)$$

This functional's minimum can be deduced equation (9):

$$K_\lambda\lambda=F_\lambda\,. \quad (9)$$

Therefore, the main work of the multibody finite element theory is to solve the equations like below:

$$\begin{cases}K_\lambda\lambda=F_\lambda\\ \bar{K}u=\bar{F}+G\lambda\end{cases}. \quad (10)$$

From these equations, we can see the possibility of the joint about these two theories, which implementation is the combination of ANSYS and ADAMS, as the upcoming section discusses.

3 Method of combined use of ANSYS and ADAMS

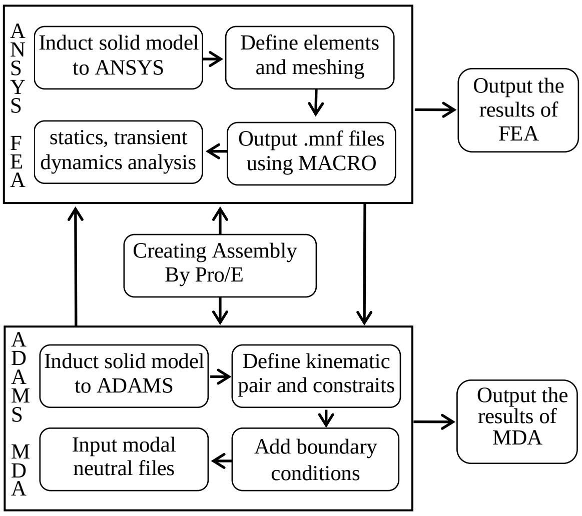

ANSYS FEA
Induct solid model to ANSYS
Define elements and meshing
statics, transient dynamics analysis
Output .mnf files using MACRO
Output the results of FEA
Creating Assembly By Pro/E
ADAMS MDA
Induct solid model to ADAMS
Define kinematic pair and constraits
Input modal neutral files
Add boundary conditions
Output the results of MDA

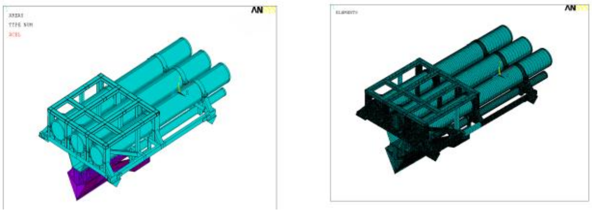
FIGURE 1 Procedure of the joint simulation

In a complex mechanical system, the flexible body will have a major impact on movement of the whole system, without which, Kinematic and Dynamic Analysis on the rigid-flexible coupling model (RFCM) will result in large errors. The movement of the whole system in turn determines the status of the force and motion state of each component. Only phasing in flexible technology can we get the accurate distribution of each part's stress and strain. In other words, only the key parts of sled are dealt with flexible bodies, can accurate dynamic simulation results be got correctly. In essence, at theory, the stress and strain analysis of flexible bodies models(FBM) need the union of the multibody dynamics and finite element method8, while in practice, the joint use of ANSYS and ADAMS (see FIGURE 1) can make the simulation results more accurate. ANSYS program automatically generates the finite element model of flexible body parts, then using the macro command “ADAMS.MAC”, we can easily output model neutral file (jobname.mnf), which is required by ADAMS, this file contains all the information in the flexible body. After defined the constraint of sled’s kinematic pair and the boundary conditions, ADAMS can do dynamics simulation reasonably

4 Creating Rigid-Flexible Coupling Model

4.1 THE NECESSARY ASSUMPTIONS TO SIMPLIFY THE MODEL

The following assumptions should be made when simulating on the rocket-propelled sled’s Rigid-Flexible Coupling Model (RFCM) [11, 12], as is shown in Figure 2:



1) Entity Mode 2) RFCM

FIGURE 2 Entity model and RFCM of rocket sled

1) Rocket propelled sled flies along a straight line and the sled body moves in parallel to the ground;
2) 1/2 model can simulate the whole sled exactly, because the sled is right-left symmetrical;
3) In addition to the elastic component we focus on, the other parts are thought as rigid bodies, which not considered those deformations.

Proewildfire4.0’s global Cartesian coordinate system, which origin placed in the midpoint of the two back slippers, X-axis pointing to flight direction, Y-axis pointing to the left direction, Z-axis pointing vertically.

4.2 THE COORDINATE SYSTEM OF THE MODEL

During the establishment of multibody model, the choice of coordinate system concerns the complexity and difficulty of solving mathematical model’s equations [13-15]. In this paper, the use of the ISO coordinate system makes this problem simpler. Rocket propelled sled models reference for the coordinates system in Pro/E.

4.3 DEFINITION OF KINEMATIC PAIRS

The principle of defining Kinematic pairs is that they must reflect the physical prototype’s true working status as far as possible, meanwhile neglecting the unnecessary ones. Table 1 defines the deferent types of Kinematic pairs in the model, and S is the abbreviation of slipper; P represents the product sled; D is on behalf of driver sled; T, on behalf of three-dimensional NC turntable; LF is the representative of the left-front position; LB, representative of the left-back position; RF, on behalf of right-front position; RB is the representative of right-front position.

TABLE 1 Types of kinematic pairs of RFCM

Index	Type	Definition in ADAMS
1	Contact pair	Contact (.RocketSled. Product-sled. Stop&. RocketSled.Ground. Spring-damper)
2	Revolute pair	Revolute (.RocketSled. Product Driver-sled &. RocketSled. Pin)
3	Cylindrical pair	Cylindrical (.RocketSled. Pin &. RocketSled. Slider)
4	Contact pair	Contact(.RocketSled. Slider & .RocketSled. Ground. Rail)
5	Fixed pair	Fixed (.RocketSled. Product-sled&. RocketSled. Turntable)
6	Contact pair	Contact (.RocketSled. Product-sled. Contactor &.RocketSled. Driver-sled. Contactor)
7	Fixed pair	Fixed (.RocketSled. Driver-sled &. RocketSled. Rocket)

Given name of rocketsled, the model’s every part is automatically added the prefix ’.rocketsled’. Driver sled and product sled are connected with the rails by pins and sliders. The sliders clench the rails firmly so as to ensure the sled cannot get away from the rails during the flight at ultra high speed. The connection of sled and pin is revolve pair, which have two groups, that is, driver sled and four pins, product sled and four pins respectively. The two contact pairs are between driver sled and product sled, and two or three groups of fixed pairs ensure that there is no relative sliding between rocket engine and driver sled. At last, we will get the rigid-flexible coupling model of the rocket-propelled sled testing system in ADAMS2005. The types of kinematic pairs are shown in Table 1 and Table 2.

5 Boundary Conditions During Flight of Sled

5.1 THE CALCULATION OF WIND LOAD

The applied wind pressure of the flying rocket Propelled sled can be equivalent to the wind loads. The way of getting the wind load is as follows: If rocket propelled sled is flying at high-speed of 300m/s, it is equivalent that there is the same high-speed wind of 300m/s blowing to the rocket propelled sled. The mechanics model of rocket propelled sled is shown in Figure 3.

TABLE 2 Detailed definition of joint type

Joint style	Joint name	First rigid body	Second rigid body
Revolute	*.Rocketsled.D-Pin-LF*	*.Rocketsled.Driver-sled*	*.Rocketsled.Pin-_D-LF*
Revolute	*.Rocketsled.D-Pin-LB*	*.Rocketsled.Driver-sled*	*.Rocketsled.Pin-_D-LB*
Revolute	*.Rocketsled.D-Pin-RF*	*.Rocketsled.Driver-sled*	*.Rocketsled.Pin-_D-RF*
Revolute	*.Rocketsled.D-Pin-RB*	*.Rocketsled.Driver-sled*	*.Rocketsled.Pin-_D-RB*
Revolute	*.Rocketsled.P-Pin-LF*	*.Rocketsled.Product-sled*	*Rocketsled.Pin-_P-LF*
Revolute	*.Rocketsled.P_-pin-_LB*	*.Rocketsled.Product-sled*	*.Rocketsled.Pin-_P-LB*
Revolute	*.Rocketsled.P_-pin-_RF*	*.Rocketsled.Product-sled*	*.Rocketsled.Pin-_P-RF*
Cylindrical	*.Rocketsled.P_-pin-_RB*	*.Rocketsled.Product-sled*	*.Rocketsled.Pin-_P-RB*
Cylindrical	*.Rocketsled.Pin-_S_D_LF*	*.Rocketsled.Pin-_D-LF*	*.Rocketsled.Slider-D-LF*
Cylindrical	*.Rocketsled.Pin-_S_D_LB*	*.Rocketsled.Pin-_D-LB*	*.Rocketsled.Slider-D-LB*
Cylindrical	*.Rocketsled.Pin-_S_D_RF*	*.Rocketsled.Pin-_D-RF*	*.Rocketsled.Slider-D-RF*
Cylindrical	*.Rocketsled.Pin-_S_D_RB*	*.Rocketsled.Pin-_D-RB*	*Rocketsled.Slider-D-RB*
Cylindrical	*.Rocketsled.Pin-_S_P_LF*	*.Rocketsled.Pin-_P-LF*	*.Rocketsled.Slider-P-LF*
Cylindrical	*.Rocketsled.Pin-_S_P_LB*	*.Rocketsled.Pin-_P-LB*	*.Rocketsled.Slider-P-LB*
Cylindrical	*.Rocketsled.Pin-S-P-RF*	*.Rocketsled.Pin-_P-RF*	*.Rocketsled.Slider-P-RF*
Cylindrical	*.Rocketsled.Pin-S-P-RB*	*.Rocketsled.Pin-_P-RB*	*.Rocketsled.Slider-P-RB*
Planar	*.Rocketsled. S-D-LF/B*	*.Rocketsled.Slider-D-LF/B*	*Rocketsled.Ground.L-rail*
Planar	*.Rocketsled. S-D-RF/B*	*.Rocketsled.Slider-D-RF/B*	*.Rocketsled.Ground.R-rail*
Planar	*.Rocketsled. S-P-LF*	*Rocketsled.Slider-P-LF*	*.Rocketsled.Ground.L-rail*
Planar	*.Rocketsled. S-P-LB*	*.Rocketsled.Slider-P-LB*	*.Rocketsled.Ground.L-rail*
Planar	*.Rocketsled. S-P-RF*	*.Rocketsled.Slider_P_RF*	*.Rocketsled.Ground.R-rail*
Planar	*.Rocketsled. S-P-RB*	*.Rocketsled.Slider_P_RB*	*.Rocketsled.Ground.R-rail*
Contact	*.Rocketsled.Damper-L*	*.Rocketsled.Driver-sled. Damper_L*	*.Rocketsled.Product-sled. Damper-L*
Contact	*.Rocketsled.Damper-R*	*.Rocketsled.Driver-sled. Damper-R*	*.Rocketsled.Product-sled. amper-R*
Fixed	*.Rocketsled.Fixed-P-T*	*.Rocketsled.Product-sled.*	*.Rocketsled.Turntable*
Fixed	*.Rocketsled.Fixed-D-R*	*.Rocketsled.Driver-sled.*	*.Rocketsled.Rocket*

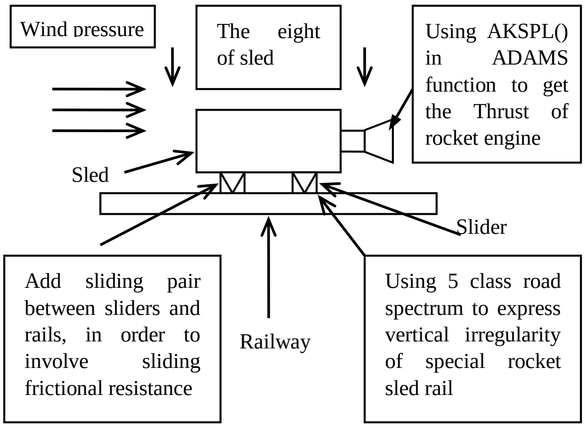

Wind pressure
The eight of sled
Using AKSPL() in ADAMS function to get the Thrust of rocket engine
Sled
Slider
Add sliding pair between sliders and rails, in order to involve sliding frictional resistance
Railway
Using 5 class road spectrum to express vertical irregularity of special rocket sled rail

FIGURE3 Mechanics model of sled

The strength of the wind is usually expressed by wind pressure. It is question in the domain of fluid mechanics. The Bernoulli equation can give us the exact wind pressure, as is shown in the following expression:

$$w=\frac{1}{2}\rho v^2=\frac{1}{2}\frac{\gamma}{g}v^2(kN/m^2)\,, \qquad (10)$$

where ρ is the density of the air (t/m^3), v is the speed of wind (m/s).

If the air pressure is the standard atmosphere pressure, $1/2(r/g)$ is about 1/1630. It is various in deferent part of the world, in china we can use 1/1600 according to china's concerned specifics.

Then, $w=1/1600\times300\times300=56.25\text{KN}$.

5.2 VERTICAL IRREGULARITY OF SPECIAL ROCKET PROPELLED SLED RAIL

As previously mentioned, the sled, either driver sled or product sled, is sliding along the rails, in other words, it is the slider that enables the sled flying in the railway and keep the sliding frictional contact between this two components. The cross-section of rail is choose the GB (national standard) size. The rails' distance is 1.2 meters, and the sleepers' space are all 0.5m. To calculate the vibrations from rails, we choose some kind of actual 5 class spectral density to simulate the vertical irregularity of the special rocket propelled sled rail, as is shown in Figure 4.

(a) The approach to take about the rail spectral density's expression in ADAMS is as below, namely, choosing about 20 sampling points in the road spectral curve, then using the spline curve function, which is in the ADAMS Function Builder to fit those 20 points, and finally simulating by Motion in the software.

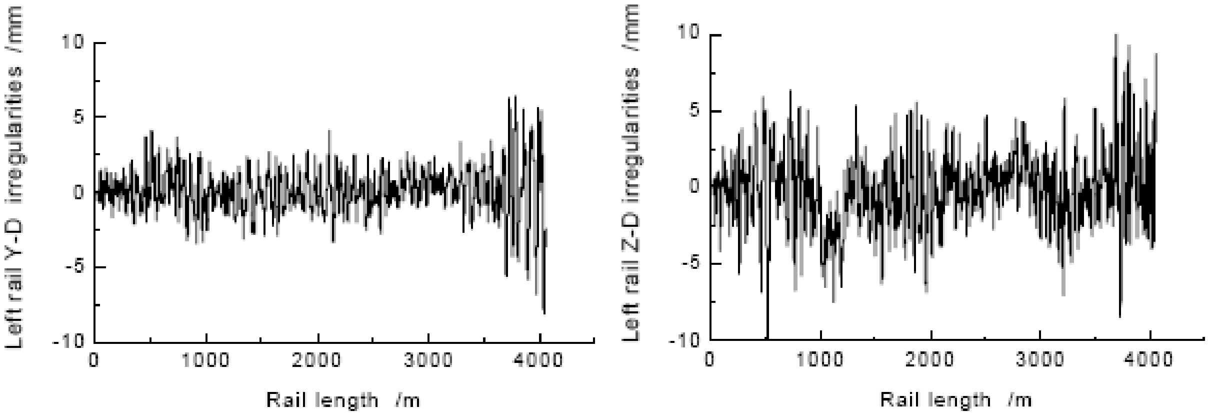


FIGURE 4 Some Actual 5 Class Spectral Density

5.3 THE THRUST OF THE ROCKET ENGINE

The rocket propelled sled's thrust on driver sled is also fit with spline curve function AKISPL() in ADAMS Function Builder. As shown below: AKISPL(Time, .Rocketsled. Driver_time, .Rocketsled. Spline_n,0) where, Time is the simulation current time, .Rocketsled.Driver_time is the time of rocket engine ignition time, .Rocketsled.Spline_n is the curve of rocket engine's thrust, which is fit by spline function in ADAMS. We can select a group of rocket engine performance thrust curve according to the needs then, and verify the performance of different rocket engines' thrust curve and its influence on the system, in order that we can select the most appropriate physical prototype testing engine models. The other useful friction parameters is: Stiffness=1.0E+008, Force Exponent=2.2, Damping= 1.0E+004， dynamic friction coefficient Dynamics=0.3, static friction coefficient Static＝0.5.

6 Some simulation results

6.1 RESPONSE OF ROCKET PROPELLED SLED ACCELERATION

Through the calculation of ADAMS, we can get the acceleration curve of three directions along three axes in Cartesian coordinate system, as shown in Figures 5 and 6.

It can be seen from the following three figures: The up-down direction's acceleration is about 35g, the front-back direction (flight direction) one is 16g. Compared with the actual test, the gap between those data is permitted, what verifies the reliability of the simulation theory and method, especially the definition of Boundary Conditions in this paper. (In a actual flight test, we used the accelerometer to measure two directions' acceleration were 15g, 36g).

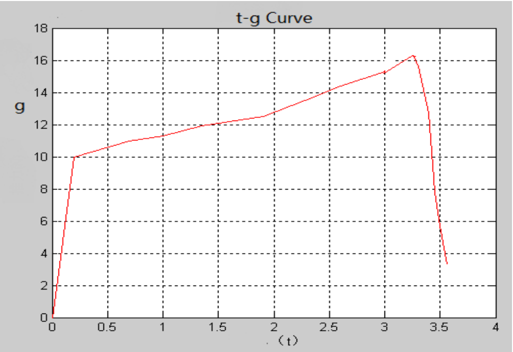


FIGURE 5 The front-back direction's acceleration

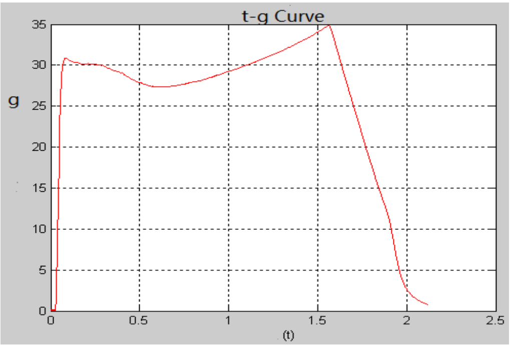
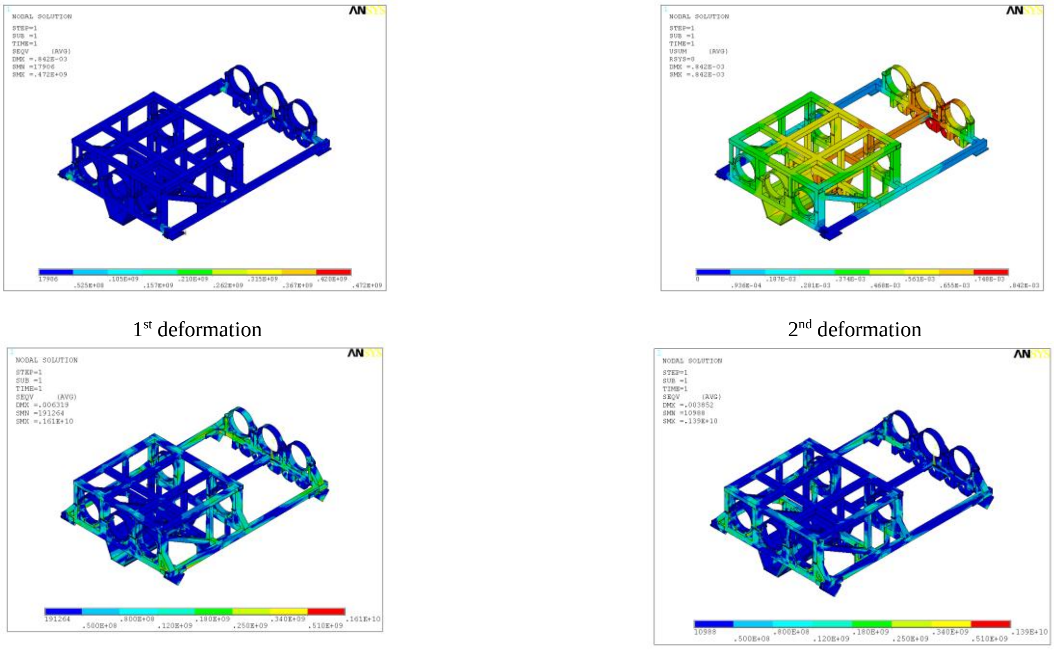


FIGURE 6 The up-down direction's acceleration

6.2 INPUT THE LOAD FILE GOTTEN IN ADAMS TO ANSYS

After getting the modal neutral file of the part we concerned, we establish the whole sled's rigid-flexible coupling mechanical system in ADAMS, with driver sled as flexible body. In this model, we establish the connecting way of rigid boy and flexible body, and do the dynamics analysis after defined the loads and boundary conditions. At the end of the analysis, we get output load files (.lod), which required by ANSYS. This file contains movement status at different time points and other load information (such as force, acceleration and so on). Because the load suffered by slippers is transferred to sled body through the pins, we add this equivalent load to sled body directly. Then we choose 4 time points and wait for ANSYS's calculation, at last get the driver sled's deformations in Figure 7.


NODAL SOLUTION
STEP=1
SUB =1
TIME=1
SEQV (AVG)
DMX =.842E-03
SMN =17906
SMX =.472E+09
17906 .525E+08 .105E+09 .157E+09 .210E+09 .262E+09 .315E+09 .367E+09 .420E+09 .472E+09
NODAL SOLUTION
STEP=1
SUB =1
TIME=1
USUM (AVG)
RSYS=0
DMX =.842E-03
SMX =.842E-03
0 .936E-04 .187E-03 .281E-03 .374E-03 .468E-03 .561E-03 .655E-03 .748E-03 .842E-03
NODAL SOLUTION
STEP=1
SUB =1
TIME=1
SEQV (AVG)
DMX =.006319
SMN =191264
SMX =.161E+10
191264 .500E+08 .800E+08 .120E+09 .180E+09 .250E+09 .340E+09 .510E+09 .161E+10
NODAL SOLUTION
STEP=1
SUB =1
TIME=1
SEQV (AVG)
DMX =.003852
SMN =10988
SMX =.139E+10
10988 .500E+08 .800E+08 .120E+09 .180E+09 .250E+09 .340E+09 .510E+09 .139E+10
ANSYS

1st deformation

2nd deformation

3rd deformation

4th deformation

FIGURE7 The Deformation of 4 time points

7 Conclusions

It is an effective way that we carry out the rocket-propelled sled testing system's simulation analysis by building a reasonable dynamics-simulating model. Through this attempt, we can overcome the difficulty of the restriction of flight test actual condition, as well as the critical lack of research funds, the shortage of real testing device. Most of all, we can do this test through a much safer way, that is, computer virtual simulation. Of cause, we finally get the integrated data and the mechanical structure's effect on the real test, which should be the necessary and critical instructions of the design of various rocket propelled sled-testing system. Moreover, there is no large difference between data of the virtual simulation analysis and those of actual test of physical prototype. Through the comparative analysis, we think that the method and the definition of boundary conditions this paper constructed is reliable and faithful.

Acknowledgments

The author was endowed by the Ministry of Transport key Laboratory of Automotive Transportation Security, Institute of Automobile, Chang'An University.

Also this paper was endowed by The Central Ministry University funding for basic research (CHD2012JC079), And the natural science foundation of Shaanxi Province (2014JM7297).

References

[1] Cinnamon J D, Palazotto A N 2009 *International Journal of Impact Engineering* **36** 254-62

[2] Szmerekovsky A G, Palazotto A N, Baker W P 2006 *International Journal of Impact Engineering* **32** 928-46

[3] Laird D J , Palazotto A N 2004 *International Journal of Impact Engineering* **30** 205-23

[4] Tachau R D M, Yew C H, Trucano T G 1995 *International Journal of Impact Engineering* **17** 825-36

[5] Gerstle Jr F P, Follansbee P S, Pearsall G W, Shepard M 1973 *Wear* **24** 97-106

[6] Graff K F, Dettloff B B 1969 *Wear* **14** 87-97

[7] Hypersonic Rail Tests 2003 *Railway Gazette International* 357

[8] Lofthouse A J *Computational Aerodynamic Analysis of the Flow Field about a Hypervelocity Test Sled* 1998

[9] Hale C S, Palazotto A N, Baker W P 2012 *Journal of Engineering Mechanics* **138** 1127-40

[10] Tarcza K R, Weldon W F 1997 *Wear* **209** 21-30

[11] Liang W-j, Su C, Wang F, Conservancy X-j T Co W 2009 Surrounding rock deformation analysis of underground caverns with multi-body finite element method *Water Science and Engineering* Engineering H, University H 210098 N and China P R *(in Chinese)*

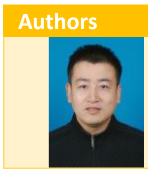
[12] Korkegi R H, Briggs R A A R L W-P A O 1968 Aerodynamics of the Hypersonic Slipper Bearing 1968 *DTIC Document*

[13] Kanas N, Salnitskiy V, Grund E M, Weiss D S, Gushin V, Kozerenko O, Sled A, Marmar C R 2001 *Acta Astronautica* **48** 777-84

[14] William F A 1999 *Wear* **235** 25-38

[15] Winter F H, James G S 1995 *Acta Astronautica* **35**, 677-98

Authors



Zhang Jianhua, born on October 31, 1973 , Yantai, Shandong Province, China

Current position, grades: Doctor of Engineering of Chang'an University, School of Science
Research interests: CAD/CAE, the vibration studying

A heuristic task deployment approach for load balancing

Gaochao Xu[1], Yunmeng Dong[1], Xiaodong Fu[1],
Yan Ding[1], Peng Liu[1], Jia Zhao[2]*

[1] *College of Computer Science and Technology, Jilin University, Changchun 130012, China*

[2] *College of Computer Science and Engineering, Changchun University of Technology, Changchun 130012, China*

Received 1 February 2014, www.tsi.lv

Abstract

The load balancing strategy, which is based on the mission deployment, has become a hot topic of green cloud data centre. For the question that currently the overloaded physical hosts in the cloud data centre causes the load imbalance of the whole cloud data centre, the proposed makes an intensive study which is about the select location question of the deployment tasks on the physical host and then this proposed a new heuristic method which is called LBC. Its main idea consists of two parts: First, based on the function, which denotes the performance fitness of physical hosts, it conducts a constraint limit to all physical hosts in cloud data centre. So a task deployment strategy with global search capability is achieved. Secondly, using clustering methods can further optimize and improve the final clustering results. Thus, the whole way achieves the long-term load balancing of the cloud data centre. The results show that compared with the conventional approach, LBC significantly reduces the number of failure of the deployment tasks, improves the throughput rate of the cloud data centre, optimizes the performance of external services of the data centre, and performs well in terms of load balancing. Besides, it makes the operation of cloud data centres be more green and efficient.

Keywords: load balancing strategy, cloud data centre, task deployment, LBC, clustering

1 Introduction

Cloud computing [1, 2] is the focus of research topic currently which is the most promising and valuable research direction following the utility computing, grid computing and distributed computing. Cloud computing provides users with infrastructure, platform and software services according to user's needs through the Internet. Infrastructure as a Service (IaaS) is the foundation of cloud computing, whose key is to make the data centre cloud computing resources to be a resource pool through virtualization technology. Besides, it allocates according to the task specifications and resource requests, which are submitted by users. In addition, it provides elastic physical or virtual computing, storage and network resources. A large number of physical hosts, which are deployed by the cloud data centre, provide services for users. However, each physical host's resource remaining amount is changing all the time. Therefore, it cannot guarantee to place every task on the physical host, which has the largest remaining amount of resources.

Currently, load balancing is the hot issue in the domain of the cloud data centre study. In order to further optimizing load balancing in the cloud data centre, this proposed presents a heuristic idea [3], load balancing strategy, which aims at finding the physical host whose deployment performance is optimal. And details are as follows: First, giving a constraint value which is based on the resource amount of requested tasks. Then cluster the physical hosts, which are greater than the constraint value in the cloud data centre. Secondly, forming a set of physical hosts by clustering whose similarities are within a certain threshold. The collection of physical hosts got after clustering is the physical hosts collection having the optimal deployment performance, which we want to find. Finally, place the tasks, which need to be processed into the physical hosts which are in the collection to conduct deployment. Clustering the physical hosts in the data centre exactly is the process of finding the physical hosts, which have optimal deployment performance. Therefore, through our deployment task strategy, it cannot only achieve load balancing in the cloud data centre, but also provide efficient external service performance for users.

This paper aims to achieve long-term load balancing in the cloud data centre and provide users with efficient external service performance. And achieving long-term load balancing in the cloud data centre must by means of deploying the task request to the resource pool in the cloud data centre efficiently and rationally. Therefore, it can achieve load balancing in the cloud data centre and improve the efficiency of the cloud data centre. Furthermore, it can show the excellent external service performance of cloud data centre to users. The load balancing strategy proposed cannot only find the physical hosts, which have optimal deployment performance efficiently, but achieve long-term load balancing in the cloud data centre.

* ***Corresponding author*** e-mail: zhaiyj049@sina.com

Other parts of this paper are organized as follows: In the second part, we briefly describe the current work that is related to the method, which can achieve the load balancing of cloud data centre. In the third section, we first point out the premise, which proposed in our questions, and then introduce the design and implementation process of our algorithm in details. In the fourth section, we will give the experiments and results, and prove that the algorithm we proposed has high efficiency. The fifth part, we summarize the full paper and future work is put forward.

2 Related works

Load balancing has been a hot research topic of cloud data centre [4] and its goal is to ensure that every computing resource can process tasks efficiently and fast, improve the utilization of resources ultimately. The question is present in a cloud computing environment. When there are some task requests in the cloud data centre, these tasks request will be deployed to the optimal physical host of the cloud data centre, so that the computing performance of cloud computing centres can achieve optimal, while cloud data centres can achieve the load balancing of entire network. Researchers have proposed a series of static, dynamic and mixed scheduling policies. In addition, there are also some studies using live migration technology of virtual machine to meet the cloud data centres' requested tasks which include performance requirements and load limitation. In fact, most problems are just deploying the requested tasks to the cloud data centre's physical hosts.

Existing load balancing strategies are generally divided into two categories: static load balancing and dynamic load balancing. Static load balancing scheduling algorithm [5-8] are commonly used round robin, weighted round robin, least connection method, weighted least connection method and so on. These static algorithms only use some static information, which cannot solve dynamic load changes among servers in cluster effectively and their adaptive ability is poor. Currently, some of the most open-source IaaS platform most use static algorithm to conduct resource scheduling. For example, Eucalyptus [9] platform uses round robin to assign virtual machines to different physical hosts in sequence to achieve load balancing. In Literature [10], Wei Q et al. used the weighted minimum link algorithm, which means that different weights indicate the performance of the physical host. Then, the virtual machine will be allocated to the physical host, which has the smallest ratio of the number and weight. The advantage of static scheduling algorithm is that it is simple to do. But facing the large-scale cloud data centre whose heterogeneous resources are strong and users' consistent demand, load balancing effect is not ideal.

Dynamic load balancing [11-13] is a NP-complete problem which is a classic combinatorial optimization problem. It is mainly used in the field of distributed parallel computing, and its main objective is how to distribute the load more rationally among multiple computers to avoid some phenomenon of calculation node overload and light load. Thus, overall system performance can be improved. Additional communication overhead produced in the process of DLB will reduce the dynamic load balancing system performance. And with the increase of network latency among each node, the influence of the restricting DLB performance of additional communication overhead will further increase. Therefore, how to reduce communication overhead furthest among each node in the process of DLB becomes an important problem, which will influence the performances of DLB. Now aiming at the problem of reducing additional communication overhead in the process of DLB, the solution is mainly using greedy algorithm to process. The LRS algorithm, which is put forward in Literature 9 using the light load preferentially received allocation pattern.

In Literature [14], Lau et al. integrated two strategies which are heavy load priority and light load priority. They put forward an adaptive load distribution algorithm, which helps reduce the load balancing communication overhead effectively. Using the greedy algorithm can solve the problem of load distribution. However, several algorithms above cannot meet greedy choice performance and sub-optimal structural property at the same time. Therefore, load distribution program was often local optima. And the effect of solving the problem of load distribution under certain special circumstances is not ideal. Cloud data centre cannot reach the entire network load balancing. Virtual machine migration placement strategy is the most widely used strategy to achieve cloud computing [15, 16] data centre load balancing currently. VMware load balancing solution is DRS (Distributed Resource Scheduling) [17]. When DRS select the physical host for the virtual machine, it will check the load status of each physical host and choose the placement method to reduce the overall load imbalance. And in the process of running a virtual machine, DRS will continue to monitor the load status of the cluster and use VMware VMotion technology to perform live migration of virtual machines among different physical servers. Thus, it can ensure load balancing and efficient use of physical resources of the entire cluster.

3 LBC algorithm design

In IaaS cloud data centre, when users have requests, the system will deploy the task request to the physical hosts, which are in the resource pool of a cloud data centre. In general, cloud data centres will select physical hosts randomly to deploy. When the requested resources are greater than the physical hosts' remaining resources, physical hosts cannot deploy the task. When the requested resources are in proximity to the physical hosts' remaining resources, it will cause overload of physical hosts. Thus, it will cause load imbalance in the cloud data centre and result in decreased efficiency and increased energy consumption. Obviously, with regard to the cloud data centre, different deployment task strategies will cause different load allocation in entire system. There is no doubt that the optimal deployment task strategy can make the entire cloud computing system produce the effect of load balancing. Therefore, it is necessary to design and implement an efficient and load-balancing deployment task strategy in the cloud data centre.

3.1 IMPLEMENTATION OF LBC ALGORITHM

The implementation of the LBC algorithm:

Step 1:

Assuming the number of physical hosts in the data centre is n. We need to do a constraint to all physical hosts in the data centre. And in order to meet the physical hosts' performance constraints. We treat the physical hosts' remaining amount of resources L_i as a metric. It is defined as follows:

$$L_i = \alpha L_c + \beta L_{mem}, \tag{1}$$

$$\alpha + \beta = 1. \tag{2}$$

L_i shows the remaining computing resources of physical host node *I*, which mean the usage of CPU and memory usage. L_c is the remaining of CPU. L_{mem} is the remaining of memory. α is the weight of CPU. β is the weight of memory. The value of α and β are obtained through BP neural network study. According to the fitness function (1) and (2) of the physical hosts' performance, it generally obtains the monitoring data of the physical hosts' various performance in the entire data centre through SNMP (simple network management protocol), including CPU and memory data. The remaining resources of n physical hosts in the cloud data centre can be calculated. The constraint value is defined as: The total amount of resources received of task request collection within time Δt, namely:

$$L_{req} = \sum_{i=1}^{n} L_{tk}^{i}. \tag{3}$$

In this equation, L_{req} is the total amount of resources of task request collection, L_{tk}^{i} is the resource of task *i* in the task request collection. There defines an empty set *S* = {}.

According to the equation (3), L_{req} can be calculated. When there is an inequation, $L_i > L_{req}$, a host *i* will be put into set *S*, otherwise, we will continue to find. The set *S*, which is got after comparing n physical hosts with the constrained value is S = { s_1 , s_2 , s_3 …., s_m }, m≤n.

Step 2:

We can get the performance values of each physical host based on the fitness function of physical hosts' performance. By restricting the constrained value, we put the physical hosts in the data centre whose performances are relatively good into the set S. Regard the remaining of the physical host's CPU as the physical host's property. Suppose S = { s_1 , s_2 , s_3 , …, s_m } as the set which contains m physical hosts. We arrange CPU remaining of physical hosts in the set S in descending order. Large CPU remaining of physical hosts is arranged in the front. Supposing that s_j is the physical host, which has the largest CPU remaining, we regard s_j as the class-centre. The equations that calculate the similarity are as follows:

$$d(\mathrm{s}_i, \mathrm{s}_j) = \sqrt{\sum_{k=1}^{d} (\mathrm{s}_i^k - \mathrm{s}_j^k)^2}, \tag{4}$$

$$s(\mathrm{s}_i, \mathrm{s}_j) = \frac{1}{d(\mathrm{s}_i, \mathrm{s}_j)}, \tag{5}$$

s_j^1 is a property of the physical host j. It can represent the physical host's CPU remaining. So according to the equations (4) and (5), the similarity of the physical host *i* and the physical host j can be calculated. $s(\mathrm{s}_i, \mathrm{s}_j)$ is the similarity of s_i and s_j.

$$s(\mathrm{s}_i, \mathrm{s}_j) = \frac{1}{\sqrt{(\mathrm{L}_{ci}^1 - \mathrm{L}_{cj}^1)^2}}. \tag{6}$$

Step 3:

Regarding s_j as the class-centre, giving a threshold value $U_{threshold}^{Similarity}$, according to the similarity, we calculate the similarity of s_j and each element of the set S. If the similarity is greater than the threshold value $U_{threshold}^{Similarity}$, we will add this element into the new set S' and not put the class-centre into S'. Then the set S selects class-centre according to the remaining of physical host CPU in descending order and calculate apart the similarity with the elements of S'. Next put the threshold which is greater than $U_{threshold}^{Similarity}$ into the set S'. When the elements of the set S' does not change, the iteration ends. The final clustering result is the set S'. $S' = \{\mathrm{s}_1', \mathrm{s}_2' \ldots \mathrm{s}_q'\}$, q≤m≤n.

Step 4:

It puts the task request received from the data centre into the collection $S^{'}$ of physical hosts, then physical hosts in the collection $S^{'}$ process the task set in a collection of physical hosts to process the requested task collection. After processing, the results will be returned to users. From the physical hosts in the collection $S^{'}$ starting processing the task until the processing is completed, the period of time is recorded as Δt. The task requirements that the data centre receives within time Δt will be the next task to be processed.

Step 5: Repeat the above process.

3.2 MODEL OF LBC ALGORITHM

Overall, the target of LBC algorithm is also in line with the idea of heuristic algorithms. It mainly because the solution that heuristic idea finds every time is not always the optimal solution. But by constant finding and revising, it can get closer to the optimal solution until infinitely close to the optimal solution. And his process meets the goal of algorithm. From the view of the goal of algorithm, we are committed to achieve the load balancing of data centre. An iteration of the algorithm can achieve load balancing. So it needs repeated iterative algorithm, and find the optimal physical host after each iterative. Therefore, users' requested tasks can be disposed by the optimal physical host in the data centre. In this way, after repeated iterative algorithm, each set of physical hosts we found can get close to the best performance. Data centres can quickly process tasks requested by the user. Thus, the data centre tends to be load-balancing and finally it achieves load balancing. This is a long process, and LBC algorithm ensures that the process can get good results within a reasonable time.

4 Evaluation

This paper uses CloudSim simulator to simulate a dynamic cloud data centre. It supports for dynamic creation of different types of entities at run-time and it can add and delete data centre. In CloudSim platform, it creates a resource pool with 100 physical hosts. These hosts have different computing resources and 50 different tasks request resources. They need different CPU and memory of physical host. LBC model we proposed calls and gets resource information and status of physical hosts in the cloud resource pool regularly. In this section, through load balancing degree, make-span, and external service performance, we compared the LBC deployment strategies we proposed with random deployment strategies and do some experiments. The results shown below:

In this scenario of experiments, we compared the Make-Span of LBC and random methods. Make-Span is the completion time of computing tasks. The results shown in Figure 1, it can be seen from the figure that the increase of two deployment methods with the increasing number of requests tasks and the Make-Span of the collection of requested tasks requests will also increase. As can be seen from Figure 1, comparing with random deployment, the LBC algorithm we proposed has a smaller Make-Span under the same condition. This experiment illustrates that the LBC algorithm not only has good load balancing effect, but has relatively good Make-Span.

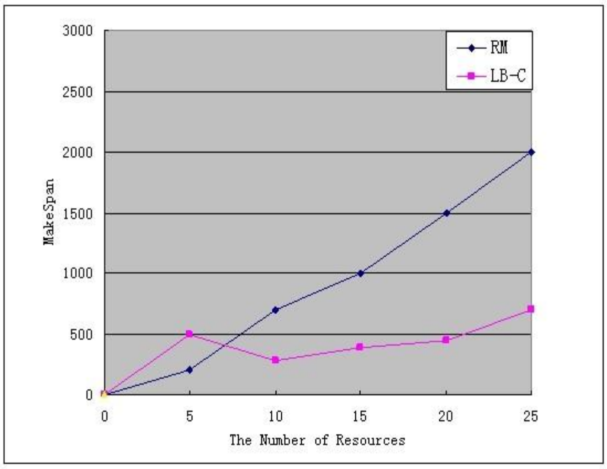


FIGURE 1 Comparison of Makespan

In this set of experiments, we compare the changes of load balancing in the cloud data centre, which influenced by LBC method and stochastic methods over time. As can be seen from Figure 2, with the time increasing during deployments, the load balance degree of stochastic methods and LBC method decrease gradually. The load balance degree of traditional random deployment strategies is always greater than LBC method. It is because the LBC method can quickly find the optimal physical host based on the required CPU resource amount of a requested task. To a certain extent, it ensures the CPU utilization of physical host is much good. From the experimental results, the LBC method we proposed has better load balancing effect. Thus, the resource utilization of the cloud data centre is more effectively improved. And it indirectly saves the power consumption for the cloud data centre.

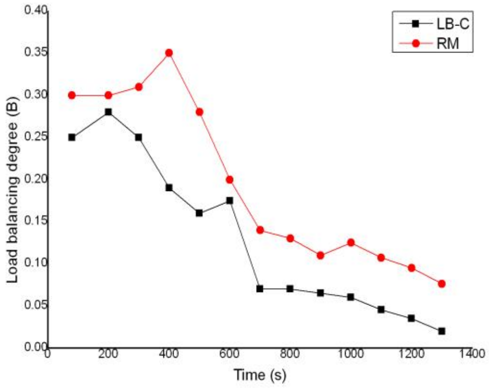


FIGURE 2 Comparison of load balancing degree

The third set of experiments verifies the LBC method from the eternal service performance of the data centre after the deployment of two methods. They selected the throughput as the evaluation criteria of the external service performance of the data centre, because the throughput is usually the overall evaluation of a system and the ability of its assembly units requested ability to process transmission data. Experimental results shown in Figure 3, it can be seen from the figure that using two different deployment methods, external service performances of the data centre are different.

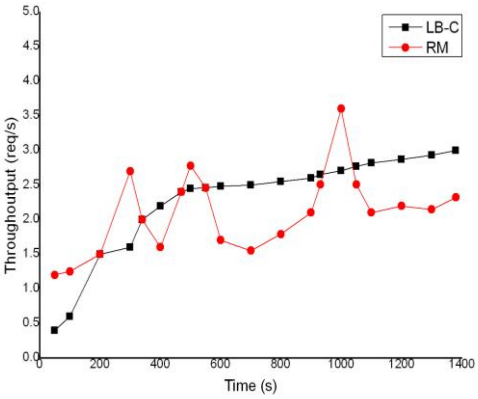


FIGURE 3 Comparison of external service performance

After the random deployment, computing performance of the data centre is much good. With the increase of response time, eternal service performance has a waved trend, and no stability. But using the LBC method to deploy tasks, the initial external service performance of the data centre is not as good as that of the data centre by using the random methods. With the increase of response time, the external service performance of the data centre gradually stabilizes. By comparing the performance of external service performance of the data centre, it can be concluded that using LBC method can be more stable and efficient than random deployment.

5 Conclusion and future work

Based on summarizing the related work, this task proposes a new load balancing strategy, LBC, which is based on task deployment and gives the main idea including process implementation and evaluation. It uses heuristic ideas, which are based on clustering. In these heuristic ideas, first it calculates the total amount of resources of requested tasks according to the fitness function of physical hosts' performance. Then, it compares the amount of tasks with the remaining resources of the physical host in the cloud data centre. And it makes the physical host which is greater than the total amount of resources to be the clustering object. Therefore, LBC has better search capability and adaptive capacity at the beginning. To assess LBC method, there are several experiments done at CloudSim platform. Through four experiments, the result shows that LBC method can deploys the real-time task requests to the resource pool of cloud data centre faster and efficiently. LBC achieves the long-time load balancing and high-efficiency computing capacity of the cloud data centre. It minimizes the number of failures of deployment tasks in the cloud data centre. And to some extent, it improves the throughput of the cloud data centre. In those LBC methods proposed, there are some open questions, which need further research, and some experimental questions, which need a lot of experiments to get a much good solution. The value of CPU weight, α , and memory weight, β , is an empirical question. It needs several experiments to obtain the optimal values so that we can get the equation: $\alpha + \beta = 1$.Therefore, LBC method can be more efficient and feasible. In this context, all the parameters are set to the appropriate value. To further improve the performance of LBC, We plan to study the robustness of LB - C in the next step. LBC method should be able to choose one to more physical hosts to deploy tasks in the collection after clustering. Thus, the cloud data centre and users can get the maximum benefit.

References

[1] Erl T, Puttini R, Mahmood Z 2013 *Cloud Computing: Concepts, Technology & Architecture* Pearson Education

[2] Garg S K, Versteeg S, Buyya R 2013 *Future Generation Computer Systems* **29**(4) 1012-23

[3] Yassa S, Chelouah R, Kadima H, Granado B 2013 Multi-Objective Approach for Energy-Aware Workflow Scheduling in Cloud Computing Environments *The Scientific World Journal* **2013** Article ID 350934 13 p

[4] Liu L, Wang H, Liu X, et al 2009 GreenCloud: a new architecture for green data center *Proceedings of the 6th international conference industry session on Autonomic computing and communications industry session* ACM 29-38

[5] Wei Qun, Xu Guangli, Li Yuling 2011 Research on duster and load balance based on linux virtual server *Information Computing and Applications* Berlin: Springer Heidelberg 169-176

[6] Di Yuan, Wang Shuai, Sun Xinya 2013 A dynamic load balancing model based on negative feedback and exponential smoothing estimation http://www.thinkmind.org/index.php?view=article&articleid=icas_2012_2_10_20043 *(24 Feb 2013)*

[7] Chen Wei, Zhang Yufang, Xiong Zhongyang 2010 Research and realization of the load balancing algorithm for heterogeneous cluster with dynamic feedback *Journal of Chongqing University* **33**(2) 2-14

[8] Song S, Lv T, Chen X 2014 A Static Load Balancing algorithm for Future Internet *TELKOMNIKA Indonesian Journal of Electrical Engineering* **12**(6)

[9] Nurmi D, Wolski R, Grzegorczyk C, et al 2009 The eucalyptus open-source cloud-computing system *Cluster Computing and the Grid, 2009 CCGRID'09 9th IEEE/ACM International Symposium on. IEEE* 124-31

[10] Wei Qun, Xu Guangli, Li Yuling 2011 Research on duster and load balance based on linux virtual server *Information Computing and Applications* Berlin: Springer Heidelberg 169-76

[11] Willebeek-LeMair M H, Reeves A P 1993 Strategies for dynamic load balancing on highly parallel computers *IEEE Transactions Parallel and Distributed* Systems **4**(9) 979-93

[12] You T, Li W, Fang Z, et al 2014 Performance Evaluation of Dynamic Load Balancing Algorithms *TELKOMNIKA Indonesian Journal of Electrical Engineering* **12**(4)

[13] Bahi J M, Contassot-Vivier S, Couturier R 2005 Dynamic load balancing and efficient load estimators for asynchronous iterative algorithms *IEEE Transactions on Parallel and Distributed Systems*, **16** (4) 289-99

[14] Lau S M, Lu Q, Leung K S 2006 Adaptive load distribution algorithms for heterogeneous distributed systems with multiple task classes *IEEE Transactions Parallel and Distributed Computing* **66**(2) 163-80

[15] Armbrust M, Fox A, Griffith R, Joseph A D, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, Zahariaa A 2010 A view of cloud computing *Communications of the ACM* **53**(4) 50-58

[16] Moreno-Vozmediano R, Montero R S, Llorente I M 2013 Key Challenges in Cloud Computing: Enabling the Future Internet of Services *Internet Computing IEEE* **17**(4) 18-25

[17] Muir S, Tavilla R, Verghese B 2012 *Vmware Technica Journal* [*EB/OL*] **1**(1)

Authors

Gaochao Xu, Wuhan, born in 1966

Current position, grades: Changchun, Professor *and* PhD supervisor of College of Computer Science and Technology, Jilin University, China.
University studies: BS, MS and PhD on College of Computer science and Technology of Jilin University in 1988, 1991 and 1995
Scientific interest: Cloud Computing, Mobile Cloud Computing
Publications: SCI 10
Experience: more than 10 national, provincial and ministerial level research projects of China Gaochao Xu was.
Research interests: distributed system, grid computing, cloud computing, Internet of things, information security, software testing and software reliability assessment, etc.

Yunmeng Dong, born in 1989, in Yushu of Jilin province of China

Current position, grades: Changchun, Master
University studies: bachelor degree in computer science at Changchun University of Technology (2012), a postgraduate candidate of the college of computer science and technology of Jilin University
Scientific interest: Virtualization, Cloud Computing, Mobile Cloud Computing
Publications: EI 1
Research interests: distributed system, cloud computing and virtualization technology

Xiaodong Fu, ChangChun

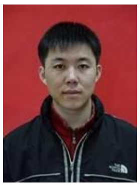
Current position, grades: Senior engineer in the College of Computer Science and Technology， Jilin University of China.
University studies: BSc degree from Jilin University.
Research interests: Distributed System, Grid Computing, Cloud Computing, Internet Things
Publications: 14 research articles

Yan Ding, born in 1988, in Yichun of Heilongjiang province of China

Current position, grades: Changchun, Master, a postgraduate candidate of the college of computer science and technology of Jilin University
University studies: bachelor degree at Jilin University in 2011
Scientific interest: Virtualization, Cloud Computing, Mobile Cloud Computing
Publications: SCI 1
Research interests: distributed system, cloud computing and virtualization technology

Peng Liu, born in 1990, in Jixi of Heilongjiang province of China

Current position, grades: Changchun, Master, is a postgraduate candidate of the college of computer science and technology of Jilin University
University studies: bachelor degree at Daqing Normal University in 2013
Research interest: Virtualization, distributed system Cloud Computing, Mobile Cloud Computing, SDN



Jia Zhao, born in 1982, in Changchun of Jilin province of China

Current position, grades: Changchun, Doctor, PhD candidate of the college of computer science and technology of Jilin University
University studies:
Scientific interest: Virtualization, Cloud Computing, Mobile Cloud Computing
Publications: SCI 4
Research interests Virtualization, Cloud Computing, Mobile Cloud Computing include distributed system, cloud computing, network technology
Experience: participated in several projects

Finite element analysis of fluid conveying pipeline of nonlinear vibration response

Gongfa Li*, Wentao Xiao, Guozhang Jiang, Jia Liu

College of Machinery and Automation, Wuhan University of Science and Technology, Wuhan 430081, China

Received 1 March 2014, www.tsi.lv

Abstract

Fluid filled pipe system was widely used in the city water supply and drainage, water power, chemical machinery, aerospace, marine engineering and the nuclear industry and other fields, it was play an important role for improving the living standards of the nation and the national economic strength. Pipe conveying fluid was easy to design and manufacture, according to the characteristics of fluid conveying pipe, transformed the axial vibration mathematical model of the fluid conveying pipe, which considerate the fluid solid coupling to the beam element model for two nodes. Using Lagrangian interpolation function, the first order Hermit interpolation function and the Ritz method to obtain the element standard equation, and then integrated a global matrix equation, obtained the response of conveying fluid pipe with the Newmark method and Matlab. With the Matlab to simulate the axial motion equation of the conveying fluid pipe, study the response of the system in two aspect of fluid pressure disturbance and the fluid velocity disturbance, and the simulation results are analysed, which provides theoretical support for the work of fluid conveying pipes.

Keywords: response, MATLAB, Numerical simulation, nonlinear vibration

1 Introduction

Study of the fluid solid coupling vibration of pipeline in our country started relatively late, there is no paper of this aspect until the mid80's in last century, and compared with the international level there are still large differences [1-2]. In recent years, with the development of China's modern industry and city modernization, the domestic scholars made a lot of significance research for the fluid solid coupling phenomenon on the long distance oil pipeline, water pipeline, large-scale city heat supply system and nuclear power plant water circulation system, which puts forward the many methods to control pipe vibration. In the aspect of modelling of the fluid conveying pipe, Yang Ke make the pipeline axial vibration of fluid solid coupling four equation model as the foundation, introduced the two order differential equations group with symmetrical "rigidity", "damping" and "quality" matrix, which regard the displacement as basic variables, both considerate the Poisson coupling and coupling of friction and damping of pipeline [3-5]. Fei Wenping established the fluid solid coupling model of a complex pipeline system by the complex modal theory, and studied theoretically. Chen Guiqing pointed out and corrected many error equation for the current mathematical modelling of the pipeline system vibration, and reclassify the pipeline according to the ground of linear pipeline, the ground of nonlinear linear pipeline and buried pipeline, last come out the most commonly used pipeline vibration differential equation. In this paper, considering the infusion pipe liquid vibration condition of small deformation, take the mathematical model of axial vibration as plane beam element, obtained the standard equation with first-order Hermit interpolation function and Ritz method. Then obtained matrix equation through made each unit assembled into the global mass matrix, global damping matrix, the global stiffness matrix and the global load matrix. Apply the finite element method to get the numerical solution for partial differential equation of higher order, and obtained the response of pipe conveying fluid system with Newmark method [11-13].

2 The establish of mathematical model for the output response of the fluid conveying pipe

Taking into account the pipe ratio for length to diameter is relatively large, the deformation of radial is the same, just only exists a certain angle difference, which can be regarded the pipe as plane beam element to consider, using two node element, as shown in Figure 1, the node number of I and J. The conveying fluid pipe is only affected by the lateral force, no axial force, so analysis with the two node element, the nodal displacement model can be defined.

* ***Corresponding author*** e-mail: ligongfa@wust.edu.cn

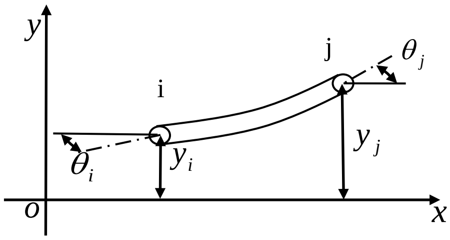


FIGURE 1 An example. Good quality with clear lettering

2.1 A SUBSECTION THE ESTABLISH OF MATHEMATICAL MODEL

The node number of I and J. The conveying fluid pipe is only affected by the lateral force, no axial force, so analysis with the two node element, the nodal displacement model can be defined [6-8]:

$$y(t) = [y_i, \theta_i, y_j, \theta_j]^T . \tag{1}$$

In the node parameter of unit, in addition to the node value of field function, also contains node value for a derivative of the field function. In order to maintain the continuity of field function derivative between the public node element, and in the end nodes to keep the derivative order is first for the field function, so the first-order Hermit interpolation polynomial is used:

$$N(\xi) = [N_1(\xi), N_2(\xi), N_3(\xi), N_4(\xi)] , \tag{2}$$

where $N_1(\xi) = 1 - 3\xi^2 + 2\xi^3$, $N_2(\xi) = \xi - 2\xi^2 + \xi^3$, $N_3(\xi) = 3\xi^2 - 2\xi^3$, $N_4(\xi) = \xi^3 - \xi^2$, ξ is the local dimensionless coordinate ($0 \le \xi \le 1$).

Taken the Ritz method interpolation functions to establish standard unit equation of the approximate solution of the lateral vibration after determine the interpolation function:

$$[M^e]\{\ddot{y}_j\} + [C^e]\{\dot{y}_j\} + [K^e]\{y_j\} = [Q^e] , \tag{3}$$

where $[M^e] = [M^e_{ij}]$, $[C^e] = [C1^e_{ij}] + [C2^e_{ij}]$, $[K^e] = [K1^e_{ij}] + [K2^e_{ij}] + [K3^e_{ij}]$,

$[Q^e] = [Q1^e_{ij}] + [Q2^e_{ij}] + [Q3^e_{ij}]$,

$$[M^e_{ij}] = \int_l N_i (m_p + m_f) N_j dx ,$$

$$[C1^e_{ij}] = \int_l N_i (2 m_f v_f) \frac{\partial N_j}{\partial x} dx ,$$

$$[C2^e_{ij}] = \int_l N_i (\frac{A_f}{c^2} \frac{\partial P}{\partial t}) N_j dx ,$$

$$[K1^e_{ij}] = \int_l \frac{\partial^2 N_i}{\partial^2 x} (EI) \frac{\partial^2 N_j}{\partial^2 x} dx ,$$

$$[K2^e_{ij}] = -\int_l \frac{\partial N_i}{\partial x} (m_f v_f^2 + (1 - 2\gamma) A_f P) \frac{\partial N_j}{\partial x} dx ,$$

$$[K3^e_{ij}] = \int_l N_i (m_f \frac{\partial v_f}{\partial t}) \frac{\partial N_j}{\partial x} dx ,$$

$$[Q2^e_{ij}] = \int_l (-\frac{v_f A_f}{c^2} \frac{\partial P}{\partial t}) N_j dx ,$$

$$[Q3^e_{ij}] = \int_l (m_p g + m_f g) N_j dx .$$

2.2 THE ENTIRETY MATRIX

There are some matrix must be appropriately expanded rewrite when the unit matrix integrated to the entirety matrix so that the matrix of all elements with uniform format, then according to the superposition to assembly.

The usually study boundary constraint conditions include fixed to hinge and fixed to fixed constraints, the mathematical expression of its boundary is given below, respectively (I) and (II).

$$\begin{cases} y(0,t) = 0, \left.\frac{\partial y}{\partial x}\right|_{x=0} = 0 \\ y(L,t) = 0 \end{cases} , \tag{I}$$

$$y(0,t) = 0, \left.\frac{\partial y}{\partial x}\right|_{x=0} = 0 . \tag{II}$$

The four kinds of boundary conditions of above given all belong to the first class constraint conditions, for this kind of constraint conditions can usually use "row column method" and "multiplied with bigger number method". The "multiplied with bigger number method" is make the main diagonal element about the specified node displacement in the overall stiffness matrix with multiply by the large number λ , at the same time, give the specified value of node displacement to the corresponding element of load matrix, then multiply by the same number as well as the main diagonal elements. Using the “multiplied with bigger number method "to deal with the boundary constraint condition by", finally forms the whole matrix:

$$[M]\{\ddot{y}\} + [C]\{\dot{y}\} + [K]\{y\} = [Q] . \tag{4}$$

The Newmark method is used to solving the flow pipeline vibration response, the Newmark method is a step-by-step integration method, the key is to establish the recurrence relations of state vector from t to $t + \Delta t$, assume at the moment of $t + \Delta t$, the $\overset{\sqcup}{y}_{t+\Delta t}$, $\overset{\sqcup}{y}_{t+\Delta t}$ and $y_{t+\Delta t}$ satisfy the dynamics equation:

$$[M]\{\ddot{y}_{t+\Delta t}\} + [C]\{\dot{y}_{t+\Delta t}\} + [K]\{y_{t+\Delta t}\} = [Q_{t+\Delta t}] . \tag{5}$$

In addition to the Newmark method assume the velocity and displacement satisfy the follow equations at the same moment:

$$\dot{y}_{t+\Delta t} = \dot{y}_t + [(1-\alpha)\ddot{y}_t + \alpha\ddot{y}_{t+\Delta t}]\Delta t,\ 0 \le \alpha \le 1, \tag{6}$$

$$y_{t+\Delta t} = y_t + \dot{y}_t\Delta t + [(\frac{1}{2}-\beta)\ddot{y}_t + \beta\ddot{y}_{t+\Delta t}]\Delta t^2,\ 0 \le 2\beta \le 1. \tag{7}$$

According to the analysis results of the algorithm stability when the $\alpha \ge 0.5$, $\beta \ge (\frac{1}{2}+\alpha)^2 \big/ 4$, the Newmark method is unconditionally stable.

The calculation steps of the Newmark method can be summarized as follows.

Step 1: forming the stiffness matrix *K*, mass matrix *M*, damping matrix *C*.

Step 2: obtaining the initial state vector $\ddot{y}_0$, $\dot{y}_0$ and y_0.

Step 3: choosing the time step Δt as well as the parameter α and β, then calculated the constant: $\gamma_0 = \dfrac{1}{\beta\Delta t^2}$, $\gamma_1 = \dfrac{\alpha}{\beta\Delta t}$, $\gamma_2 = \dfrac{1}{\beta\Delta t}$, $\gamma_3 = \dfrac{1}{2\beta} - 1$, $\gamma_4 = \dfrac{\alpha}{\beta} - 1$, $\gamma_5 = \dfrac{\Delta t}{2}(\dfrac{\alpha}{\beta} - 2)$, $\gamma_6 = \Delta t(1-\alpha)$, $\gamma_7 = \alpha\Delta t$.

Step 4: calculation of the effective stiffness matrix:

$$\tilde{K} = K + \gamma_0 M + \gamma_1 C, \tag{8}$$

Step 5: calculation of the effective load vector at the moment of $t+\Delta t$:

$$\hat{Q}_{t+\Delta t} = Q_{t+\Delta t} + M(\gamma_0 y_t + \gamma_2\dot{y}_t + \gamma_3\ddot{y}_t) + C(\gamma_1 y_t + \gamma_4\dot{y}_t + \gamma_5\ddot{y}_t), \tag{9}$$

Step 6: calculation of the displacement at the moment of:

$$\tilde{K}y_{t+\Delta t} = \hat{Q}_{t+\Delta t}, \tag{10}$$

Step 7: calculation of the acceleration and velocity at the moment of

$$\ddot{y}_{t+\Delta t} = \gamma_0(y_{t+\Delta t} - y_t) - \gamma_2\dot{y}_t - \gamma_3\ddot{y}_t, \tag{11}$$

$$\dot{y}_{t+\Delta t} = \dot{y}_t + \gamma_6\ddot{y}_t + \gamma_7\ddot{y}_{t+\Delta t}. \tag{12}$$

3 Simulation and analysis of the axial vibration

For the axial vibration, the constraint of the fixed to hinge and constraint of the fixed - overhanging belonging to the same constraints, which is decided by its displacement and load form, While the boundary constraint of the hinge to hinge, because the axial without any restraint, this pipeline system is in an unstable state, the modal and response could not be calculated. So the axial vibration simulation of the fluid conveying pipeline was the main considerate the constraint of the fixed to hinged and constraint of the fixed - overhanging. In the simulation process, the two end points of pipe as the supporting point, and assumed to be rigid constrain. Then make the pipe length divided into 100 equal parts, a total of 101 nodes, in the process of analysis with room temperature water as the fluid and rolling copper as pipe material [9-10].

3.1 A SUBSECTION THE VIBRATION RESPONSE OF THE FLUID CONVEYING PIPELINE

The four order mode of vibration in two kinds of boundary conditions. The first four order mode of vibration of the constraint of fixed to hinge as shown in Figure 2. The first four order mode of vibration of the constraint of fixed to fixed hinge as shown in Figure 3.

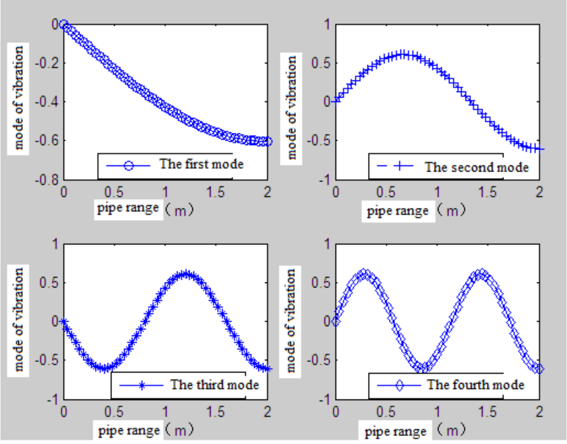


FIGURE 2 The first four order mode of vibration of the constraint of fixed to hinge

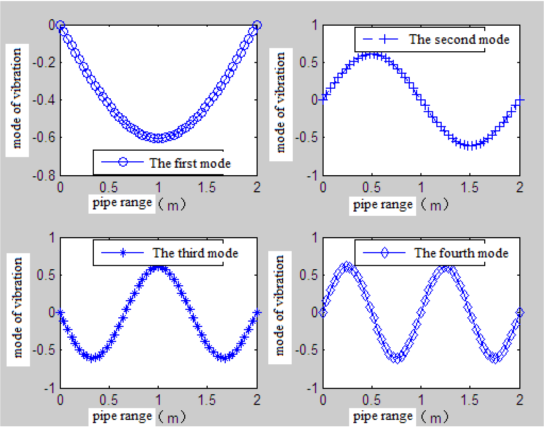


FIGURE 3 The first four order mode of vibration of the constraint of fixed to fixed to hinge

3.2 THE TWO BOUNDARY CONSTRAINTS VIBRATION RESPONSE

As shown in Figure 4 is the vibration response of same node in different time for the constraint of the fixed to hinge, we can get the vibration cycle of the tenth nodes, fifty-first nodes and hundred and first nodes from the figure shows, the amplitude becomes larger from tenth nodes, fifty-first nodes, hundred and first nodes. At the end of pipe the hinge hundred and first nodes is the maximum amplitude with match the actual situation. As shown in Figure 5 is the vibration response of same node in different time for the constraint of the fixed to fixed, compared with the fixed to hinge node, the node vibration graphs is same, but for the vibration amplitude the latter is small, the vibration cycle is small, in addition to the fifty-first node is the maximum vibration amplitude of the constraint of the fixed to fixed, that match with the actual situation.

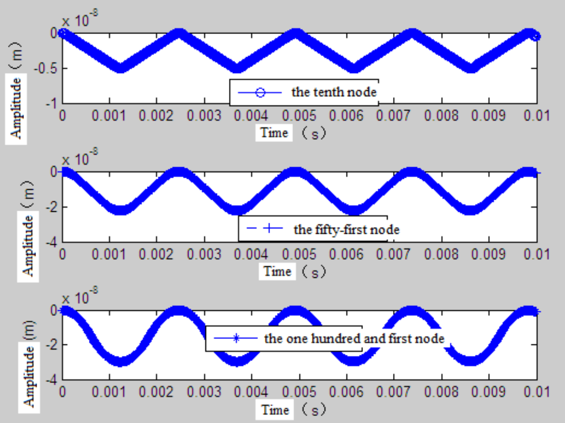


FIGURE 4 The vibration response of same node in different time for the constraint of the fixed to hinge

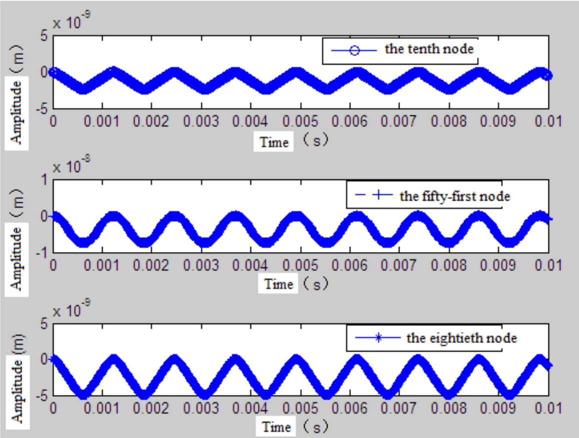


FIGURE 5 The vibration response of same node in different time for the constraint of the fix to fix

3.3 THE EFFECT OF VELOCITY AND PRESSURE FOR THE VIBRATION RESPONSE

The effect of fluid velocity perturbation for response of the fluid conveying pipeline The effect of fluid velocity perturbation for system response of the constraint of the fixed to hinge as shown Figure 6.

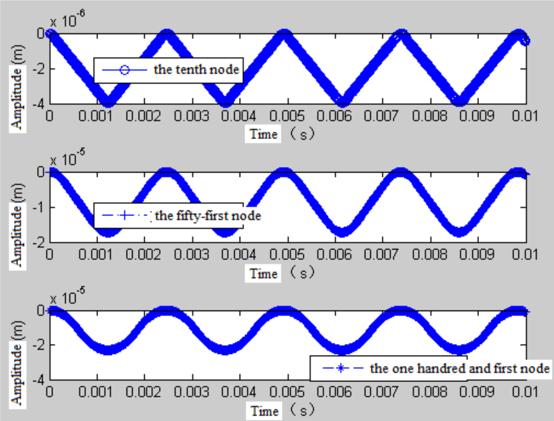


FIGURE 6 The effect of fluid velocity perturbation for response of the constraint of the fix to fix

From the Figure 6 we can see the response of node periodic change in velocity perturbation, when the fluid velocity and pressure of nodes is a constant value, the vibration amplitude increases about 100 times with the time increased. The boundary conditions for the constraint of the fix to fix, the response of nodes has a cycle changes under the disturbance velocity, which the vibration amplitude increased about 100 times.

The effect of the fluid pressure disturbance is for response of the fluid conveying pipeline. From the Figure 7 we can see the response of node periodic change in velocity perturbation in the constraint of the fix to hinge, when the fluid velocity and pressure of nodes is a constant value, the amplitude of vibration is increased, but not very strong, the response is much smaller than the velocity perturbation. The boundary conditions for the response of the constraint of fix to fix constraint nodes is almost to the same with fix to hinge.

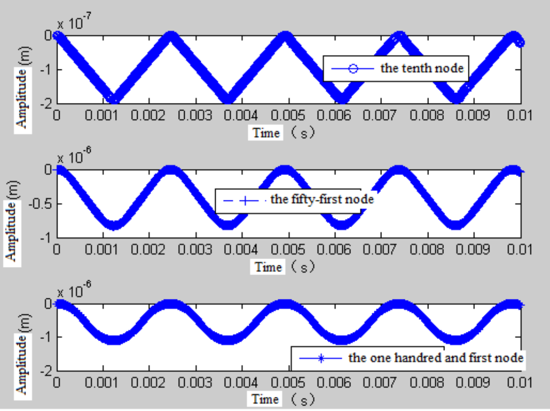


FIGURE 7 The effect of the fluid pressure disturbance for response of the constraint of the fix to hinge

3.4 CHARACTERISTICS COMPARATIVE ANALYSIS OF AXIAL VIBRATION OF PIPELINE

The Table 1 gives the first four order natural frequency for response of the fluid conveying pipeline of the constraint of the fix to hinge and the fix to fix. Can see from the table, the axial vibration natural frequency of the fluid conveying pipeline system is bigger, the first-order natural frequency of the constraint of the fix to fix is two times that of the fix to hinge.

TABLE 1 The axial vibration of pipeline system in two kinds of boundary characteristics

Frequency	Fixed hinge constraint	Fixed constraint
The first-order natural frequency (Hz)	407.0157	814.0566
The second-order natural frequency (Hz)	1221.1476	1628.314
The third-order natural frequency (Hz)	2035.5808	2442.9732
The fourth-order natural frequency (Hz)	2850.5163	3258.2352

The effect for response of the fluid conveying pipeline of velocity and pressure for the vibration response is bigger, in a general it will increase 100 times.

4 Conclusion

The simulation for the response of the fluid conveying pipe axial motion summarized and analysed. In two kinds of boundary conditions, analysed the effect for the vibration response of the fluid conveying pipeline of velocity and pressure disturbance, verified the correctness of the established vibration model of the fluid conveying pipeline, and obtained the vibration characteristics of the fluid conveying pipe of velocity and pressure disturbance in the constraint of the fix to fix and fix to hinge, which provides theoretical support for the work of fluid conveying pipes.

References

[1] Kubenko V D, Koval'chuk P S, Kruk L A 2011 *J International Applied Mechanics* **47** 636-44

[2] Kubenko V D, Koval'chuk P S, Kruk L A 2011 *J International Applied Mechanics* **47** 1-9

[3] Li Shuai-Jun, Liu Gong-Min, Kong Wei-Tao 2014 *Nuclear Engineering and Design* **266** 78-88

[4] Zhu Qing-Jie, Chen Yan-Hua, Liu Ting-Quan, Dai Zhao-Li 2008 Finite element analysis of fluid-structure interaction in buried liquid-conveying pipeline *Journal of Central South University of Technology* **1**(15) 307-10 *(in Chinese)*

[5] Zhai Hong-Bo, Wu Zi-Yan, Liu Yong-Shou, Yue Zhu-Feng *2011 Nuclear Engineering and Design* **241** 2744-9

[6] Yi-min Huang, Yong-shou Liu, Bao-hui Li, Yan-jiang Li, and Zhu-feng Yue 2010 *Nuclear Engineering and Design* **240** 461-7

[7] Wang Xiaoyang, Zhou, Zheng, Liu Xingpei *Advanced Materials Research* **655** 620-4

[8] Bao Ri-Dong, Jin Zhi-Hao and Wen Bang-Chun 2008 Analysis of nonlinear dynamic characteristics of commonly supported fluid conveying pipe *Zhendong yu Chongji/Journal of Vibration and Shock.* **27** 87-90 *(in Chinese)*

[9] He Er-Ming, Liu Feng, Hu Ya-Qi, Zhao Zhi-Bin 2013 Nonlinear vibration response analysis and fatigue life prediction of a thin-walled structure under thermal-acoustic loading *Zhendong yu Chongji/Journal of Vibration and Shock* **24** 135-9 *(in Chinese)*

[10] Dong Xing-Jian, Peng Zhi-Ke, Zhang Wen-Ming, Meng Guang 2013 Parametric characteristic of the random vibration response of nonlinear systems *Acta Mechanica Sinica/Lixue Xuebao* **29**(2) 267-83

[11] Yang Ke, Zhang Lixiang, Wang Bindi 2005 A symmetric model of liquid fiied pipe in axial vibration *Journal of hydrodynamics* **20** 8-13 *(in Chinese)*

[12] Li Hai liang, YAN Jin li, Wang Xu feng 2013Fluid structure coupling analysis of turbine runner based on ANSYS *Journal of Mechanical & Electrical Engineering* **30** 1093-6 *(in Chinese)*

[13] Chen Guo, He Li dong, Han Wan fu, Pei Zheng wu 2013 Vibration analysis of butylenes centrifugal pump and plunger pump pipeline and research of vibration-damping technology *Journal of Mechanical & Electrical Engineering* **30** 167-70 *(in Chinese)*

Authors

Gongfa Li, born on July 10, 1979, Honghu, China

Current position, grades: Associate professor, College of Machinery and Automation, Wuhan University of Science and Technology
Scientific interests: computer aided engineering, mechanical CAD/CAE, Modelling and optimal control of complex industrial process.
Publications: 60
Experience: Dr. Gongfa Li received the Ph.D. degree in mechanical design and theory from Wuhan University of Science and Technology in China. Currently, he is an associate professor at Wuhan University of Science and Technology, China. His major research interests include modelling and optimal control of complex industrial process. He is invited as a reviewer by the editors of some international journals, such as Environmental Engineering and Management Journal, International Journal of Engineering and Technology, International Journal of Physical Sciences, International Journal of Water Resources and Environmental Engineering, etc. He has published nearly twenty papers in related journals.

Wentao Xiao, born in May 11, 1989, Tianmen, China

Current position, grades: Currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology
Scientific interests: mechanical CAD/CAE, signal analysis and processing.
Experience: Wentao Xiao was born in Hubei province, P. R. China, in 1989. He received B.S. degree in mechanical engineering and automation from City College of Wuhan University of Science and Technology, Wuhan, China, in 2013. He is currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology. His current research interests include mechanical CAD/CAE, signal analysis and processing.

Guozhang Jiang, born in December 15, 1965, Tianmen, China

Current position, grades: Professor of Industrial Engineering, and the Assistant Dean of the college ofmachinery and automation, Wuhan University of Science and Technology.
University studies: He received the B.S. degree in Chang' an University, China, in 1986, and M.S. degree in Wuhan University of Technology, China, in 1992. He received the Ph.D. degree in mechanical design and theory from Wuhan University of Science and Technology, China, in 2007.
Scientific interests: computer aided engineering, mechanical CAD/CAE and industrial engineering and management system.
Publications: 100
Experience: Guozhang Jiang was born in Hubei province, P. R. China, in 1965. He is a Professor of Industrial Engineering, and the Assistant Dean of the college of machinery and automation, Wuhan University of Science and Technology. Currently, his research interests are computer aided engineering, mechanical CAD/CAE and industrial engineering and management system.

Jia Liu, born in 1990, Shanxi, China

Current position, grades: Currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology
Scientific interests: mechanical CAD/CAE, signal analysis and processing.
Experience: Jia Liu was born in Shanxi province, P. R. China, in 1990. He received B.S. degree in mechanical engineering and automation from Wuchang institute of Technology, Wuhan, China, in 2012. He is currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology. His current research interests include mechanical CAD/CAE, signal analysis and processing.

Birkhoff normal forms for the wave equations with nonlinear terms depending on the time and space variables

Yi Wang[*]

School of Mathematics and Quantitative Economic, Shandong University of Finance and Economics, Jinan, Shandong 250014, P. R. China

Received 1 March 2014, www.tsi.lv

Abstract

The one-dimensional (1D) quasi-periodically forced nonlinear wave equation with periodic boundary conditions is considered. It is proved that there is a real analytic and symplectic change of coordinates, which can transform the Hamiltonian to the Birkhoff normal form.

Keywords: Infinite dimensional Hamiltonian system, quasi-periodically forced nonlinear wave equation, quasi-periodic solution, periodic boundary condition, Birkhoff normal form

1 Introduction and main results

In this paper, we are concerned with the quasi-periodically forced nonlinear wave equation

$$u_{tt} - u_{xx} + \mu u + \varepsilon g(\omega t, x)h(u) = 0, \quad \mu > 0, \quad x \in \mathbb{T} = \mathbb{R}/2\pi\mathbb{Z} \tag{1.1}$$

under the periodic boundary conditions

$$u(t,x) = u(t, x+2\pi), \tag{1.2}$$

where ε is a small positive parameter; the function $g(\omega t, x) = g(\vartheta, x)$, $(\vartheta, x) \in \mathbb{T}^m \times \mathbb{T}$ is real analytic in (ϑ, x) and quasi-periodic in t with frequency vectors $\omega = (\omega_1, \omega_2 \ldots, \omega_m) \in [\varrho, 2\varrho]^m$ for some constant $\varrho > 0$; and the nonlinearity h is a real analytic function of the form $h(u) = u^3 + \mathcal{O}(u^4)$.

The technology of the Birkhoff normal forms has been widely used in the study of the dynamics of Hamiltonian systems close to elliptic equilibrium points. For example, obtaining Birkhoff normal forms of the Hamiltonians is the most important step of the KAM approach, which is one of the main tools to deal with the existence of periodic and quasi-periodic solutions of nonlinear PDEs.

This paper is devoted to transform the Hamiltonians of a kind of wave equations to the four-order Birkhoff normal forms. This kind of systems contains nonlinear terms with quasi-periodically forcing and the space variable. We obtain a quantitative description about the Hamiltonian's proposition in a ball of a Sobolev type phase space. The result in this paper provides a basis for the forthcoming research of the existence of periodic or quasi-periodic solutions. The method in this paper can be considered as an idea to deal with the infinite-dimensional systems whose nonlinear terms depend on the time or space variables.

For the Birkhoff normal forms of wave equations under Dirichlet boundary conditions, the reader is referred to [1-4]. However, the partial differential equations with periodic boundary conditions are more complicated since the eigenvalues are not distinct but multiple. This fact would bring a lot of trouble in constructing normal forms. The reason mainly lies in the notorious “small divisor problem”, which makes it difficult to obtain the regularity of the symplectic transformations. In [5], the author studied the completely resonant nonlinear wave equation under periodic boundary conditions. But the difficulty caused by the multiplicity of eigenvalues was avoided since the author only considered the even solutions. Articles [6] and [7] succeeded in constructing Birkhoff normal forms of wave equations with periodic boundary conditions and proved that the existence of quasi-periodic solutions. However, their results cannot be used in equations with constant potential.

In this paper, we are interested in the nonlinear wave equations with constant potential and with the nonlinear terms depending on time or space variables. In fact, Berti and Procesi [8] considered the periodically forced wave equations: $\begin{cases} v_{tt} - v_{xx} + f(\omega_1 t, v) = 0 \\ v(t,x) = v(t, x+2\pi), \end{cases}$ with the nonlinear forcing term: $f(\omega_1 t, v) = a(\omega_1 t)v^{2d-1} + \mathcal{O}(v^{2d})$, $d > 1$, $d \in \mathbb{N}_+$ being $2\pi/\omega_1$ -periodic in time t.

Zhang and Si [9] focused on the quasi-periodically forced nonlinear wave equations:

[*] ***Corresponding author*** e-mail: yiwang8080@126.com

$u_{tt} - u_{xx} + \mu u + \varepsilon\phi(t)h(u) = 0,\quad \mu > 0$ with Dirichlet boundary conditions, where ϕ is real analytic quasi-periodic function and $h(u) = \eta_1 u + \eta_{2r+1}u^{2r+1} + \sum_{k\ge r+1}\eta_{2k+1}u^{2k+1}$, $\eta_1, \eta_{2r+1} \neq 0,\quad r \in \mathbb{N}$.

In the above equations, one needs to deal with essentially finite small divisors. Moreover, the above equations exclude those cases where the nonlinear terms contain the space variable, while in this paper; we provide an idea to deal with those cases. Factually, in those cases, the important “compactness property” cannot hold. Thus, one would confront essentially infinite small divisors. To overcome this point, we truncate the unperturbed term as well as the perturbed term. Therefore, although the “compactness property” is not satisfied, we can also estimate the measure of the small divisors. Our main Theorem 3.1 proves that there is a canonical transformation, which can change the Hamiltonian to a four-order Birkhoff normal form.

The paper is organized as follows. In section 2, we will give the expression of Hamiltonian. Section 3 is devoted to the Birkhoff normal form of the Hamiltonian.

2 Hamiltonian setting

Throughout this paper, we assume that:

(H) $g_0 := \lim_{T\to\infty}\frac{1}{T}\int_0^T g(\omega t, x)dt \equiv const.\quad 0 \neq g_0 \in \mathbb{R}.$

For $f \equiv 0$, the equation (1.1) becomes:

$$u_{tt} - u_{xx} + \mu u = 0. \tag{2.1}$$

The operator $A = -\frac{d^2}{dx^2} + \mu$ with periodic boundary conditions admits a complete orthogonal basis of eigenfunctions $\phi_j \in L^2([0,2\pi]),\quad j \in \mathbb{Z},$ with corresponding eigenvalues $\zeta_j = j^2 + \mu,$ if one sets $\phi_0 = 1/\sqrt{2\pi}$ and for $j \ge 1,\quad \phi_j(x) = \frac{1}{\sqrt{\pi}}\cos(jx),$ $\phi_{-j}(x) = \frac{1}{\sqrt{\pi}}\sin(jx)$.

Every solution of the linear wave equation (2.1) can be written as a super-position of the basic modes ϕ_j, namely, for $\mathcal{I}$ any subset of $\mathbb{Z}$ and $\mu_j := \sqrt{\zeta_j}$, $u(x,t) = \sum_{j\in\mathcal{I}}\xi_j\cos(\mu_j t + \theta_j)\phi_j(x)$, with amplitudes $\xi_j > 0$ and initial phases θ_j.

In the whole of this paper, we denote by C the universal constants if we do not care their values. For some $\sigma_1 > 0$ and $\sigma > 0$, we suppose that g analytically in ϑ, x extends to the domain $D_1(\sigma_1)\times D(\sigma)$, where $D_1(\sigma_1) = \{\vartheta \,|\, |\mathrm{Im}\vartheta| < \sigma_1\}$ and $D(\sigma) = \{x \,|\, |\mathrm{Im}x| < \sigma\}$.

We rewrite the wave equation (1.1) as follows:

$$\dot{u} = v,\quad \dot{v} + Au = -\varepsilon g(\omega t, x)h(u), \tag{2.2}$$

where $A = -d^2/dx^2 + \mu, t \in \mathbb{R}$. As is well known, the equation (2.2) can be studied as an infinite dimensional Hamiltonian system by taking the phase space to be product of the Sobolev spaces $H_0^1([0,2\pi])\times L^2([0,2\pi])$ with coordinates u and $v = \partial_t u$. The Hamiltonian for (2.2) is then $H = \frac{1}{2}(v,v) + \frac{1}{2}(Au,u) + \varepsilon\int_{\mathbb{T}}\chi(u,x,\omega t)dx$, where $\chi(u,x,\omega t) = g(\omega t, x)[\frac{1}{4}u^4 + \mathcal{O}(u^5)]$, and $(\cdot,\cdot)$ denotes the usual scalar product in $L^2([0,2\pi])$.

We introduce the coordinates $q = (q_0, q_1, q_{-1}, \ldots)$ and $p = (p_0, p_1, p_{-1}, \ldots)$ by setting $u(t,x) = \sum_{j\in\mathbb{Z}} q_j(t)\phi_j(x)$, $v = \sum_{j\in\mathbb{Z}} p_j(t)\phi_j(x)$.

The coordinates are taken from some Banach space $l_b^s (s > 0)$ of all real valued bi-infinite sequences $q = (q_0, q_1, q_{-1}, \ldots)$ with finite norm $\|q\|_s = \sum_{j\in\mathbb{Z}}(j)^s|q_j|$, where $(j) = \max(1, |j|)$. We can obtain the Hamiltonian: $H = \Lambda + G$, where

$$\Lambda = \frac{1}{2}\sum_j(\mu_j^2 q_j^2 + p_j^2),\quad G = \varepsilon\int_{\mathbb{T}}\chi\left(\sum_{j\in\mathbb{Z}} q_j(t)\phi_j(x), x, \omega t\right)dx$$

and $\mu_j = \sqrt{\zeta_j}$.

The equations of motion are: $\dot{q}_j = \frac{\partial H}{\partial p_j} = p_j,\quad \dot{p}_j = -\frac{\partial H}{\partial q_j} = -\mu_j^2 q_j - \frac{\partial G}{\partial q_j}$ with respect to the symplectic structure $\sum dp_i \wedge dq_i$ on $l_b^s \times l_b^s$.

To make the system turn into an autonomous system, we introduce a pair of action-angle variables $(J,\vartheta)\in\mathbb{R}^m\times\mathbb{T}^m$ $(\mathbb{T}^m := \mathbb{R}^m/2\pi\mathbb{Z}^m)$ by assuming that $\vartheta = \omega t$. Then, $\dot{q}_j = \frac{\partial H}{\partial p_j},\quad \dot{p}_j = -\frac{\partial H}{\partial q_j},\quad \dot{\vartheta} = \omega,$ $\dot{J} = -\frac{\partial G}{\partial\vartheta} = -\varepsilon\frac{\partial\int_{\mathbb{T}}\chi dx}{\partial\vartheta}$ can be written as a Hamiltonian system (with respect to the symplectic structure $(d\vartheta\wedge dJ + \sum dp_i \wedge dq_i)$ with the Hamiltonian:

$$H = <\omega, J> + \frac{1}{2}\sum_j(\mu_j^2 q_j^2 + p_j^2) + G(q,\vartheta). \tag{2.3}$$

To continue our investigation of the Hamiltonian (2.3), we need to establish the regularity of the nonlinear Hamiltonian vector field X_G associated to G, where $<\cdot,\cdot>$ is the standard inner product in $\mathbb{C}^m$.

To this end, let l_b^2 and L^2, respectively, be the Hilbert spaces of all bi-infinite, square summated sequences with complex coefficients and all square-integrable complex-valued functions on $[0,2\pi]$. Let $\mathcal{F}: l_b^2 \to L^2,\ \ q \mapsto \mathcal{F}q = \sqrt{\frac{1}{\pi}}\sum_j q_j e^{ijx}$ be the inverse discrete Fourier transform, which defines an isometry between the two spaces. Let $s \ge 1$. The subspaces $l_b^s \subset l_b^2$ consist, by definition of all bi-infinite sequences with the finite form $\|q\|_s = \sum_j (j)^s |q_j|$. Through $\mathcal{F}$ we define subspaces $H^s[0,2\pi] \subset L^2[0,2\pi]$ that are normalized by setting $\|\mathcal{F}q\|_s = \|q\|_s$.

The following lemma was proved in [7], we only give the result.

Lemma 2.1 For all $s > 0$, the space l_b^s is a Banach algebra with respect to convolution of the sequences $(q*p)_j := \sum_k q_{j-k}p_k$, and $\|p*q\|_s \le 2^s \|q\|_s \|p\|_s$.

Using the above lemma, we can prove the following lemma.

Lemma 2.2 For all $s \ge 1$, the gradient $\partial_q G$ is real analytic as a map from some neighbourhood of origin in $l_b^s \to l_b^s$, with $\|\partial_q G\|_s = \varepsilon\mathcal{O}(\|q\|_s^3)$.

Proof Let $q \in l_b^s$. Consider as a function on $[0,2\pi]$, $u = \sum q_j\phi_j$ is in H^s with $\|u\|_s \le \|q\|_s$. From Assumption **(H)**, we assume

$$g(\vartheta,x) = g_0 + \sum_{|k|\ge1}\left[\sum\nolimits'_{\tau} g_k^{\tau} e^{\mathrm{i}\tau x}\right] e^{\mathrm{i}<k,\vartheta>}$$
$$= \sqrt{\frac{1}{\pi}}\left\{\sqrt{\pi}g_0 + \sum\nolimits'_{\tau}\sqrt{\pi}\left[\sum_{|k|\ge1} g_k^{\tau} e^{\mathrm{i}<k,\vartheta>}\right] e^{\mathrm{i}\tau x}\right\},$$

where the prime symbol in the summation sign indicates that the sum runs over all $\tau \in \mathbb{Z}$. By using of Lemma A.1 in [10], $|g_k^{\tau}| \le \|g(\vartheta,x)\|_{D(\sigma_1)\times D(\sigma)} e^{-|k|\sigma_1} e^{-|\tau|\sigma}$.

Furthermore, for $\vartheta \in D(\frac{\sigma_1}{2})$,

$$\left|\sum_{|k|\ge1} g_k^{\tau} e^{\mathrm{i}<k,\vartheta>}\right| \le \|g(\vartheta,x)\|_{D(\sigma_1)\times D(\sigma)} e^{-|\tau|\sigma} \sum_{|k|\ge1} e^{-|k|\sigma_1} e^{|k|\frac{\sigma_1}{2}}$$
$$\le C\|g(\vartheta,x)\|_{D(\sigma_1)\times D(\sigma)} e^{-|\tau|\sigma},$$

because of the convergence of the series $\sum_{|k|\ge1} e^{-|k|\frac{\sigma_1}{2}}$.

Hence, for $(\vartheta,x) \in D(\frac{\sigma_1}{2}) \times D(\frac{\sigma}{2})$,

$$\|g(\vartheta,x)\|_s = |\sqrt{\pi}g_0| + C\sqrt{\pi}\,\|g(\vartheta,x)\|_{D(\sigma_1)\times D(\sigma)} \cdot \sum\nolimits_{|\tau|\ge0} (\tau)^s e^{-|\tau|\sigma} e^{|\tau|\frac{\sigma}{2}} \le C, \tag{2.4}$$

because of the convergence of the series $\sum_{|\tau|\ge0} (\tau)^s e^{-|k|\frac{\sigma}{2}}$, where C depends on g, σ_1, s and σ. That is $g(\cdot,\cdot) \in H^s[0,2\pi]$. By the algebra property and the analyticity of g and h from (2.4), the function $g(\vartheta,x)h(u)$ also belongs to $H^s[0,2\pi]$ with $\|g(\vartheta,x)h(u)\|_s \le C\|q\|_s^3$ in a sufficiently small neighbourhood of the origin, where C depends on s, σ, σ_1 and g. On the other hand, since $\frac{\partial G}{\partial q_j} = \varepsilon\int_{\mathbb{T}} g(\vartheta,x)h(u)\phi_j(x)dx$.

The components of G_q are the Fourier coefficients of $g(\vartheta,x)h(u)$, so G_q belongs to l_b^s, with $\|G_q\|_s \le \varepsilon\cdot\|g(\vartheta,x)h(u)\|_s \le \varepsilon C\|q\|_s^3$. The regularity of G_q follows from the regularity of its component and its local boundedness.

3 Partial Birkhoff normal forms

Since $\chi(u,x,\vartheta) = g(\vartheta,x)[\frac{1}{4}u^4 + \mathcal{O}(u^5)]$ and $u = \sum_j q_j\phi_j$, we find that

$$G(q,\vartheta) = \frac{\varepsilon}{4}\sum_{i,j,d,l}\int_{\mathbb{T}} g(\vartheta,x)\phi_i\phi_j\phi_d\phi_l dx q_i q_j q_d q_l + \varepsilon\mathcal{O}(\|q\|_s^5).$$

From (**H**), we can get that

$$g(\vartheta,x) = g_0 + \sum_{|k|\ge1} g_k(x) e^{\mathrm{i}<\mathrm{k},\vartheta>}. \tag{3.1}$$

It follows from (3.1) that

$$G(q,\vartheta) = \frac{\varepsilon}{4}\sum_{i,j,d,l} G_{ijdl} q_i q_j q_d q_l + \frac{\varepsilon}{4}\sum_{|k|\ge1,i,j,d,l} G_{k,ijdl} e^{\mathrm{i}<\mathrm{k},\vartheta>} q_i q_j q_d q_l + \varepsilon\mathcal{O}(\|q\|_s^5),$$

where

$$G_{ijdl} = g_0\int_{\mathbb{T}}\phi_i\phi_j\phi_d\phi_l dx \quad \text{and} \quad G_{k,ijdl} = \int_{\mathbb{T}} g_k(x)\phi_i\phi_j\phi_d\phi_l dx, \quad |k| \ge 1. \tag{3.2}$$

An easy computation shows that $G_{ijdl} = 0$ unless $i \pm j \pm d \pm l = 0$ for at least one combination of plus and minus signs. In particular, we have

$$G_{ijij}=\begin{cases}\frac{g_0(2+\delta_{ij})}{4\pi}, & if \quad ij>0\\ \frac{g_0(2-\delta_{i(-j)})}{4\pi}, & if \quad ij<0\\ \frac{g_0}{2\pi}, & if \quad ij=0,\end{cases}$$, where

$\delta_{ij}=\begin{cases}1, & i=j\\ 0, & i\neq j.\end{cases}$ This will play an important role later on.

Given a fixed finite subset of indices $\mathcal{I}_N=\{n_1,\dots,n_N\}\subset\mathbb{Z}$ with $|n_i|\neq|n_j|$, if $i\neq j$, we decompose the Hamiltonian (2.3) as $H=H_N+H_\infty$, where

$$H_N=\Lambda_N+\varepsilon G_N, \tag{3.3}$$

$$H_\infty=\Lambda_\infty+\varepsilon G_\infty, \qquad \Lambda_N=<\omega,J>+\frac{1}{2}\sum_{j\in\mathcal{I}_{\mathbb{N}}}(\mu_j^2q_j^2+p_j^2),$$

$$\begin{aligned}G_N(q,\vartheta)&=\frac{1}{4}\sum_{i,j,d,l\in\mathcal{L}_N,i\pm j\pm d\pm l=0}G_{ijdl}q_iq_jq_dq_l\\&+\frac{1}{4}\sum_{|k|\geq1,i,j,d,l\in\mathcal{L}_N}G_{k,ijdl}e^{i<k,\vartheta>}q_iq_jq_dq_l+\mathcal{O}(|q|^5),\end{aligned} \tag{3.4}$$

$$\begin{aligned}&\Lambda_\infty=\frac{1}{2}\sum_{j\notin\mathcal{I}_N}(\mu_j^2q_j^2+p_j^2),\\&and \quad G_\infty=G(q,\vartheta)-G_N(q,\vartheta),\end{aligned} \tag{3.5}$$

where $\mathcal{O}(|q|^5)$ denotes the five order terms, in which all the subscripts of q belong to the subset $\mathcal{I}_N$ and noticing that N is finite.

We introduce the complex coordinates z_j , $j=1,2,\dots,N$ by $z_j=\frac{1}{\sqrt{2\mu_{n_j}}}(\mu_{n_j}q_{n_j}+\mathrm{i}p_{n_j})$ and define $|z|^2=|z_1|^2+\dots+|z_N|^2$ for a vector $z=(z_1,\dots,z_N)$. So we obtains the Hamiltonian:

$$\begin{aligned}&H_N(z,\overline{z})=\Lambda_N+\varepsilon G_N=\\&<\omega,J>+\sum_{j=1,2,\dots,N}\mu_{n_j}|z_j|^2+\varepsilon G_N(z,\overline{z},\vartheta),\end{aligned} \tag{3.6}$$

with symplectic structure $d\vartheta\wedge dJ+\mathrm{i}\sum_j dz_j\wedge\overline{z}_j$, where

$$\begin{aligned}G_N(\overline{z},z,\vartheta)&=\frac{1}{4}\sum_{n_i\pm n_j\pm n_d\pm n_l=0,\,i,j,d,l=1,\dots,N}G_{n_in_jn_dn_l}\\&\cdot\frac{(z_i+\overline{z}_i)}{\sqrt{2\mu_{n_i}}}\frac{(z_j+\overline{z}_j)}{\sqrt{2\mu_{n_j}}}\frac{(z_d+\overline{z}_d)}{\sqrt{2\mu_{n_d}}}\frac{(z_l+\overline{z}_l)}{\sqrt{2\mu_{n_l}}}\\&+\frac{1}{4}\sum_{|k|\geq1}\sum_{i,j,d,l=1,\dots,N}G_{k,n_in_jn_dn_l}e^{\mathrm{i}<\mathrm{k},\vartheta>}\\&\cdot\frac{(z_i+\overline{z}_i)}{\sqrt{2\mu_{n_i}}}\frac{(z_j+\overline{z}_j)}{\sqrt{2\mu_{n_j}}}\frac{(z_d+\overline{z}_d)}{\sqrt{2\mu_{n_d}}}\frac{(z_l+\overline{z}_l)}{\sqrt{2\mu_{n_l}}}+\mathcal{O}(|z|^5)\end{aligned} \tag{3.7}$$

By using the method in [7], for the remaining coordinates, one introduces the notation, for $\nu\geq1$,

$$x'_\nu=\begin{cases}(q_\nu,q_{-\nu})\in\mathbb{R}^2, & \text{if} \quad \nu,-\nu\notin\mathcal{I}_N,\\ q_{-\tilde{\nu}}\in\mathbb{R}, & \text{if} \quad \nu=|\tilde{\nu}| \quad \text{for some} \quad \tilde{\nu}\in\mathcal{I}_N,\end{cases}$$

and similarly for $p_j, j\notin\mathcal{I}_N$, denoted in term of $y_\nu\in\mathbb{R}^{N_\nu},\nu\geq1$, with N_ν as above, namely, $N_\nu=2$ if both $\nu,-\nu\notin\mathcal{I}_N$ and $N_\nu=1$ otherwise. For d_k, $k\geq1$, a sequence of strictly positive integers uniformly bounded by some $\overline{d}<\infty$, let $\mathcal{R}^\infty$ denote the set of infinite sequences $x'=(x'_1,x'_2,\dots)$ with $x'_k\in\mathbb{R}^{d_k}$. Then we can introduce the following family of Banach spaces $\mathcal{R}_s^\infty$, $s\in\mathbb{R}$, $\mathcal{R}_s^\infty=\{Z\in\mathcal{R}^\infty \,\|\, Z|\equiv\sum_{k\geq1}k^s|Z_k|_{\mathbb{R}^{d_k}}\}$. Clearly, for $q,p\in l_b^s$ one has $x',y\in\mathcal{R}_s^\infty$, and $H_\infty(z,\overline{z},p,q)$ reads in these notations

$$H_\infty(z,\overline{z},x',y,\vartheta)=\Lambda_\infty(x',y)+\varepsilon G_\infty(z,\overline{z},x',\vartheta),$$

$$\Lambda_\infty(x',y)=\frac{1}{2}\sum_{\nu\geq1}(\mu_\nu^2|x'_\nu|^2+|y_\nu|^2), \tag{3.8}$$

and $|G_\infty|=\mathcal{O}(\sum_{l=0}^{3}|z|^l\,\|x'\|_s^{4-l})+\mathcal{O}(\sum_{l=0}^{4}|z|^l\,\|x'\|_s^{5-l})$.

Theorem 3.1 Choose ϵ_0 small enough. Consider the Hamiltonian H_N. For each fixed subset $\mathcal{I}_N,N<\infty$, satisfying $|n_i|\neq|n_j|$ when $i\neq j$, there is a subset $\Omega\subset[\varrho,2\varrho]^m$ with $meas\Omega>0$ such that for any $\omega\in\Omega$, and there is a real analytic, symplectic change of coordinates Ψ_N in a complex neighbourhood: $\vartheta\in D(\frac{\sigma_1}{2}):=\{\vartheta|\ |\mathrm{Im}\vartheta|<\frac{\sigma_1}{2},\sigma_1>0\}$ of the tour $\mathbb{T}^m$ and a neighbourhood of the origin in $\mathbb{C}^N$ such that for all $\mu>0$, the Hamiltonian (3.3) can be transformed into $H_N\circ\Psi_N=\Lambda_N+\varepsilon\overline{G}_N+\varepsilon\hat{G}_N+\varepsilon^2K_N+\varepsilon\mathcal{O}(|z|^5)$, where $\hat{G}_N=\epsilon_0\mathcal{O}(|z|^4),K_N=\mathcal{O}(|z|^6)$, and $\overline{G}_N(z,\overline{z})=\frac{1}{2}\sum_{i,j=1}^{N}\overline{g}_{ij}|z_i|^2|z_j|^2$ with uniquely determined coefficient

$$\bar{g}_{ij}=\begin{cases}\dfrac{3g_0}{4\pi\mu_{n_{|i|}}\mu_{n_{|j|}}}, & \text{if}\quad i\neq j;\\ \dfrac{9g_0}{16\pi\mu_{n_{|i|}}\mu_{n_{|i|}}}, & \text{if}\quad i=j\quad\text{and}\quad n_{|i|}=n_{|j|}\neq 0;\\ \dfrac{3g_0}{8\pi\mu_{n_{|i|}}\mu_{n_{|i|}}}, & \text{if}\quad i=j\quad\text{and}\quad n_{|i|}=n_{|j|}=0.\end{cases}$$

Furthermore, setting $\Psi_\infty=\Psi_N\oplus\mathbb{1}_{\mathcal{R}_s^\infty\times\mathcal{R}_s^\infty}$, one has $H_\infty{}^\circ\Psi_\infty=\Lambda_\infty+\varepsilon K_\infty$ with

$$|K_\infty|=\mathcal{O}(\sum_{l=0}^{3}|z|^l\,\|x'\|_s^{4-l})+\mathcal{O}(\sum_{l=0}^{4}|z|^l\,\|x'\|_s^{5-l}),\tag{3.9}$$

where $\mathbb{1}$ denotes the identity map.

Before the proof of the above theorem, we first prove the following lemmas.

Lemma 3.1 There is a set $\underline{\Omega}\subset[\varrho,2\varrho]^m$ $(\varrho>0)$ such that for any $\omega\in\underline{\Omega}$ satisfying that

$$|<k,\omega>|\geq\frac{\varrho\varepsilon}{|k|^{m+1}},\quad for\quad all\quad 0\neq k\in\mathbb{Z}^m\tag{3.10}$$

And $meas\underline{\Omega}\geq(1-C_1\varepsilon)\varrho^m$, where the constant C_1 depends on m.

Proof Let $0\neq k\in\mathbb{Z}^m$, $\mathcal{R}_k^1=\left\{\omega\in[\varrho,2\varrho]^m:|<k,\omega>|<\dfrac{\varrho\varepsilon}{|k|^{m+1}}\right\}$, and $\mathcal{R}^1=\bigcup_{0\neq k\in\mathbb{Z}^m}\mathcal{R}_k^1$. Consider two hyperplanes $<k,\omega>=\pm\dfrac{\varrho\varepsilon}{|k|^{m+1}}$. We have

$$\text{meas}\mathcal{R}_k^1\leq m|k|^{-1}(\sqrt{2}\varrho)^{m-1}\frac{2\varrho\varepsilon}{|k|^{m+1}}\leq\frac{2(\sqrt{2})^{m-1}m\varepsilon}{|k|^{m+2}}\varrho^m.$$

It follows that

$$\text{meas}\mathcal{R}^1\leq\sum_{0\neq k\in\mathbb{Z}^m}\text{meas}\mathcal{R}_k^1\leq 2(\sqrt{2})^{m-1}m\varepsilon\varrho^m\sum_{0\neq k\in\mathbb{Z}^m}\frac{1}{|k|^{m+2}}$$
$$\leq C_1\varepsilon\varrho^m\sum_{p=1}^{\infty}(2p+1)^{m-1}p^{-(m+2)}\leq C_1\varepsilon\varrho^m,$$

because the series $\sum_{p=1}^{\infty}(2p+1)^{m-1}p^{-(m+2)}$ is convergent. Therefore, this lemma is true when we assume that $\underline{\Omega}=[\varrho,2\varrho]^m\setminus\mathcal{R}^1$.

Now, we use the notation $\mu_{i'}=\text{sgn}i\cdot\mu_{n_{|i|}}$.

Lemma 3.2 Assume that $n_{|i|},n_{|j|},n_{|d|},n_{|l|}\in\mathcal{I}_N$ are integers, $i,j,d,l\in\{1,-1,2,-2,...,N,-N\}$ and $1\leq|k|\leq K_0$, where we choose $K_0>K_0':=\dfrac{4}{\sigma_1}ln(\epsilon_0^{-1})$.

Then, for the parameter set $[\varrho,2\varrho]^m$, there is a subset $\bar{\Omega}\subset[\varrho,2\varrho]^m$ with

$$meas\bar{\Omega}\geq\varrho^m\left(1-\frac{C_2\varepsilon}{\ln(\epsilon_0^{-1})}\right),\tag{3.11}$$

satisfying that, for any $\omega\in\bar{\Omega}$,

$$\left|\mu_i'+\mu_j'+\mu_d'+\mu_l'+<k,\omega>\right|\geq\frac{\varrho\varepsilon}{K_0^{m+1}},\tag{3.12}$$

where C_2 is a constant depending on N, m, and σ_1.

Proof Assume $\mathcal{R}_{ijdl,k}^2=\{\omega\in[\varrho,2\varrho]^m:|\mu_i'+\mu_j'+\mu_d'+\mu_l'+<k,\omega>|<\dfrac{\varrho\varepsilon}{K_0^{m+1}}\}$ and $\Omega^2=\bigcup_{1\leq|k|\leq K_0}\bigcup_{i,j,d,l}\mathcal{R}_{ijdl,k}^2$. It follows that, by using of the same method in the proof of Lemma 3.1, for fixed i,j,d,l and k,

$$\text{meas}\mathcal{R}_{ijdl,k}^2\leq C\frac{\varrho^m\varepsilon}{K_0^{m+1}|k|}\leq C\frac{\varrho^m\varepsilon}{K_0^{m+1}},\tag{3.13}$$

where C depends on *m.* It is well known that the number

$$\sharp\{k\in\mathbb{Z}^m:|k|=l\}\leq 2^m l^{m-1}.\tag{3.14}$$

So

$$\sharp\{k\in\mathbb{Z}^m:1\leq|k|\leq K_0\}\leq 2^m\sum_{l=1}^{K_0}l^{m-1}\leq 2^m K_0^m.\tag{3.15}$$

It yields that, from (3.13), $\text{meas}\Omega^2=\text{meas}\bigcup_{1\leq|k|\leq K_0}\bigcup_{1\leq|i|,|j|,|d|,|l|\leq N}\mathcal{R}_{ijdl,k}^2$ $\leq\dfrac{C\varrho^m\varepsilon}{K_0^{m+1}}(2N)^4 2^m K_0^m\leq\dfrac{C_2\varrho^m\varepsilon}{K_0}\leq C_2\dfrac{\varrho^m\varepsilon}{\ln(\epsilon_0^{-1})}$, where C_2 is a constant depending on N, σ_1, and m. Finally, we only need to assume $\bar{\Omega}=[\varrho,2\varrho]^m/\Omega^2$. This completes the proof.

Now we prove Theorem 3.1.

Proof of Theorem 3.1 We always suppose, that $n_{|i|}, n_{|j|}, n_{|d|}, n_{|l|} \in \mathcal{I}_N$. It is convenient to adopt the notation $z_j = w_j$, $\overline{z}_j = w_{-j}$, $j = 1, 2, ..., N$, in which H_N reads, from (3.6) and (3.7),

$$H_N = \Lambda_N + \varepsilon G_N$$

$$=< \omega, J > + \sum_{j=1}^{N} \mu_{n_j} w_j w_{-j} + \frac{\varepsilon}{16} \sum_{i,j,d,l,\ n_{|i|} \pm n_{|j|} \pm n_{|d|} \pm n_{|l|} = 0} {}' g_{ijdl}$$

$$\cdot w_i w_j w_d w_l + \frac{\varepsilon}{16} \sum_{|k| \geq 1} \sum_{i,j,d,l} {}' g_{k,ijdl} e^{\mathrm{i} < k, \vartheta >} w_i w_j w_d w_l + \varepsilon \mathcal{O}(| w |^5),$$

where

$$g_{ijdl} = \frac{G_{n_{|i|} n_{|j|} n_{|d|} n_{|l|}}}{\sqrt{\mu_{n_{|i|}} \mu_{n_{|j|}} \mu_{n_{|d|}} \mu_{n_{|l|}}}},$$
$$g_{k,ijdl} = \frac{G_{k, n_{|i|} n_{|j|} n_{|d|} n_{|l|}}}{\sqrt{\mu_{n_{|i|}} \mu_{n_{|j|}} \mu_{n_{|d|}} \mu_{n_{|l|}}}}, \tag{3.16}$$

and the prime symbol in the summation sign indicates that the sum runs over all indices $i, j, d, l \in \{1, -1, ..., N, \ -N\}$.

Consider a Hamiltonian function $\mathcal{F} = \varepsilon F = \varepsilon \sum_{i,j,d,l} {}' F_{ijdl} w_i w_j w_d w_l + \varepsilon \sum_{0 < |k| \leq K_0} \sum_{i,j,d,l} {}' F_{k,ijdl} e^{\mathrm{i} < k, \vartheta >} w_i w_j w_d w_l$, where we define, for $k = 0$, $\mathrm{i} F_{ijdl} = \dfrac{g_{ijdl}}{16(\mu_i' + \mu_j' + \mu_d' + \mu_l')}$, if $\mu_i', \mu_j', \mu_d', \mu_l' \not\equiv \{a, -a, b, -b\}$ and $n_{|i|} \pm n_{|j|} \pm n_{|d|} \pm n_{|l|} = 0$ and $\mathrm{i} F_{ijdl} = 0,$ if $\mu_i', \mu_j', \mu_d', \mu_l' \equiv \{a, -a, b, -b\}$ and $n_{|i|} \pm n_{|j|} \pm n_{|d|} \pm n_{|l|} = 0$, or $n_{|i|} \pm n_{|j|} \pm n_{|d|} \pm n_{|l|} \neq 0$; for $k \neq 0,$

$$\mathrm{i} F_{k,ijdl} = \begin{cases} \dfrac{g_{k,ijdl}}{16 < k, \omega >}, & \text{if} \quad \mu_i' + \mu_j' + \mu_d' + \mu_l' = 0, \\ \dfrac{g_{k,ijdl}}{16(\mu_i' + \mu_j' + \mu_d' + \mu_l' + < k, \omega >)}, & \textit{otherwise.} \end{cases}$$

In the same way with [3] and [7], we can prove that, for the integers $n_{|i|}, n_{|j|}, n_{|d|}, n_{|l|} \in \mathcal{I}_N$ satisfying $n_{|i|} \pm n_{|j|} \pm n_{|d|} \pm n_{|l|} = 0,$ the following inequality

$$| \mu_i' + \mu_j' + \mu_d' + \mu_l' | \geq \frac{c\mu}{(M^2 + \mu)^{3/2}} > 0, \tag{3.17}$$

holds, where c is some absolute constant and $M = \min \{n_{|i|}, n_{|j|}, n_{|d|}, n_{|l|}\}$.

Let $\Psi_N = X_{\mathcal{F}}^1$ be the time-1 map of the vector-field of the Hamiltonian $\mathcal{F}$. Expanding at $t = 0$ and using Taylor's formula we can obtain that

$$H_N \circ \Psi_N = H_N + \{H_N, \mathcal{F}\} + \int_0^1 (1 - t)\{\{H_N, \mathcal{F}\}, \mathcal{F}\} \circ X_{\mathcal{F}}^t dt$$
$$= \Lambda_N + \varepsilon \tilde{G}_N + \varepsilon\{\Lambda_N, F\} + \varepsilon \hat{G}_N + \varepsilon^2 \{G_N, F\}$$
$$+ \varepsilon^2 \int_0^1 (1 - t)\{\{H_N, F\}, F\} \circ X_{\mathcal{F}}^t dt,$$

$$\tilde{G}_N(w, \vartheta) = G_N(w, \vartheta) - \hat{G}_N(w, \vartheta)$$

where $= \dfrac{1}{16} \sum_{i,j,d,l,\ n_{|i|} \pm n_{|j|} \pm n_{|d|} \pm n_{|l|} = 0} {}' g_{ijdl} w_i w_j w_d w_l$

$$+ \frac{1}{16} \sum_{0 < |k| \leq K_0} \sum_{i,j,d,l} {}' g_{k,ijdl} e^{\mathrm{i} < k, \vartheta >} w_i w_j w_d w_l + \mathcal{O}(| w |^5),$$

and $\{\cdot, \cdot\}$ is the Poisson bracket of smooth functions:

$$\{G_1, G_2\} = \frac{\partial G_1}{\partial \vartheta} \frac{\partial G_2}{\partial J} - \frac{\partial G_1}{\partial J} \frac{\partial G_2}{\partial \vartheta} + \mathrm{i} \sum_{j=1}^{N} (\frac{\partial G_1}{\partial z_j} \frac{\partial G_2}{\partial \overline{z}_j} - \frac{\partial G_1}{\partial \overline{z}_j} \frac{\partial G_2}{\partial z_j}).$$

Now let us compute $\{\Lambda_N, F\}$:

$$\{\Lambda_N, F\} =$$
$$-\mathrm{i} \sum_{i,j,d,l,\ n_{|i|} \pm n_{|j|} \pm n_{|d|} \pm n_{|l|} = 0} {}'(\mu_i' + \mu_j' + \mu_d' + \mu_l') F_{ijdl} w_i w_j w_d w_l$$
$$-\mathrm{i} \sum_{0 < |k| \leq K_0} \sum_{i,j,d,l} {}'(\mu_i' + \mu_j' + \mu_d' + \mu_l' + < k, \omega >)$$
$$\cdot F_{k,ijdl} e^{\mathrm{i} < k, \vartheta >} w_i w_j w_d w_l$$

Hence

$$\tilde{G}_N + \{\Lambda_N, F\}$$
$$= \sum_{i,j,d,l,\ n_{|i|} \pm n_{|j|} \pm n_{|d|} \pm n_{|l|} = 0} {}'(\frac{1}{16} g_{ijdl} - \mathrm{i}(\mu_i ' + \mu_j ' + \mu_d ' + \mu_l ') F_{ijdl})$$
$$\cdot w_i w_j w_d w_l$$
$$+ \sum_{i,j,d,l} {}' \sum_{0 < |k| \leq K_0} (\frac{1}{16} g_{k,ijdl} - \mathrm{i}(\mu_i ' + \mu_j ' + \mu_d ' + \mu_l ' + < k, \omega >)$$
$$F_{k,ijdl}) \cdot e^{\mathrm{i} < \mathrm{k}, \vartheta >} w_i w_j w_d w_l + \mathcal{O}(| w |^5)$$
$$= \overline{G}_N + \mathcal{O}(| w |^5) = \frac{1}{2} \sum_{i,j=1}^{N} \overline{g}_{ij} w_i w_{-i} w_j w_{-j} + \mathcal{O}(| w |^5),$$

where if $i \neq j$, $\overline{g}_{ij} = \dfrac{24}{16} g_{i-ij-j} = \dfrac{3g_0}{4\pi\mu_{n_{|i|}} \mu_{n_{|j|}}}$; if $i = j$ and $n_{|i|} = n_{|j|} \neq 0$, $\overline{g}_{ij} = \dfrac{12}{16} g_{i-ii-i} = \dfrac{9g_0}{16\pi\mu_{n_{|i|}} \mu_{n_{|i|}}}$; if $i = j$ and $n_{|i|} = n_{|j|} = 0$, $\overline{g}_{ij} = \dfrac{12}{16} g_{i-ii-i} = \dfrac{3g_0}{8\pi\mu_{n_{|i|}} \mu_{n_{|i|}}}$.

The uniqueness can be proved in the classical way as same as in [11]. Hence, we have

$$H_N \circ \Psi_N = \Lambda_N + \varepsilon \overline{G}_N + \varepsilon \hat{G}_N + \varepsilon^2 \{G_N, F\}$$
$$+ \varepsilon^2 \int_0^1 (1 - t)\{\{H_N, F\}, F\} \circ X_{\mathcal{F}}^t + \varepsilon \mathcal{O}(| w |^5).$$

CLAIM. The vector-field of the Hamiltonian X_F is real analytic in a complex neighbourhood $\vartheta \in D(\frac{\sigma_1}{2})$ of

$\mathbb{T}^m$ and some neighbourhood of the origin in $\mathbb{C}^N$, and satisfies $|F_w| = \mathcal{O}(|w|^3)$. In fact, letting $w = (w_1, w_2, ..., \ w_N) \in \mathbb{C}^N$ $\vartheta \in D(\sigma_1)$, we have that $|w|^2 = |w_1|^2 + ... + \ |w_N|^2$, and from (3.17), (3.16) and (3.2), we have that

$$|F_{\imath jdl}| = \left| \frac{g_{\imath jdl}}{16(\mu'_\imath + \mu'_j + \mu'_d + \mu'_l)} \right| \leq C |g_{\imath jdl}|$$

$$\leq \frac{C}{\mu} |G_{n_{|\imath|} n_{|j|} n_{|d|} n_{|l|}}| \leq C |g_0| \leq C.$$

Now we let $\Omega = \bar{\Omega} \cap \underline{\Omega}$. By using of Lemmas 3.1 and 3.2, it is obvious, that $meas\Omega \geq \varrho^m (1 - C_1\varepsilon - \frac{C_2\varepsilon}{ln(\epsilon_0^{-1})})$. So $meas\Omega > 0$ when ε is small enough. In addition, for $(\vartheta, x) \in D(\sigma_1) \times D(\sigma)$, by Lemma A.1 in [10] and from (3.1):

$$|g_k(x)| \leq \|g(\vartheta, x)\|_{D(\sigma_1)\times D(\sigma)} e^{-|k|\sigma_1} \tag{3.18}$$

is always true. Therefore, when $\omega \in \Omega$, if $\mu'_\imath + \mu'_j + \mu'_d + \mu'_l = 0$, from (3.10), (3.2) and (3.16), we can get

$$|F_{k,\imath jdl}| = \left| \frac{g_{k,\imath jdl}}{16 < k, \omega >} \right| \leq C \frac{|k|^{m+1}}{\varrho\varepsilon} |g_{k,\imath jdl}|$$

$$\leq C \frac{|k|^{m+1}}{\mu\varrho\varepsilon} |G_{k, n_{|\imath|} n_{|j|} n_{|d|} n_{|l|}}| \leq C \frac{|k|^{m+1}}{\mu\varrho\varepsilon} \|g(\vartheta, x)\|_{D(\sigma_1)\times D(\sigma)} e^{-|k|\sigma_1}$$

$$\leq C |k|^{m+1} \|g(\vartheta, x)\|_{D(\sigma_1)\times D(\sigma)} e^{-|k|\sigma_1}.$$

If $\mu'_\imath + \mu'_j + \mu'_d + \mu'_l \neq 0$, we have that, from (3.12), (3.2) and (3.16),

$$|F_{k,\imath jdl}| = \left| \frac{g_{k,\imath jdl}}{16(\mu'_\imath + \mu'_j + \mu'_d + \mu'_l + < k, \omega >)} \right| \leq C \frac{K_0^{m+1}}{\varrho\varepsilon} |g_{k,\imath jdl}|$$

$$\leq C \frac{K_0^{m+1}}{\mu\varrho\varepsilon} |G_{k, n_{|\imath|} n_{|j|} n_{|d|} n_{|l|}}| \leq C \frac{K_0^{m+1}}{\mu\varrho\varepsilon} \|g(\vartheta, x)\|_{D(\sigma_1)\times D(\sigma)} e^{-|k|\sigma_1}$$

$$\leq CK_0^{m+1} \|g(\vartheta, x)\|_{D(\sigma_1)\times D(\sigma)} e^{-|k|\sigma_1}.$$

It follows that, by using of (3.15),

$$|F_{w_\imath}| \leq 4 \sum_{j,d,l,n_{|\imath|} = n_{|j|} \pm n_{|d|} \pm n_{|l|}} {}' |F_{\imath jdl}| |w_j w_d w_l| +$$

$$4 \sum_{0<|k|\leq K_0} \sum_{j,d,l} {}' |F_{k,\imath jdl} e^{\mathrm{i}<\mathrm{k},\vartheta>}| |w_j w_d w_l|$$

$$\leq C \sum_{j,d,l} {}' |w_j w_d w_l| + \sum_{j,d,l} {}' \sum_{0<|k|\leq K_0} CK_0^{m+1} \|g(\vartheta, x)\|_{D(\sigma_1)\times D(\sigma)}$$

$$\cdot e^{-|k|\sigma_1} e^{|k|\sigma_1} |w_j w_d w_l| \leq C |w|^3 + C \|g(\vartheta, x)\|_{D(\sigma_1)\times D(\sigma)}$$

$$\cdot \sum_{1\leq|k|\leq K_0} K_0^{m+1} |w|^3 \leq C |w|^3 + CK_0^{m+1} 2^m K_0^m |w|^3 \leq C |w|^3,$$

where C depends on $m, g, N, \sigma_1, \sigma, K_0, \varrho, \varepsilon$ and μ. Therefore, we can get that

$$|F_w| = \sqrt{\sum_{\imath=1}^{N} |F_{w_\imath}|^2} \leq C(|w|)^3, \tag{3.19}$$

where C depends on $m, g, N, \sigma_1, K_0, \varrho, \varepsilon, N$, and μ. Similarly, we can prove that, for $(\vartheta, x) \in D(\frac{\sigma_1}{2}) \times D(\sigma)$,

$$|(G_N)_{w_\imath}| \leq \frac{1}{4} \sum{}'_{j,d,l,n_{|\imath|} \pm n_{|j|} \pm n_{|d|} \pm n_{|l|} = 0} |g_{\imath jdl}| |w_j w_d w_l|$$

$$+ \frac{1}{4} \sum_{|k|\geq 1} \sum{}'_{j,d,l} |g_{k,\imath jdl} e^{\mathrm{i}<k,\vartheta>}| |w_j w_d w_l| + C(|w|^4)$$

$$\leq C \sum{}'_{j,d,l,n_{|\imath|} \pm n_{|j|} \pm n_{|d|} \pm n_{|l|} = 0} |w_j w_d w_l| + C \sum_{|k|\geq 1} \sum{}'_{j,d,l} |g_{k,\imath jdl}|$$

$$e^{|k|\frac{\sigma_1}{2}} |w_j w_d w_l| + C(|w|^4) \leq C |w|^3$$

$$+ C \sum_{|k|\geq 1} \sum{}'_{j,d,l} |G_{k, n_{|\imath|} n_{|j|} n_{|d|} n_{|l|}}| e^{|k|\frac{\sigma_1}{2}} |w_j w_d w_l| + C(|w|^4).$$

It follows from (3.18) and (3.14) that, for $|w| \leq 1$,

$$|(G_N)_{w_\imath}| \leq C |w|^3 + C \sum{}'_{j,d,l} \sum_{|k|\geq 1} \|g(\vartheta, x)\|_{D(\sigma_1)\times D(\sigma)}$$

$$\cdot e^{-|k|\sigma_1} e^{|k|\frac{\sigma_1}{2}} |w_j w_d w_l|$$

$$\leq C |w|^3 + C |w|^3 \sum_{|k|\geq 1} e^{-|k|\frac{\sigma_1}{2}}$$

$$\leq C |w|^3 + C |w|^3 \sum_{l\geq 1} 2^m l^{m-1} e^{-l\frac{\sigma_1}{2}} \leq C |w|^3,$$

by using of convergence of series $\sum_{l\geq 1} 2^m l^{m-1} e^{-l\frac{\sigma_1}{2}}$, where C depends on $m, g, N, \sigma_1, \sigma$, and μ. Thus

$$|(G_N)_w| = \mathcal{O}(|w|^3). \tag{3.20}$$

Suppose that

$$K_N = \{G_N, F\} + \int_0^1 (1-t)\{\{H_N, F\}, F\} \circ X_{\mathcal{F}}^t dt.$$

By using of (3.20) and (3.19), we get that

$$|\{G_N, F\}| = \mathcal{O}(|w|^6). \tag{3.21}$$

Using of Cauchy estimates for the fact that $|\{\Lambda_N, F\}| = |\bar{G}_N + \mathcal{O}(|w|^5) - \tilde{G}_N| = \mathcal{O}(|w|^4)$ and from (3.19), it results

$$|\{\{\Lambda_N, F\}, F\}| = \mathcal{O}(|w|^6) \tag{3.22}$$

For $\frac{|w|}{2} \leq \frac{1}{2}$. Moreover, using Cauchy estimates for (3.21), it derives from (3.19) that $|\{\{G_N, F\}, F\}| = \mathcal{O}(|w|^8)$. It follows from (3.22) and (3.3) that $|\{\{H_N, F\}, F\}| = \mathcal{O}(|w|^6)$. Therefore, $|K_N| = \mathcal{O}(|w|^6)$ holds for $\frac{|w|}{2} \le \frac{1}{2}$.

At the end of this proof, we estimate $\hat{G}_N$. By the definition of $g_{k,ijdl}$, from (3.2), (3.14) and (3.16), for $(\vartheta, x) \in D(\frac{\sigma_1}{2}) \times D(\sigma)$ and $K_0 > K_0{}'$, we have

$$|\sum_{|k|>K_0} g_{k,ijdl} e^{\mathrm{i}<k,\vartheta>}| \le \sum_{|k|>K_0} |g_{k,ijdl}| e^{|k|\frac{\sigma_1}{2}}$$

$$\le \sum_{|k|>K_0} \frac{1}{\mu} |G_{k,n_{|i|}n_{|j|}n_{|d|}n_{|l|}}| e^{|k|\frac{\sigma_1}{2}}$$

$$\le \frac{1}{\mu} \sum_{|k|>K_0} |\int_{\mathbb{T}} g_k(x) \phi_{n_{|i|}} \phi_{n_{|j|}} \phi_{n_{|d|}} \phi_{n_{|l|}} dx| e^{|k|\frac{\sigma_1}{2}}$$

$$\le C \sum_{|k|>K_0} \left\| g(\vartheta, \mathrm{x}) \right\|_{D(\sigma_1)\times D(\sigma)} e^{-|k|\sigma_1} e^{|k|\frac{\sigma_1}{2}}$$

$$\le C \sum_{|k|>K_0} e^{-|k|\frac{\sigma_1}{2}} \le C \sum_{l>K_0} 2^m l^{m-1} e^{-l\frac{\sigma_1}{2}}$$

$$\le C K_0^m e^{-K_0\frac{\sigma_1}{2}} = C \frac{K_0^m}{e^{K_0\frac{\sigma_1}{4}}} e^{-K_0\frac{\sigma_1}{4}}$$

$$\le C e^{-K_0\frac{\sigma_1}{4}} \le C e^{-\frac{\sigma_1}{4}\frac{4}{\sigma_1}\ln(\epsilon_0^{-1})} = C\epsilon_0,$$

where C depends on g, σ_1, σ and μ, as ϵ_0 small enough. It follows that

$$|\hat{G}_N| = |\frac{1}{16} \sum\nolimits_{|k|>K_0} \sum\nolimits'_{i,j,d,l} g_{k,ijdl} e^{\mathrm{i}<k,\vartheta>} w_i w_j w_d w_l|$$

$$\le \frac{1}{16} \sum\nolimits'_{i,j,d,l} |\sum_{|k|>K_0} g_{k,ijdl} e^{\mathrm{i}<k,\vartheta>}| |w_i w_j w_d w_l|$$

$$\le C\epsilon_0 \sum\nolimits'_{i,j,d,l} |w_i w_j w_d w_l| \le C\epsilon_0 |w|^4,$$

where C depends on $g, \sigma_1, \sigma, \mu,$ and N. Hence, $\hat{G}_N = \epsilon_0 \mathcal{O}(|w|^4)$. This completes the proof.

Acknowledgments

This work is partially supported by Program for Soft Science Research Project of Shandong Province (No. 2013RKB01253); this work is also partially supported by NSFC grants (No. 61340043 and No. 61273230).

References

[1] Wayne C E 1990 *Comm. Math. Phys.* **127** 479-528
[2] Poschel J 1996 *Ann. Sc. Norm. Super. Pisa Cl. Sci. IV Ser.* **23** 119-48
[3] Poschel J 1996 *Comm. Math. Helv.* **71** 269-96
[4] Bambusi D 2003 *Comm. Math. Phys.* **234** 253-85
[5] Yuan X 2005 *J. Differential Equations* **230** 213-74
[6] Chierchia L, You J 2000 *Commun. Math. Phys.* **211** 497-525
[7] Bricmont J, Kupiainen A, Schenkel A, 2001 *Commun. Math. Phys.* **211** 101-40
[8] Berti M, Procesi L 2006 *Comm. in Partial Differential Equations* **31** 959-85
[9] Zhang M, Si J 2009 *Physica D* **238** 2185-215
[10] Poschel J 2001 *Proc. Symp. Pure Math.* **69** 707-32
[11] Siegal C L, Moser J K 1971 *Lectures on Celestial Mechanics* vol. Grundlehren **187** Springer: Berlin

Author

Yi Wang, born in May, 1980, Shandong, China

Current position: associate professor of mathematics
University studies: Ph. D from Shandong University in 2012
Scientific interest: differential equations, difference equations, and applications of mathematics in economics
Publications: Yi Wang has published many papers of high scientific quality, including “A result on quasi-periodic solutions of a nonlinear beam equation with a quasi-periodic forcing term” in *Zeitschrift fur Angewandte Mathematik und Physik*, “Quasi-periodic solutions of a quasi-periodically forced nonlinear beam equation” in *Communications in Nonlinear Science and Numerical Simulation*, etc.

Wireless sensor networks optimization covering algorithms based on genetic algorithms

Sun Zeyu[1, 2], Yang Tao[1], Shu Yunxing[1]*

[1] *Computer and Information Engineering, Luoyang Institute of Science and Technology, Luoyang 471023, China*

[2] *Electrical and information Engineering, Xi'an JiaoTong University Xi'an 710061, China*

Received 1 February 2014, www.tsi.lv

Abstract

This paper starts with two methods applied widely of computational intelligence; Evolutionary computing and swarm intelligence. It makes the Genetic Algorithms (GA) that is classic in evolutionary computing and genetic algorithm that is representative in swarm intelligence as its study foundation. It presents theory and characteristic of the two methods to seek the application of intelligent optimization in engineering practice. In application, in view of the feature that wireless sensor network (WSN) must possess auto-organization, auto-adaptation and robustness, especially, energy of WSN is very limited, this paper fully utilizes the advantages of computational intelligence, marries together both the research focuses. It proposes some methods and ideas for applying computational intelligence to solve optimization problems of WSN. This paper depicts coverage problem of WSN, for the feature that this problem is the problem of multi objective optimization, under the topology control of GA, it applies GA based on sorting to solve the problem, then improves this algorithm to maintain population diversity and obtain high-quality, well distributed solutions. The algorithm it proposes realizes the aim that using the least number of sensor nodes to achieve the best coverage, which is able to save energy of the network, decrease the interference between signals and prolong the network life-time.

Keywords: Wireless Sensor Network (WSN), Coverage rate, Sensor node, Genetic Algorithms

1 Introduction

With the progress of science and the development of the times, problems that people encounter in industrial production and engineering practice, with more and more large-scale, complex, constraint, nonlinearity, uncertainty, etc., in the production practice! Economic management and scientific research in many fields have a lot of questions that people in urgent need of a large and complex space to find optimal or near-optimal solution. Computational intelligence as a new optimization technique solves the difficulties encountered in conventional optimization algorithm, the algorithm is relatively simple and easy to understand, easy to implement, more importantly, and computational intelligence methods mostly have implicit parallelism, Self-organization, adaptive characteristics, effectively promote their application in optimizing the production of various areas of the system efficiency, reducing energy consumption, rational use of resources and improve the economic efficiency has an important role and significance.

WSN (Wireless Sensor Network) is a kind of self-organization network system which consists of large number of inexpensive sensor nodes, and its nodes are characterized by a certain sensing ability, computing power and communication capabilities. It is widely used in the fields of defence and military, environmental monitoring, rescue works and etc. WSN works in such a way that following way: large numbers of sensor nodes are distributed in discrete form within the coverage area, and data is sent to or collected from nodes directly or indirectly. Usually the target node is covered in a manner that sensor nodes are high density deployed to monitor the target area, and to improve the quos of network, information is exchanged among sensor nodes to achieve target node coverage and information processing. But there're some defects, first, deployment of larger number of sensor nodes in target area results in existence of considerable amount of redundant nodes, which consume much network energy and reduce the network Qos.

The second, due to the excessive consumption of node energy, and non-rechargeable feature of nodes, the network tends to collapse quickly. How to distribute sensor nodes in target area reasonably to determine the minimum point set under certain coverage requirement, and how to limit the power consumption maximally, become key problems, which influence the network lifetime directly. In summary, the solving of energy issues and coverage problem means monitoring the given area at the minimum nodes number and low energy cost, meanwhile, the quality of coverage should be guaranteed. It is also the study focus of this paper.

*~~~~***Corresponding author** e-mail: lylgszy@163.com

2 Related works

In the 1990 of the 20th century, Italy scholar M Dorig, who was inspired from the mechanism of biological evolution, Ant routing behaviour by simulating the natural world, proposing a new simulated evolution of Ant Colony algorithm (Ant Colony algorithm ACA). Early was widely used in the travelling salesman problem (Travelling Salesman Problem, TSP) solution. Travelling salesman problem is a typical combinatorial optimization problem, but also a NP hard problem. As the problem grows, ant colony algorithm in a limited number of cycles is difficult to find the exact solution of the problem, and can easily fall into local optimal solution, causing the system to run the cycle is too long, slow convergence and the emergence of stagnation. University of Michigan in 1975, Professor John H. Holland proposed genetic algorithm (Genetic Algorithm, GA) can be initialized from a start node traversal, to avoid initialization from a single node caused the most easy to fall into local optimal solution of the iterative process that converges to a greater probability of the optimal solution, which has a better ability to solve the global optimal solution. However, in solving complex nonlinear problems there too premature, convergence is slow, resulting in a lot of redundant code and other shortcomings, thus making the solution accuracy is too low [1-3]. Wireless sensor network coverage problem in the field of wireless sensor network (WSN) is one of the focuses of research problems. Wireless sensor network characteristics can be summarized as: small size, low cost, low energy consumption, has a certain calculation, processing and communication capabilities. In the process of wireless sensor network coverage, needs to solve two problems: first, the coverage, how to reasonably and effectively reduce the node energy consumption, and try our best to prolong the network life cycle; Second: using mobile node scheduling strategy and parameter dynamic change, reduce the mobile node to cover the amount of work area, to achieve the goal of local area cover effectively, enhanced the topology of the network, reducing redundant data generated at the same time improve the quality of network service. Covered, therefore, how to meet certain conditions, the use of minimum sensor node to specify the local area covered and effectively inhibit the node energy consumption of too fast is a challenging topic.

The methods of deployment of the wireless sensor network nodes can be divided into deterministic deployment and randomness deployment. Usually the deterministic deployment method is adopted when the network is small, and a good monitoring regional condition can be guaranteed. The advantage of this method is that by controlling the position of each node through artificially deployment, the optimal solution meeting the network coverage requirement can be achieved. On the contrary, if artificial way is not feasible, usually aircraft or other tools is used to randomly distribute sensor nodes in a certain area, because of the uncertainty of nodes' positions, more nodes will be needed compared to the deterministic method, then the node redundancy problem comes out. Thus, the problem of energy consumption and node coverage become one of the major research topics in the fields of wireless sensor network. In references [4], by exploiting the Force Filed Theory of mobile network and Round Coverage Thinking in wires sensor network, VFA algorithm is proposed. When the nodes in a wireless sensor network are distributed unevenly, this algorithm can be used to scatter the intensive nodes in order to effectively cover the target area, but the energy consumption problem of whole network is not fully considered. In references [5], a sensor network coverage and connectivity probability model in the case that nodes are random scattered is proposed. By exploiting this model, the node numbers which meet different coverage and connectivity requirement can be calculated and the calculation is simple; but this model is only studied with the complete coverage case, meanwhile, the connectivity rate under multi-network coverage is not considered. In references [6] proposed perception coverage and connectivity restore study in mobile sensor network, the idea is study the coverage area and connectivity issues as a whole, Coverage Conscious connectivity Restoration is used to restore one or more nodes from the failure nodes, thus the connectivity is restored and this node at initial position in coverage area are monitored. Because the energy consumed for data collection every time is not equal, and it is not suitable to re-divide the intersecting coverage set during the recovering process of the failure nodes [7-9]. To this end, by using the Gaussian density function and the coverage area probability function the quantitative comparison between the node-sensing radius and the number of nodes is given in the minimum nodes set theory model, so that the coverage for the target area is done.

3 Mathematical models of genetic algorithms

In practical applications, it is often encountered in multi-criteria or objectives, design and decision making problems, "such as securities investment issues, investors in order to get higher returns, you need to select the best stocks to invest in, in general, an outstanding shares have the following characteristics: good performance, low price-earnings ratio, growth higher, but usually these goals are in conflict, such as the current domestic steel industry generally better performance of listed companies, earnings are relatively low, but the steel industry is not sunrise industry, the company's growth is not high; while some small and medium sized companies although growth is high, but the performance is poor, the high price-earnings ratio, and thus to be able to choose a good stock, you need to make investment decisions among these goals a balanced approach, that more than a numerical target in a given region of the optimization problem is known as multi-objective optimization.

In order to solve multi-objective optimization problem, we need to create a general mathematical model, we must first determine its decision variables, the general case, the decision variables dimensional Euclidean space as a point, namely:

$$x=\left(x_1,x_2,x_3\cdots x_n\right)\in E^n. \tag{1}$$

The second one is the objective function, in general it can be assumed with objective functions and decision variables are all about function, namely:

$$f\left(x\right)=\left[f_1\left(x\right),f_2\left(x\right),\cdots f_p\left(x\right)\right]^T. \tag{2}$$

Finally, its constraints, from a mathematical point of view, there are two constraints: inequality constraints and equality constraints, constraints can be defined as the m inequality constraints and k equality constraints:

$$\begin{cases} g_i\left(x\right)\le 0 & i=1,2,3\cdots m \\ h_j\left(x\right)=0 & j=1,2,3\cdots k \end{cases}. \tag{3}$$

If all are the minimization of the objective function value, the multi-objective optimization problem can be described as the following mathematical model:

$$\begin{cases} min\ f\left(x\right)=\left[f_1\left(x\right),f_2\left(x\right),\cdots f_p\left(x\right)\right]^T \\ x_i^\alpha \le x_i \le x_i^\beta \end{cases}, \tag{4}$$

where, x is the decision variable, $f\left(x\right)$ is the objective function, X represents the decision vector formed by the decision space x, $g_i\left(x\right)$ and $h_j\left(x\right)$ constraints x feasible decision variables to determine the range, min represents A Minimization Vector, namely, a vector target $f\left(x\right)=\left[f_1(x),f_2(x),\cdots f_p(x)\right]^T$ in certain constraints as far as possible the various sub-objective function minimization. It can be seen when the $p=1$, the mathematical model for a single objective optimization problem mathematical model.

3.1 DEFINITION MULTI-OBJECTIVE OPTIMIZATION

Multi-objective optimization problem is that people in the production or frequently encountered problems in life, in most cases, due to multi-objective optimization problem in all its goals are in conflict, a sub-target improvement may cause the performance of other sub-goals reduced, in order to make optimal multiple targets simultaneously is impossible, and thus in solving multi-objective optimization problem for each sub-goal can only be coordinated and compromise treatment, so that each sub-objective functions are optimal as possible multi-objective optimization problem with a single objective optimization problem is essentially different, in order to properly solve multi-objective optimization problem the optimal solution, we must first multi-objective optimization of the basic concepts of a systematic exposition.

Definition 1: N Viola Space:

$$\begin{cases} x=\left(x_1,x_2,x_3\cdots x_n\right)^T \\ y=\left(y_1,y_2,y_3\cdots y_n\right)^T \\ x=y \quad Iff\ x_i=y_i \quad \forall i=1,2,3\cdots n \\ x>y \quad Iff\ x_i>y_i \quad \forall i=1,2,3\cdots n \end{cases}. \tag{5}$$

Definition 2: Let $X\subseteq R^m$ is a multi-objective optimization model of the constraint set, $f\left(x\right)\in R^p$ is a vector objective function, $x_1\in X$, $x_2\in X$, (a) $f_k\left(x_1\right)<f_k\left(x_2\right)$ better solution called solution, x_2. (b) x_1 weak solution of $f_k\left(x_1\right)\le f_k\left(x_2\right)$ called superior solution x_2. (c) $f_k\left(x_1\right)\ge f_k\left(x_2\right)$ solution called indifference to solution x_1, x_2.

Definition 3: Let $X\subseteq R^m$ be a multi-objective optimization model constraint set, $f\left(x\right)\in R^p$ is a vector objective function, $x^n\in X$ and x^n than the X all the other points are superior, called x^n is the multi-objective minimization model optimal solution. By definition, multi-objective optimization problem is to make the optimal solution x-vector objective function $f\left(x\right)$ for each sub-goal is to achieve the most advantages of the solution, obviously, in most cases; the optimal multi-objective optimization problem solution does not exist.

Definition 4: Pareto optimal solution: Let $X\subseteq R^m$ be a multi-objective optimization model constraint set, $f\left(x\right)\in R^p$ is the vector of the objective function. If $\xi\in X$, ξ and there is no more than the superiority of x, then ξ is a minimal model of multi-objective Pareto optimal solution, or non-inferior solution.

Definition 5: No inferior set with the front end: Let $X\subseteq R^m$ be a multi-objective optimization model constraint set, $f\left(x\right)\in R^p$ is a vector objective function. $\lambda\in X$ Is a minimal model of multi-objective Pareto optimal solution set, then λ is called non-inferior set of X, $Y=f\left(\lambda\right)$ is called Pareto optimal front.

Seen from the above definition: (a) Multi-objective optimization problem with a single objective optimization problem is essentially different, in general, multi-objective optimization problem Pareto "optimal solution is a collection of the Mu most cases, similar to the single-objective optimization problem in a multi-objective optimal solution optimization problem does not exist, there is only Pareto optimal "multi-objective optimization problem is just a Pareto optimal solution acceptable" not bad "solution, and usually most multi-objective optimization problem with multiple Pareto optimal solution. (b) If a multi-objective optimization problem optimal solution exists, then the optimal solution must be Pareto optimal solution, and the Pareto optimal solution is also the optimal solution by only composed of these, do not contain other solutions, so can be so say, Pareto optimal solution is a multi-objective optimization problem reasonable solution set. (c) For practical application, must be based on the level of understanding of the problem and the decision-makers of personal preference, from a multi-objective optimization problem Pareto optimal solution set of one or more selected solution as a multi-objective optimization problem of optimal solution, so seeking more objective optimization problem the first step is to find all its Pareto optimal.

3.2 NETWORK MODEL AND HYPOTHESIS

The following hypotheses are advanced on the network model:

Hypothesis 1: the monitored area is much larger than the sensor node sensing area, not considering the boundary factors on the monitoring of regional influence.

Hypothesis 2: sensor node sensing radius and radius of communication will appear a disk shape and the communication radius greater than or equal to 2 times the radius of perception.

Hypothesis 3: each sensor node can be through their own information to their location information.

Hypothesis 4: the initial state, and each sensor of node energy is the same, all sensor nodes have the same processing capacity, and equal status.

Definition 6*:* the distance between any two nodes $d(i,j)$ are called nodes i and j Euclidean distance, when $d(i,j)<2R$ referred to the neighbour node, node i and j.

Definition 7: in the monitoring of the target area, when a target node is K sensor node coverage, called K heavy cover.

Definition 8: in the monitoring of the target area, all sensor nodes coverage Union and all sensor nodes range and then, called network covering efficiency:

$$EA = \bigcup_{N=1,2\ldots N} S_i \Big/ \sum_{N=1,2\ldots N} S_i \ . \tag{6}$$

Definition 9: Covering the region of coverage for:

$$p(s_i,s_j)=\begin{cases} 0 & \text{if} \quad R_s \le d(s_i,s_j) \\ e^{-\varepsilon d} & \text{if} \quad (R_s-R_e)<d(s_i,s_j)<R_s \\ 1 & \text{if} \quad d(s_i,s_j)\le(R_s-R_e) \end{cases} \cdot \tag{7}$$

Among them: ε is sensor node physical parameters; R_e said sensor node monitoring dynamic parameters in the said sensor nodes; $d(s_i,s_j)$ Euclidean distance; when $d(s_i,s_j)\le(R_s-R_e)$, this time node s_i is detected, it is not detected.

Theorem 1: when and only when the three equal circles intersect at one point, and form an equilateral triangle length of a side is $\sqrt{3}$, covering the efficiency of *EA* maximum, That is: $EA\le 82.73\%$

Proof*:* As shown in Figure1:

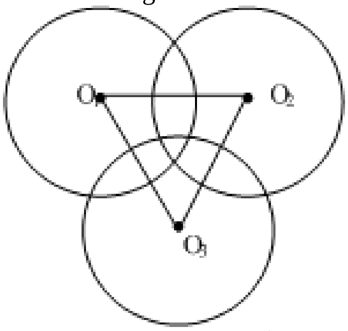
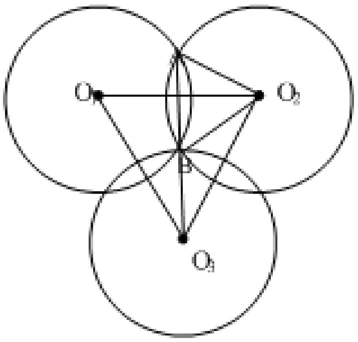


FIGURE 1(a) Any intersection Of two circles



FIGUER 1(b) Any three circles intersect at one point

Firstly, Figure 1(a) was analysed. Two intersect, and the intersection region are equal, so $\triangle O_1O_2O_3$ is an equilateral triangle, with side length O_1O_2 is r, an equilateral triangle $\triangle O_1O_2O_3$ three interior angles are respectively $\pi/3$, $\angle O_2O_1O_3=\pi/3$, $S_{\triangle O_1O_2O_3}=\frac{1}{2}(r^2\sin{}^{\pi}\!/_{3})$, since three, round two intersection, and completely covered on the plane the Euclidean distance, $d_i<2r$, let the equilateral triangle $\triangle O_1O_2O_3$ the maximum length to keep the $S_{\triangle O_1O_2O_3}$ area is the largest, the three circle intersect at a point B, as shown in Figure 1 (b) as shown, connect to the O_3B and extended to two points to A, connecting the O_2A, set three the radius of the circle of 1, $S_{\triangle ABO_2}={}^{\sqrt{3}}\!/_{4}$, $S_{\mathrm{ABO_2}}={}^{\pi}\!/_{6}$, according to the formula (1) we get $EA=S_{\triangle ABO_2}/S_{ABO_2}$, $EA = {}^{3\sqrt{3}}\!/_{2\pi}=82.73\%$ namely in the completely covered cases the maximum coverage of the efficiency value is 82.73%.

Theorem 2: a sensor node monitoring area A, the monitoring of regional node density λ, a monitoring area A node number X subject to node K probability density: $P(X=k)=e^{-\lambda A}\cdot(\lambda A)^k/k!$

Proof: the monitoring area is S, in the monitoring region of arbitrary nodes subordinated to the K node distribution probability of $p=A/S$, when the number of nodes of n probability obeys two type distributions is:

$$P(X=k)=C_n^k p^k(1-p)^{n-k}. \tag{8}$$

According to the node density formula $\lambda=n/S$ into arbitrary node distribution probability of P:

$$p=A\lambda/n. \tag{9}$$

Equation (9) into the formula (8):

$$\begin{aligned}P(X=k)&=C_n^k(A\lambda/n)^k(1-A\lambda/n)^{n-k}\\&=\frac{n!(A\lambda)^k(1-A\lambda/n)^n}{(n-k)!k!(n-A\lambda)^k}\end{aligned} \tag{10}$$

When $n\to\infty$ and its limit available:

$$\begin{aligned}P(X=k)&=\lim_{n\to\infty}\left(\frac{n!(A\lambda)^k(1-A\lambda/n)^n}{(n-k)!k!(n-A\lambda)^k}\right).\\&=e^{-\lambda A}(A\lambda)^k/k!\end{aligned} \tag{11}$$

That is: $P(X=k)=e^{-\lambda A}\cdot(\lambda A)^k/k!$

4 Coverage control and scheduling of nodes

In order to achieve the efficient coverage on monitoring region by minimal node, the purpose is to better extend network existence period. Make the network lifetime maximization is the basic method to make the network system of the node energy minimization. That is to say, in the network monitoring region to let each sensor node to consume all their energy as much as possible, but in practical application process exists the position difference, the sensor node energy consumption is not the same; for example: in close proximity to the base station node for forwarding a large amount of data and the formation of excessive energy consumption and rapid death. Therefore, the node exists between energy consumption disequilibrium phenomenons, which require the deployment of nodes, considering the different regional deployed nodes is also different. Its purpose is to balance each sensor node's energy has to balance the network deployment, while the network effectively covering algorithm finally, can be achieved on the node energy consumption effectively resist, the lower energy nodes not too quick death, thereby extending the network cycle.

When the target into a cluster head monitoring area, to the neighbour cluster head node sends a packet containing the target information, all the monitoring to the target cluster are dynamically in the target around to form a group, cluster member nodes only with the cluster node communication, the cluster head and between cluster heads can be mutually communication. Involved in tracking the cluster number depends on the size of the radius of the grid. For example, if the access grid side length equal to the radius of communication node, then a maximum of only four cluster capable of simultaneously monitoring to the target. When at the same time two or more than two cluster head and monitoring to the target, we select these clusters in a cluster head node as a leader node, cluster head first to the neighbour hair to send their and monitoring the distance between the target data information, if the cluster head received a distance closer to the target hair to information, give up campaign to become leader node. Selection criteria for: first, choose from the closest cluster head node; second, if there is two or more than two cluster head node and the target and the distance between the same, residual energy larger the lead node. All the monitoring to the target cluster head node will be sent to a leader node data first, and then by the leading node calculation and data fusion are transmitted to a data centre node. As shown in figure 2:

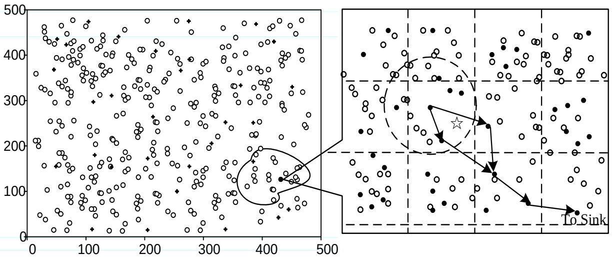


FIGURE 2 The target node coverage area diagram

When the mobile target leading away from the node, because of the need to transmit data over long distances to the leader node, or a new cluster head node monitoring to the target, then a leader node is no longer applicable acts as a leader node, fast the election of a new leader node is very necessary. Here we shall, when there is a new cluster head node joins the mobile target tracking, under the leadership of node selection rules, in all involved in tracking the cluster head node selects a distance to a target the nearest cluster head node as its new leader node, data reported by the new leader node is sent to a data centre.

5 Simulation experiment

In order to evaluate the feature of the algorithm, this paper MATLAB6.5 is adopted as a simulation platform in this paper, the sensor nodes are randomly deployed in different network areas, the parameters are included in table1.

TABLE 1 Simulation parameters

Parameter	Value	Parameter	Value
dimension 1	100*100m^2	ε_{amp}	20(pJ/b)/m^2
dimension 2	200*200m^2	$E_{R\text{-}elec}$	30nJ/b
dimension 3	400*400m^2	E_{min}	0.02J
Number	180	Header	20B
R_s	2m	Initial energy	2J
$E_{T\text{-}elec}$	50nJ/b	broadcast	20B
ε_{fs}	10(pJ/b)/m^2	each round	100ms

The wireless communication models for Sensor node transmitting data and receiving data are respectively the following:

$$E_{Tr}(k,d)=E_{T-elec}k+E_{amp}(k,d) = \begin{cases} E_{T-elec}k+\varepsilon_{fs}d^2k & d<d_0 \\ E_{T-elec}k+\varepsilon_{amp}d^4k & d\ge d_0 \end{cases} . \qquad (12)$$

In the above formula, $E_{T\text{-}elec}$ and $E_{R\text{-}elec}$ denote the energy consumption of wireless transmitting module and wireless receiving module; ε_{fs} and ε_{amp} stand for the energy consumption parameters of spatial model and multiple attenuation models; d_0 is a constant.

Experiment I. The first case is, with the same respective parameters, execute 50 times and get the mean value, then execute for 400 to compare with the LEACH protocol the quantitative relationship between number of remaining nodes and the number of turns, as shown in Figure 3.

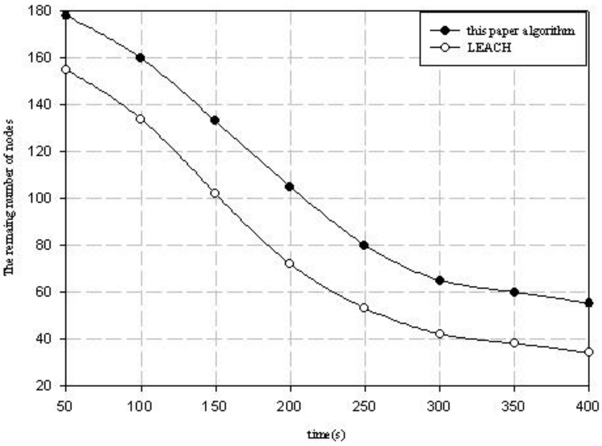


FIGURE 3 Remaining nodes and the round number

As can be seen from Figure 2, with increasing of time, the number of remaining nodes of proposed algorithm is higher than the LEACH protocol, and then the conclusion that with the increasing of time, the energy consumption of the proposed algorithm is lower than that of LEACH protocol, and the network lifetime is extended, also the network resources are optimized.

Experiment II. In order to achieve the scale of network coverage, and thus better evaluate the performance of the model in different sizes, which mainly reflect the minimum number of nodes needs to by deploy in different network coverage, each simulation experiment executed 50 times at average. Curve of node coverage changes is shown in Figure 4.

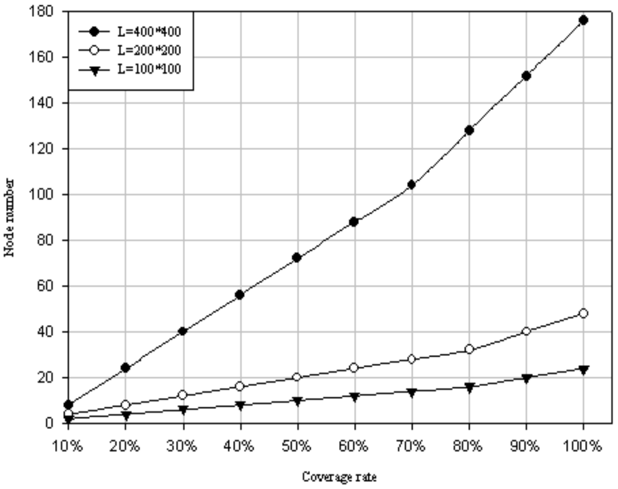


FIGURE 4 Coverage rate for different coverage area

Figure 4 shows the graph of the number of sensor nodes needed to deploy to achieve different node coverage under different network dimensions. The figure shows that, with the expansion of the network, to meet the demand for network coverage, the number of nodes required to be deployed will increase, and the higher the coverage of the network, the number of nodes need to be deployed increases can be obtained from Figure more fast, so that the concern target node can achieve complete coverage.

Experiment III. Figure 5 shows a diagram of the number of sensor nodes need to be deployed for the same network size 400 * 400m^2 under different node coverage requirement, and compare with the experiments of literature [10] SCCP algorithm, to meet certain demand for network coverage, the number of nodes deployed will be gradually increased as time progresses, and the network coverage will also increase, so that completely coverage is achieved for the same coverage area and different nodes coverage for target area, as shown in Figure 5.

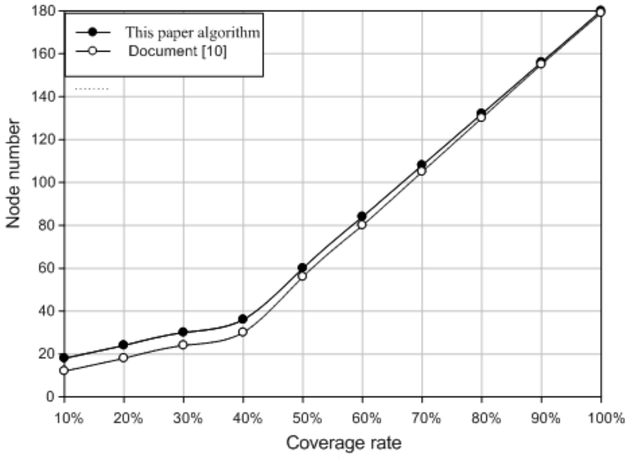


FIGURE 5 Coverage comparison of proposed algorithm and [10]

5 Conclusions

Computational Intelligence approach is that people learn and use a variety of principles and mechanisms of natural phenomena in nature or organisms developed a new method of adaptive environmental capacity and has, because of its efficiency to optimize performance, no problem specific information, etc., in has been successfully applied in many fields. coverage problem for WSN are described, for it has the characteristics of multi-objective optimization, genetic algorithm based on Pareto applied to solve this sort of problem, "in algorithm design, the choice of the operator has been improved, Meanwhile, the introduction of external groups to save Pareto optimal solutions generated by each generation, and using quick sort method based on Euclidean distance to external groups to be updated to maintain population diversity and individual differences between algorithm to achieve the ultimate purpose of using as few sensor nodes in order to achieve the greatest possible degree of target coverage, so WSN energy consumption can be balanced.

Acknowledgments

Scientific Research Program Funded by Henan Provincial Education Department Project No.2014B520099; Natural Science and Technology Research of Foundation Project of Henan Province Department of Science under Grant Nos. 142102210471.

References

[1] Cai Guo qiang, Jia Li min, Xing Zong yi 2008 Design of Ant Colony Algorithm based Fuzzy Classification System *Fuzzy Systems and Mathematics* **4**(22) 87-98

[2] Feng Zu hong, Xu Zong ben 2002 A Hybrid Ant Colony Algorithm for Solving TSP *Journal of Engineering Mathematics* **4**(19) 35-9

[3] Gao Shang, Zhang Xiao ru 2009 Ant Colony Optimization Genetic Hybrid Algorithm *Mathematics in Practice and Theory* **24**(39) 93-6

[4] Zou Y, Charrabarty K 2003 *Sensor deployment and tarter localization base on virtual forces proceeding* Elsevier Ad HocNetwork 286-97

[5] Bahi J, Makhoul A, Mostefaoui A 2008 *Computer Communications* **31** 770-81

[6] Tamboli N, Younis M 2010 Journal of Network and Computer Applications **33** 363-74

[7] Xing G, Wang X, Zhang Y 2005 *ACM Transactions on Sensor Networks* **1**(1) 36-72

[8] Chan Y, Zhao Q 2005 *IEEE Communications Letters* **11** 55-9

[9] Zhang H, Ho J 2005 ACM Trans on Sensor Networks **2** 272-9

[10] Xing X, Wang G, Wu J, Li J 2009 Square Region-Based Coverage and Connectivity Probability Model in Wireless Sensor Networks *Proc. of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing* (CollaborateCom 2009) Washington DC USA November 2009

Authors

Sun Zeyu, born in 1977, Changchun city, Jilin province

Current positions, grades: Master of Science, PhD student of Xi 'An Jiaotong university
University study: Master of Science, Lanzhou university, 2010; Xi 'An Jiaotong university study for a doctorate at present
Research interests: wireless sensor networks, parallel computing and Internet
Experience: a lecturer in Luoyang institute of technology of computer and information engineering, a member of China computer society

Yang Tao, born in 1982, Luoyang City, Henan Province

Current position, positions: a lecturer at the Luoyang Institute of Computer and Information Engineering
University study: Master of Science at Lanzhou University, 2009
Research interests: distributed computing, and network information security

Shu YunXing, born in 1962, Nantong, Jiangsu province

Current position, positions: , professor Luoyang institute of technology
University study: master's degree graduated at Tsinghua university computer college with in engineering, 1994; PhD at Wuhan university of science and engineering, 2008
Research interests: computer modelling and simulation, complex networks and parallel computing
Experience: participated in two national natural science foundation of China, completed the 12 key projects of Henan province department of science, academic leaders in Henan province, Luoyang outstanding experts.

An improved light-weight trust model in WSN

Na Wang[1, 2]*, Yanxia Pang[2]

[1] *MoE Engineering Center for Software/Hardware Co-design Technology and its Application, East China Normal University, No.3663 North Zhongshan Rd, Shanghai 200062, China*

[2] *School of Computer and information, Shanghai Second Polytechnic University, No. 2360 Jinhai Rd, Shanghai 201209, China*

Received 1 March 2014, www.tsi.lv

Abstract

WSN is often deployed in unattended or even hostile environments. Therefore, providing security in WSN is a major requirement for acceptance and deployment of WSN. Furthermore, establishing trust in a clustered environment can provide numerous advantages. We proposed a light-weight trust model which considers data aggregation and communication failure due to wireless channels. It computes retransmission rate to get success, failed and uncertain value, and details the data in parameters to depend against attacks. With comparing our model with LDTS and Model using Trust Matrix, we conclude that our model has implemented a trade-off between detection rate and communication consumption.

Keywords: direct trust, light-weight, trust matrix, retransmission rate, indirect trust

1 Introduction

A large amount of applications ranging from health, home, environmental to military and defence make use of sensor nodes for collection of appropriate data. The sensor nodes comprising of data collecting, processing, and transmitting units are very small in size and can be densely deployed owing to their low cost [4]. Cluster WSN such as LEACH is broadly used. Clustering algorithms can effectively improve network scalability and throughput. Using clustering algorithms, nodes are grouped into clusters, and within each cluster, a node with strong computing power is elected as a cluster head (CH). CHs together form a higher-level backbone network. After several recursive iterations, a clustering algorithm constructs a multi-level WSN structure [5].

However, WSN is often deployed in unattended or even hostile environments. The wireless and resource-constraint nature of a sensor network makes it an ideal medium for attackers to do any kinds of vicious things. Therefore, providing security in WSN is a major requirement for acceptance and deployment of WSN [6]. Establishing trust in a clustered environment provides numerous advantages, such as enabling a CH to detect faulty or malicious nodes within a cluster. In the case of multi-hop clustering, a trust system aids in obtain correct data aggregation.

The rest of the paper is organized as follows. The models and definitions are proposed in section 3. The detailed trust model is depicted in Section 4. The comparison and evaluation of our trust model with other models are given in Sections 5. The related work and our conclusions are presented in Sections 2 and 6.

2 Related work

Research on trust management systems for WSN received considerable attention from scholars. A number of studies have proposed such systems for WSNs. However, these systems suffer from various limitations such as the incapability to meet the resource constraint requirements of the WSNs, more specifically, for the large-scale WSN. Recently, a few trust management systems have been proposed for clustered WSNs, such as GTMS [1], Model using Trust Matrix [3], a light-weighted Trust Model [16]. To our best knowledge, a universal trust system designed for clustered WSNs to achieve light-weight remains lacking.

In Group based Trust Management Scheme [1], the authors proposed a new light weight trust management scheme for WSN. It works with two different topologies: intragroup and intergroup, where distributed trust management and centralized trust management is adopted respectively. And the trust states are represented as Trusted, Untrusted and Uncertain respectively. The advantage of the scheme is that, it evaluates the trust for the group of nodes rather than a single node in the cluster. However, GTMS relies on a broadcast-based strategy to collect feedback from the CMs of a cluster, which requires a significant amount of resources and power.

In a Fault-Event Detection Model Using Trust Matrix in WSN (DMUTM) [3], the author proposed a method of fault and event detection using trust model in WSN based on similarity matrix. They used similarity matrix which is based on data aggregation distinguish groups from each other in one cluster to detect fault. The trust was calculated by cluster head either directly or indirectly. When in indirectly case, the head calculated the trust by

**Corresponding author* e-mail:wnoffice@126.com transitivity algorithm. However, the trust transitivity required a high complexity, which leads to amount of power consumption.

In [8], Xiao proposed a trust system LDTS for WSNs, which employ clustering algorithms. First, a lightweight trust decision-making scheme is proposed based on the nodes' identities in the clustered WSNs. Then a dependability-enhanced trust evaluating approach is defined for co-operations between CHs. Moreover, a self-adaptive weighted method is defined for trust aggregation at CH level. But the method focuses on transmit process but not considers data property in the network. Therefore, it can only depend against Garnished attack and bad mouthing attack.

In [13], the author proposes a trust-based defending model against multiple attacks. Considering the characteristics of resource-constrained sensor nodes, trust values of neighbouring nodes on the routing path can be calculated through the Dirichlet distribution function, which is based on data packets' acknowledgements in a certain period instead of energy-consuming monitoring. But the data packets' acknowledgements may consume much energy.

In A light-weighted Trust Model [16], the authors proposed a trust model based on data aggregation and detailed the data in parameters to depend against attacks. But it did not consider the retransmission rate and use only data similarity to make a trust decision while omit the transmission quality.

Therefore, it is necessary to build a light-weight trust model which consider data aggregation and detailed the data in parameters to depend against more attacks. Work in this paper is an improvement of our former work [16], the contributions are:

1) Use retransmission rate to compute success value.
2) Create an improved light-weight trust model based on our former work.
3) Combine transmission and data similarity to evaluate the total trust of a node to another.
4) Compare our model with LDTS, Model using Trust Matrix and our former work.

3 Models and definitions

3.1 NETWORK MODEL

WSN in a two dimensional plane with n sensors, denoted by a set N = (n_1, n_2, ……, n_n), where n_i is the ith sensor. These sensors are placed in an area and the transmission radius is r_s. Each node maintains its ID, sensing data and location. In such a network, we use LEACH protocol to create clusters. A node in the clustered WSN model can be identified as a CH, or a CM. Members of a cluster can communicate with their CH directly. A CH can forward the aggregated data to the central BS through other CHs.

3.2 TRUST MODEL

Trust models are classified into two categories that are node trust models and data trust models [6].

A data trust model is proposed to distinguish forged data of illegal nodes from innocent data of legal nodes. Sensor nodes evaluate trustworthiness of their neighbour nodes by cross checking the neighbour nodes' redundant sensing data with their own result. The trust value is calculated through a light-weighted method, and the data considering is a structure composed of three parameters: the consistency value of sensing data, the communication ability and the remained lifetime of a node. After the trust assertion, inconsistent data from malicious or compromised nodes can be detected.

3.3 DEFINITION OF TRUST MATRIX

When consider a cluster, we get a G = (V, E, s) consists of vertexes V, edges E and similarity weight s. Each vertex is a node and each edge is the connection of two neighbours. We compute the similarity among sensor nodes as Eq. (1) where node i and node j is adjacent in location. X is the sensing data of node. If $s_{i,j}$>0.9, we set new $s_{i,j}$ as 1, otherwise as 0.

$$s_{i,j} = \left[\frac{10 * X_i X_j}{X_i^2 + X_j^2 - X_i X_j} \right]. \tag{1}$$

We consider a window of time Δt. Thus, as time elapses, the window deletes old experiences but adds newer experiences. The trust value between two nodes can be calculated according to (2):

$$\mathrm{DSTi,k}(\Delta t) = \left[\left(\frac{10 * s_{x,y}(\Delta t)}{s_{x,y}(\Delta t) + d_{x,y}(\Delta t)} \right) \left(\frac{1}{\sqrt{d_{x,y}(\Delta t)}} \right) \right], \tag{2}$$

where $\left[\left(\frac{10 * s_{x,y}(\Delta t)}{s_{x,y}(\Delta t) + d_{x,y}(\Delta t)} \right) \left(\frac{1}{\sqrt{d_{x,y}(\Delta t)}} \right) \right]$ is the nearest integer function. $s_{i,k}(\Delta t)$ is the total number of similar data comparison of node i with k in Δt time, and $d_{i,k}(\Delta t)$ is the total number of dissimilar data comparison. Specially, if $d_{i,k}(\Delta t) = 0$, we set $ST_{i,k}(\Delta t)$=10.

The cluster head will periodically broadcast the request packet within the cluster. In response, all CMs in the cluster will forward their data values to CH. Then, CH will maintain these values in a matrix as shown below where the real number is the similarity of node i for node j and 1is a default value presenting the similarity of the node for itself.

$$\begin{matrix} DST_{1,1} & \cdots & DST_{1,n} \\ \vdots & \ddots & \vdots \\ DST_{n,1} & \cdots & DST_{n,n} \end{matrix}. \tag{3}$$

3.4 DEFINITION OF COMMUNICATION TRUST

The trust value based on communication between two nodes can be calculated according to (4):

$$\mathrm{DCT}_{i,k}(\Delta t)=\left\lceil\left(\frac{10*s_{i,k}(\Delta t)}{s_{i,k}(\Delta t)+f_{i,k}(\Delta t)}\right)\left(\frac{1}{\sqrt{f_{i,k}(\Delta t)}}\right)\right\rceil, \tag{4}$$

where $\left\lceil\left(\frac{10*s_{i,k}(\Delta t)}{s_{i,k}(\Delta t)+f_{i,k}(\Delta t)}\right)\left(\frac{1}{\sqrt{f_{i,k}(\Delta t)}}\right)\right\rceil$ is the nearest integer function. $s_{i,k}(\Delta t)$ is the success number of communication between node i and k in Δt time, and $f_{i,k}(\Delta t)$ is the failed number of communication. Specially, if $f(\Delta t) = 0$, we set $\mathrm{CT}_{i,k}(\Delta t)=10$ [8].

CH will maintain a matrix as shown in Eq. (5) where the number is the direct trust of node i for node k based on communication and 1is a default value presenting the trust toward itself.

$$\begin{matrix} DCT_{1,1} & \cdots & DCT_{1,n} \\ \vdots & \ddots & \vdots \\ DCT_{n,1} & \cdots & DCT_{n,n} \end{matrix}. \tag{5}$$

4 A light-weighted trust model

4.1 CALCULATE DIRECT TRUST

A CM’s trust value can be calculated by direct and indirect observation. Direct trust is evaluated by the number of successful and unsuccessful interactions, similar or dissimilar data comparison. In this work, interaction refers to the cooperation of two CMs and comparison refers to data aggregation. Indirect trust is evaluated by aid of similarity matrix in CH. That is, if node x wants to calculated the trust value for node y, first it checks whether it has a valid interaction with y during a specific time interval. If a past valid interaction record exists, then it compares its data value with y. Otherwise, if its remaining energy is less than ten percent, it will send a request to its CH. The model considers the consistency value of sensing data, the communication ability and the remained lifetime of a node. The process can be depicted in Figure 1.

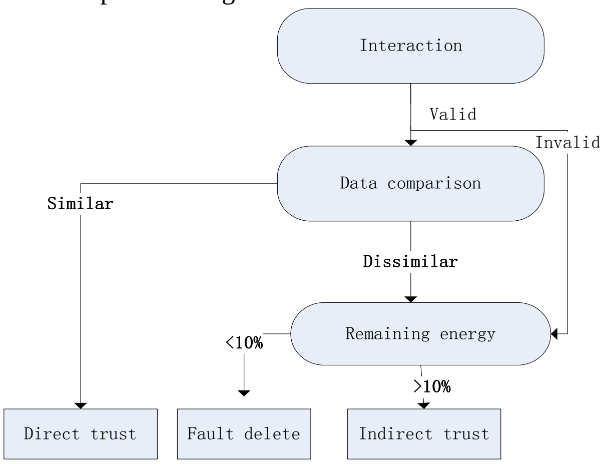


FIGURE 1 Process of the model

If the interaction $\mathrm{DCT}_{i,k}$ is more than 5 according to Eq. (4), it is regarded as valid. Then start second stage to calculate data comparison using Eq. (1). If data similarity is more than 5, the direct similar trust is as Eq. (2). The combined trust is as Eq. (6). Otherwise, check the remaining energy to decide whether to calculated the indirect trust or assert the node is fault to delete from the network:

$$\mathrm{DT}_{i,k=}\ \left\lceil\frac{DCT_{i,k}*DST_{i,k}}{10}\right\rceil. \tag{6}$$

If the interaction $\mathrm{DCT}_{i,k}$ is less than 5, it is regarded as invalid. Then we check the remaining energy to do the same work as above.

4.2 COMPUTE RETRANSMISSION RATE

After node i sends a data packet to its neighbouring node j in one-hop transmission range, it should receive an acknowledgement from node j. Otherwise, node i will retransmit the data packet. Retransmission in the link layer is supposed to be caused by some non-malicious factors such as the quality of wireless channels, node malfunction, etc., and by attacks in the routing layer. For node i, the non-malicious impact factor is calculated in (7) based on the retransmission rates of all its neighbours:

$$\theta=\frac{\sum_{K=1}^{N} t_{i,k}}{N}, \tag{7}$$

where N represents the number of node i's neighbouring nodes, and for node i the retransmission rate of the neighbouring node k within a certain period is denoted as t_{ik}, which is calculated by (8):

$$t_{i,k}=\frac{l}{m}, \tag{8}$$

where l represents the number of packets retransmitted from node i to node k, and m represents the total number of packets sent by node i to node k.

During Δt, if node i receives an acknowledgement from its neighboring node k, node i considers that the data packet has been successfully forwarded to the destination node through node k, and the number of successful forwarding times for node k is added by 1. Otherwise, the number of failed forwarding attempts for node k is added by 1. But the retransmission may compensate part of failed communication, so the real failed communication should be calculated again.

Since $cs_{i,k}(\Delta t)$ is the success number of communication between node i and k in Δt time, we can detail failed communication as uncertain communication as (9) and failed communication as (10):

$$cu_{i,k}(\Delta t)= cf_{i,k}(\Delta t)\ \theta\ , \tag{9}$$

$$cf_{i,k}(\Delta t)= cf_{i,k}(\Delta t)\ (1-\theta)\ . \tag{10}$$

4.3 CALCULATE INDIRECT TRUST

When entering the stage of calculating indirect trust, node i cannot determine the trust on k, it will request to CH for a feedback that can calculate the probability expectation based on data. We use the beta probability density functions to compute the indirect trust as Eq. (11) based on Eq. (3).

$$IST_{ch,k}=\left\lceil 10 * \frac{s_{i,k}+1}{s_{i,k}+d_{i,k}+2} \right\rceil. \qquad (11)$$

Here, $s_{i,k}$ denotes the number of similar feedback to node k except itself and $d_{i,k}$ denotes the number of dissimilar data to node k in a period Δt. For example, as shown in Figure 3, which is deduced from Figure 2 with setting the threshold as 9, we want to calculate indirect trust of node 1. The value is a real number of 6.7.

$$\begin{bmatrix} 10 & 9 & 0 & 9 & 9 \\ 9 & 10 & 1 & 0 & 9 \\ 0 & 1 & 10 & 0 & 1 \\ 9 & 0 & 0 & 10 & 9 \\ 9 & 9 & 1 & 9 & 10 \end{bmatrix}$$

FIGURE 2 Similarity matrix

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \end{bmatrix}$$

FIGURE 3 Trust matrix

At the same time, the feedback can also calculate the probability expectation based on communication. We use the beta probability density functions to compute the indirect trust as Eq. (12) based on matrix (5).

$$ICTch,\ k=\left\lceil 10 * \frac{cs_{i,k}+1}{cs_{i,k}+cf_{i,k}+cu_{i,k}+3} \right\rceil, \qquad (12)$$

where, $cs_{i,k}$ denotes the number of success communication to node k except itself, $f_{i,k}$ which is calculated from Eq. (10) denotes the number of failed communication to node k and $u_{i,k}$ denotes the number of uncertain communication in a period Δt.

Then the total indirect trust can be described as (13):

$$IT_{i,k}=\left\lceil \frac{ICT_{i,k}*IST_{i,k}}{10} \right\rceil. \qquad (13)$$

5 Evaluations

Our experiment uses *ns3* to design. Fifty sensor nodes are distributed in a space of 500×700, and the communication radius is set as 60. Each node has two to five neighbours in the experiment and the node's location is already known. The detailed value is shown in Table 1. Each node maintains a structure as shown in Table 2.

TABLE 1 Values in evaluation

Symbol	Description	Values
N	Number of nodes	50
n	Number of CMs in a cluster	6-8
m	Number of CM's neighbours	4-6

TABLE 2 Structure of nodes

<table>
<tr><th>Node</th><th colspan="3">The number of interaction</th><th colspan="2">The number of similar</th></tr>
<tr><th>ID</th><th>$s_{x,y}$</th><th>$f_{x,y}$</th><th>$u_{x,y}$</th><th>$s_{x,y}$</th><th>$d_{x,y}$</th></tr>
<tr><td>2 bytes</td><td>2 bytes</td><td>2 bytes</td><td>2 bytes</td><td>2 bytes</td><td>2 bytes</td></tr>
<tr><td colspan="3">Direct trust</td><td colspan="3">Indirect trust</td></tr>
<tr><td colspan="2"><i>communication</i></td><td><i>similar</i></td><td colspan="2"><i>communication</i></td><td><i>similar</i></td></tr>
<tr><td colspan="2">0.5 bytes</td><td>0.5 bytes</td><td colspan="2">0.5 bytes</td><td>0.5 bytes</td></tr>
</table>

We only consider the communication overhead with ignoring calculation cost. It is also assumed that the route is reliable without considering the case of route failure. We compare the communication consumption and error detection rate for simulation to LDTS and DMUTM.

It is shown in Figure 4 that when data error rate is changed, DMUTM maintains an average of packets by 2800. But LDTS and ours algorithm gets an increased average of packets as the data error rate growing. This is due to the calculating of indirect trust, which will consume more communication. While for LDTS, it only considers interaction, so the probability of calculating indirect trust is less than ours since our method considers both the interaction and data similarity.

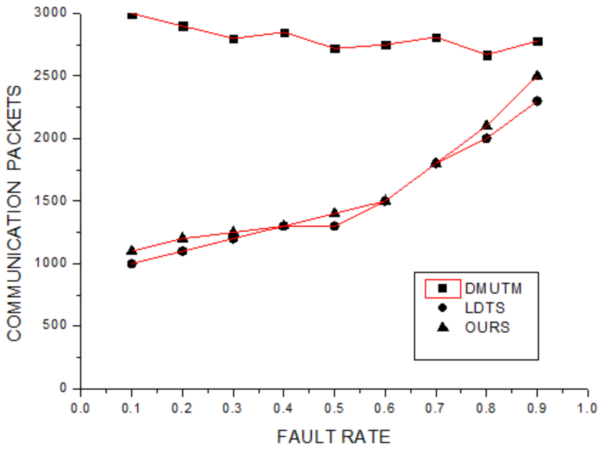


FIGURE 4 Comparison of communication consumption

Except for energy consumption, error detection rate is another important merit to measure a trust algorithm. We define error detection rate as f_s/f, where f_s is the number of fault nodes that have been detected and f is the total number of fault nodes.

Simulation result shown in Figure 5 indicates that the detection rate of DMUTM is higher than the other two methods because it handles all cases with indirect trust. And the higher detection rate is an exchange for communication consumption. LDTS has a lower detection rate than ours since it omits the data fault. And our former model has a lower detection rate than the current model since it omits the retransmission to regard fine node as error. Our model implements a balance between detection rate and communication consumption.

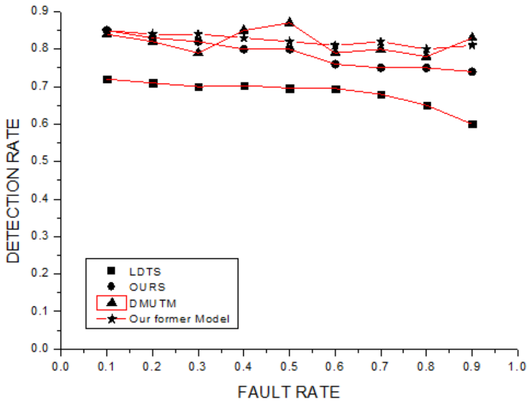


FIGURE 5 Comparison of fault detection rate

Although the advantage of our model, we can see from the structure that the memory overhead is double that of the LDTS.

6 Conclusions

In this paper, we investigate a method of light-weight trust calculating. A CM's trust value can be calculated by direct and indirect observation. Direct trust is evaluated by the number of successful and unsuccessful interactions, similar or dissimilar data comparison. Indirect trust is evaluated by aid of similarity matrix and communication matrix stored in CH. In order to distinguish uncertain from failed, we introduced a retransmission factor to make the communication result more explicit. This model can successfully detect fault nodes but consume more memory since it stored both data information and communication information. We did a series of simulations to test the performance of our proposed model. The simulation results have showed that our model has implemented a trade-off between detection rate and communication consumption.

In future, we should make full use of data information such as aggregating data at the same time when the trust is updated.

Acknowledgements

This work is supported by the National Basic Research Program of China (Grant No. 2011CB302802) and the National Natural Science Foundation of China (Grant No 61370100). Wang acknowledges the supported of the key discipline of Shanghai second polytechnic University named software engineering (No.XXKZD1301). The authors would like to thank the referees for their invaluable comments and suggestions.

References

[1] Riaz A S, Jameel H, d'Auriol B J, Sungyoung Lee H, Song Y-J *IEEE Transactions on Parallel and Distributed Systems* **20**(11) 1698-712

[2] Ganeriwal S, Srivastava M B 2004 Reputation-Based Framework for High Integrity Sensor Networks *Proceedings of ACM workshop security of ad hoc and sensor networks (SASN '04)* 66–7

[3] Wang Na, Chen YiXiang 2013 *Sensors &Transducers* **158-159**(12) 190-4

[4] Krasniewski M, Varadharajan P, Rabeler B, Bagchi Saurabh 2005 TIBFIT: Trust Index Based Fault Tolerance for Arbitrary Data Faults in Sensor Networks *Proceedings of the International Conference on Dependable Systems and Networks* 2005 672-81

[5] Ganeriwal S, Balzano L K, Srivastava M B 2008 *ACM Trans. Sensor Networks* **4**(3) 1–37

[6] Han Guangjie, Jiang Jinfang, Shu Lei, Niu Jianwei, Chao Han-Chieh 2014 Management and applications of trust in Wireless Sensor Networks: A survey *Journal of Computer and System Sciences* **80**(3) 602–17

[7] Sarma Dhulipala V R, Karthik N, Chandrasekaran R M 2013 A Novel Heuristic Approach Based TrustWorthy Architecture for Wireless Sensor Networks *Wireless Pers Commun* **70** 189–205

[8] Li Xiaoyong, Zhou Feng, Du Junping 2013 *IEEE Transactions on Information Forensics and Security* **8**(6) 924-35

[9] Chen Yixiang, Bu TianMing, Zhang Min, Zhu 2010 Measurement of Trust Transitivity in Trustworthy Networks *Hong Journal of Emerging Technologies in Web Intelligence* **2**(4) 319-25 *(in Chinese)*

[10] Liu Chen-xu, Liu Yun, Zhang Zhen-jiang 2013 Improved Reliable Trust-Based and Energy-Efficient Data Aggregation for Wireless Sensor Networks *International Journal of Distributed Sensor Networks* **2013** Article ID 652495 11 pages

[11] Wu Chunxue, Feng Bin 2007 Based on Single-hop Flow Control Scheme for Wireless Sensor Networks *IET Conference on Wireless, Mobile and Sensor Networks 2007 December 2007*

[12] Wu Chunxue 2006 Practical models and control methods with data packets loss on NCS *The IET International Conference on Wireless Mobile and Multimedia Networks January 2006*

[13] Zhang Guanghua, Zhang Yuqing, Chen Zhenguo 2013 Using Trust to Secure Geographic and Energy Aware Routing against Multiple Attacks *PLOS ONE* **8**(10) Oct 21 2013

[14] Wang Na, Wu YuePing 2013 Data aggregation for failure tolerance in wireless sensor network *Applied Mechanics and Materials* **347-350** 965-9

[15] Wang Na, Liu Dongqian, He Kangli 2013 A formal description for protocols in WSN based on STeC language *Proceedings of the 8th International Conference on Computer Science and Education, ICCSE 2013* 921-4

[16] Wang Na, Gao Liping, Wu Chunxue 2014 A Light-Weighted Data Trust Model in WSN *International Journal of Grid and Distributed Computing* **7**(2)

Authors

	Na Wang, born on June 6, 1979 He Nan **Current position, grades:** PhD candidate , a lecturer of Shanghai second University **University studies:** East China Normal University, Shanghai Second Polytechnic University **Scientific interest:** Wireless sensor networks **Publications:** 12
	Yanxia Pang, born on December 23, 1980 Shan Dong **Current position, grades:** PhD candidate, a lecturer of Shanghai second University **University studies:** Shanghai Second Polytechnic University **Scientific interest:** Data mining **Publications:** 5

An approach of VM image customized through Linux from scratch on cloud platform

Gaochao Xu[1, 2], Yushuang Dong[1*], Bingyi Sun[1], Xiaodong Fu[1], Jia Zhao[1]

[1] *Department of Computer Science and Technology, JiLin University, Changchun 130012, China*

[2] *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China*

Received 1 March 2014, www.tsi.lv

Abstract

The cloud platform provides abundant resources and services for users. More and more mobile users began to use the cloud services. They have higher real-time demands on service. The size of traditional virtual machine (VM) operating system is basically large. It will take many resources in deployment and communication process, and always affect the real-time performance of system. To reduce communication overhead and improve deployment speed of VMs, this paper proposes an approach of customized VM image with LFS. LFS can reduce the size of VM image efficiently and enable flexible customization of the VM image by incremental customization. The experimental results show us that the size of VM image generated by the proposed method is smaller than the one generated by kernel tailoring technology in system overhead. Meanwhile it is also faster in running speed.

Keywords: Cloud computing, Linux from scratch, Customized virtual machine

1 Introduction

Cloud computing is at the forefront of the information technology. With the development of mobile devices and the increasing requirement of mobile user, cloud computing begin to be applied in providing service for mobile users. The size of traditional VM image is so large that it has a profound impact on real-time. If the size of VM image could be decreased, we can realistically reduce system overhead and communication overhead for cloud platform. With cloud computing application development on mobile platforms and other devices, the micro-kernel technology will be more demanded under cloud computing environment. This paper will analyze how to get customized VM image through LFS. The full name of LFS is "Linux from Scratch" [1]. LFS refers to building a Linux system manual or an idea rather than a release version of Linux. Unlike ordinary Linux installation, LFS guides to compile the open-source software packages into a needed smallest, fastest Linux system through the host system. Users can control all features of the new system during the build process such as Installation Directory and File organization form, parameters and permissions settings, etc. We can improve the real-time performance effectively by using the customized system.

In section 2 of this paper, we will give a brief introduction of related work about current VM image customized methods. In section 3, the design and implementation of customize VM image through LFS will be presented in detail. Than in section 4, experiments undertaken and results obtained will be shown which demonstrate that the new method provides an effective solution. Finally, we will conclude the paper in section 5.

2 Relevant work

Because of the large size of the traditional VM image, the real-time performance of system is poor and too many system resources are employed. It becomes necessary to reduce the system size. Although many Linux kernel-cutting methods can efficiently decrease the size of Linux kernel, these methods are usually used to customize the Linux kernel for embedded system [2, 3] and many other specific fields [4]. These methods are not appropriate for cloud computing and micro-kernel technology is not yet universally applied in the cloud computing.

Application of virtualization technology proposed software as a service model. As App-V [5] of Microsoft, ThinApp [6] of VMware and Citrix XenApp [7], individual does not need to consider the process of installation, maintenance and upgrade. These operations can be completed by software service providers. Users obtain the right to use the software via user identity authentication mechanism. They distinguish the application and the OS, the centralized management, the maintenance and upgrading of the software. Software provider on-demand provides the software for user; users can use them without installation. However, most of the proposed plans are business plans for windows, and their application needs to run with network support. Therefore, OpenAppV [8] proposed on-demand Customized Virtual Machine Instance System. These methods realized software customization, but did not reduce the system

* ***Corresponding author*** e-mail: yushuangdong@gmail.com

size. In general, VM developed by the traditional methods is often difficult to be understood and modified. It is not satisfied enough in extensibility and reusability. It is very difficult to dynamically change and extend the VM. It needs frequently to full-manually override the existing VM [9, 10]. Reference [11] proposed An Implementation Approach to Custom-Built Virtual Machines. However, this method is relatively slow and not suitable for high real-time systems. To reduce the system size and to improve the real-time, we will analyze getting the customized VM image through LFS in this paper.

3 VM image customized with LFS

At present, jhalfs can realize the automatic installation of LFS by extracting the command from the XML of lfsbook. jhalfs can choose flexibly whether LFS process tests the installation and optimizes the system etc. by setting common/config file under the jhalfs. In order to meet users' requirements, we need generate customized VM image with the minimum matching image incremental installation, for which, we need to get the configuration information of users' requirements through the management process of the cloud platform. Therefore, this paper analyzes how to realize automatic installation of LFS, and the application software and services based on our own shell scripts.

GCC is a GNU [12] compiler containing C / C + + compiler, which can compile the source code program to generate the target file in combination with other basic tools. Binutils [14] is a set of binary tools including connectors, assembler and other tools used in the target file and it can transform the target file into an executable file. Make is a program similar to the shell script, which can control the entire compile-link process through reading a makefile that contains the source code and document library dependencies and rules. Glibc, the crucial link in the production process of LFS, is also very important for Linux. All dynamic linker must use it. In the process, aforementioned four tools are dependent on the C Runtime library glibc.

The toolchain is a temporary build environment. It is used to generate the target system that includes all necessary tools such as GCC and binutils. Their versions have exactly the same as the tools of the target system.

As the Linux kernel source code we want to compile requires a specific toolchain and glibc version and an environment that are inconsistent with the host system. Therefore, we have to first build a specific toolchain and glibc to get the target system. For this purpose, all the tools on the toolchain are compiled by their respective source codes. GCC, binutils, make, common tools in the compilation process are dependent on glibc. Similarly, the glibc needed is a particular one rather than the one in the host system. This particular glibc is compiled by another tool chain, which we call a temporary toolchain. While this temporary toolchain is compiled by host systems obtained through a set of own compilation tools, glibc, other C libraries and Linux kernel, etc.

The implementation process of LFS shown in Figure 1:

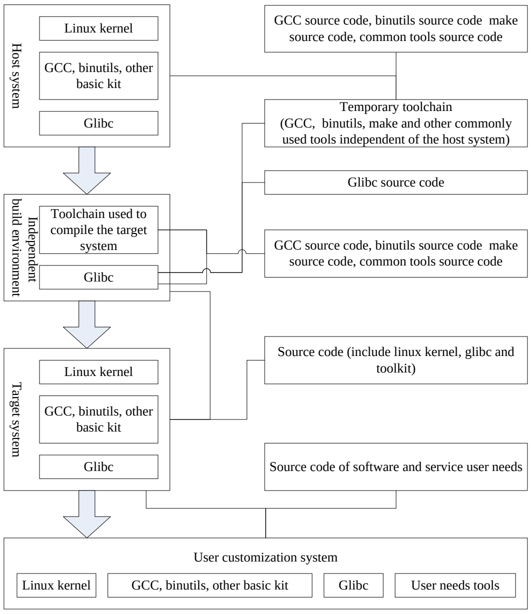

Host system
Linux kernel
GCC, binutils, other basic kit
Glibc
GCC source code, binutils source code make source code, common tools source code
Temporary toolchain (GCC, binutils, make and other commonly used tools independent of the host system)
Glibc source code
Independent build environment
Toolchain used to compile the target system
Glibc
GCC source code, binutils source code make source code, common tools source code
Target system
Linux kernel
GCC, binutils, other basic kit
Glibc
Source code (include linux kernel, glibc and toolkit)
Source code of software and service user needs
User customization system
Linux kernel
GCC, binutils, other basic kit
Glibc
User needs tools

FIGURE 1 LFS implementation process

1) Temporary tool chain compiled by source code of the tool set on the host system;

2) Get an independent glibc library by compile the glibc source code using the temporary tool chain?

3) Use the independent glibc library in the host system environment to build independent toolchain, independent build environment completed;

4) Compile the Linux source code package in the independent build environment and build the Linux kernel for the operating system.

5) Get the user request information and compiled required software and services in the target system according to user requests. Then generate customized system image.

6) Finally, do some final changes work as removing the source package, temporary toolchain, and independent build environment, and restore the original system parameters and configuration files.

The process of customizing VM image as showed in Fig. 2:

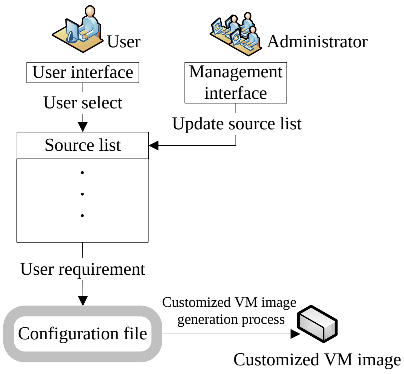
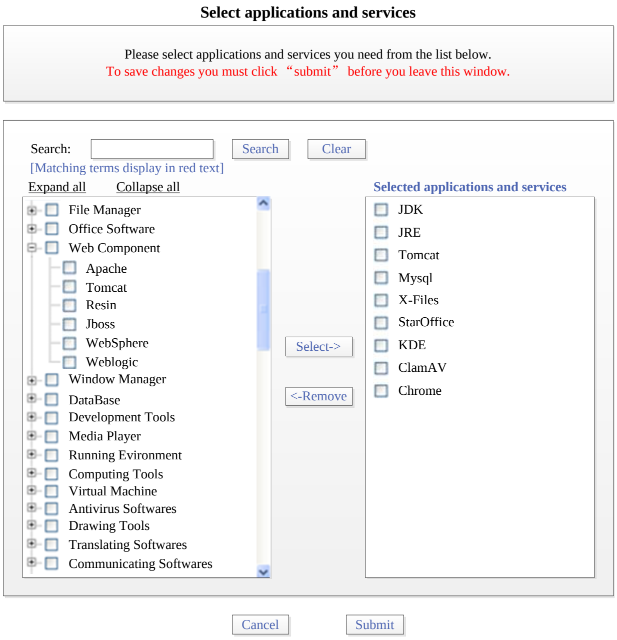


FIGURE 2 LFS implementation process

3.1 USER INTERFACE

It is easy to access and operate by using the interactive interface of traditional web, and gives a list of application software and services in a platform. User can log in the platform for selecting their needs through a user interactive interface, as showed in Fig. 3. Sent through the network management process to the cloud platform, and user request information is stored in the management side.



Select applications and services

Please select applications and services you need from the list below.
To save changes you must click "submit" before you leave this window.

Search: Search Clear
[Matching terms display in red text]
Expand all Collapse all

□ File Manager
□ Office Software
□ Web Component
□ Apache
□ Tomcat
□ Resin
□ Jboss
□ WebSphere
□ Weblogic
□ Window Manager
□ DataBase
□ Development Tools
□ Media Player
□ Running Evironment
□ Computing Tools
□ Virtual Machine
□ Antivirus Softwares
□ Drawing Tools
□ Translating Softwares
□ Communicating Softwares

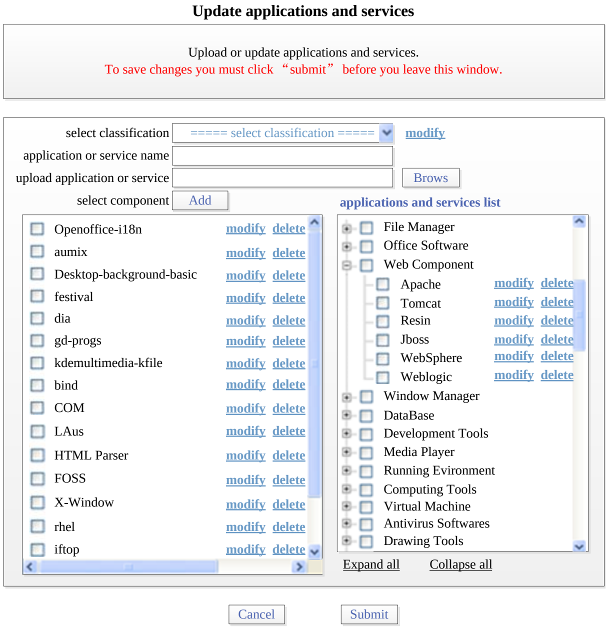
Select->
<-Remove

Selected applications and services
□ JDK
□ JRE
□ Tomcat
□ Mysql
□ X-Files
□ StarOffice
□ KDE
□ ClamAV
□ Chrome

Cancel Submit

FIGURE 3 User interface

Shown in Fig. 4, the applications and components provided by the management side of the cloud platform also can be dynamically updated when new applications or components need to be updated to the cloud platform. Administrators use an interface to upload applications or components. When the list of components is missing from the system components that applications relied on, it needs administrators upload the missing system components and the management process generated script code used to complete the installation of the customization process. Administrators of cloud platform can dynamically update or delete the existing applications and components. Management process records the information when administrators upload applications and components. The information mainly includes five parts: (a) id of application or component, (b) name of application or component, (c) size of application or component, (d) components information that application relied on, (e) installation script information.



Update applications and services

Upload or update applications and services.
To save changes you must click "submit" before you leave this window.

select classification ===== select classification ===== modify
application or service name
upload application or service Brows
select component Add

□ Openoffice-i18n	modify	delete
□ aumix	modify	delete
□ Desktop-background-basic	modify	delete
□ festival	modify	delete
□ dia	modify	delete
□ gd-progs	modify	delete
□ kdemultimedia-kfile	modify	delete
□ bind	modify	delete
□ COM	modify	delete
□ LAus	modify	delete
□ HTML Parser	modify	delete
□ FOSS	modify	delete
□ X-Window	modify	delete
□ rhel	modify	delete
□ iftop	modify	delete

applications and services list

□ File Manager		
□ Office Software		
□ Web Component		
□ Apache	modify	delete
□ Tomcat	modify	delete
□ Resin	modify	delete
□ Jboss	modify	delete
□ WebSphere	modify	delete
□ Weblogic	modify	delete
□ Window Manager		
□ DataBase		
□ Development Tools		
□ Media Player		
□ Running Evironment		
□ Computing Tools		
□ Virtual Machine		
□ Antivirus Softwares		
□ Drawing Tools		

Expand all Collapse all

Cancel Submit

FIGURE 4 Applications and services management interface

3.2 CONFIGURATION FILE

Users select their needs through user interface and send the request information to the management process. The management process receives the request information and extracts it to form a customized XML configuration file. The information of the configuration file mainly includes three parts: (a) hardware requirements include CPU, hard disk, memory, and number of computing nodes, (b) applications requirements, (c) components requirements. We use id to record applications and components that user requested in configuration file. Configuration file fragment is shown as follows:

```
1: <user-request>
2:   <hardware name=”userid”>testuser</hardware>
3:   <hardware name=”vmnum”>20</hardware>
4:   <hardware name=”cpu”>1</hardware>
5:   <hardware name=”ram”>1024</hardware>
6:   <hardware name=”disk”>80</hardware>
7:   <application name=”id”>2, 6, 8, 9, 11</application>
8:   <component name=”id”>4, 7, 14</component>
9: </user-request>
```

Lines 3-6 record hardware requirements include the size of virtual cluster, frequency of VM CPU, memory size of VM and disk size of VM. Lines 7-8 record the id of applications and components that satisfy the user requirements.

3.3 CUSTOMIZED IMAGE

The management process extracts the information I_{conf} from the configuration file. The information of I_{conf} includes the user request. According to I_{conf}, user needs K

VM nodes. There are N exist VM image copies stored on cloud platform. The VM image copy is used to generate VM image that users customized previously. The initial minimized image is P_0. The match degree is calculated by matching configuration file information of VM image copy with I_{conf}. The process traverses the copy and matches the copy of node c ($0 =< c < N$) with I_{conf}. If the copy node c exists other applications or services beyond I_{conf}, the node cannot be a matching node. If the copy of node c does not exist other applications or services and the match degree is greater than the former one, then $P_{target} = P_c$. If the copy node c matches exactly with I_{conf}, then $P_{target} = P_c$, now finish the traversal process to get customized VM image P_{target}. If there is not an exactly matching node, we need to completely traverse the copy in order to get P_c with the largest match degree to obtain customized VM image P_{target} and to install the application software and services in I_{conf} that P_{target} does not have to get customized VM image P_{target}. Then we store it in a virtual node and update the copy of the image to store this node information in that copy. The customized VM image generation process is as follows:

```
1:  $P_0 \leftarrow$ initial VM image copy
2:  $S_0 \leftarrow$ size of $P_0$
3:  $I_0 \leftarrow$ configuration file information of $P_0$
4:  $M_0 \leftarrow$ match degree of $P_0$
5:  $P_{target} \leftarrow$ target VM image copy
6:  $S_{target} \leftarrow$ size of $P_{target}$
7:  $M_{target} \leftarrow$ match degree of $P_{target}$
8:  $P_{target} = P_0, S_{target} = S_0, M_{target} = M_0$
9:  $c = 0$
10: While $c < N$ do
11:     If $I_c$ exactly matches with $I_{conf}$
12:         $P_{target} = P_c$
13:         Break
14:     Else
15:         If $I_c$ exist other applications or components information beyond $I_{conf}$
16:             $S_c = 0, M_c = -1$
17:         Else
18:             If $M_c > M_{target}$
19:                 $P_{target} = P_c, S_{target} = S_c, M_{target} = M_c$.
20:             End if
21:             If $M_c = M_{target}$ and $S_{target} < S_c$
22:                 $P_{target} = P_c, S_{target} = S_c, M_{target} = M_c$
23:             End if
24:         End if
25:     End if
26:     $c = c + 1$
27: End while
28: If $c == N$
29:     $P_{temp} \leftarrow$ According to $I_{conf}$, complete the $P_{target}$ installing
30:     $P_{target} = P_{temp}$
31: End if
32: store the $P_{target}$ and $I_{target}$ on cloud platform, set the host as target host
33: generate customize VM image through $P_{target}$, create $K$ VM nodes
```

Lines 1-4 give VM image copy P_0 as input that has no extra applications or components. P_0 is generated by LFS process. Lines 5-8 initialize the target VM image copy P_{target} as output. Lines 10-27 find the target VM image copy that mostly satisfies the user's requirements. If the VM image copy exists other applications or components beyond the user's requirements, it is not the VM image we need. If there are two or more VM image copies mostly satisfy the user's requirements, we should compare the size of these VM image copies and choose the biggest one as the target VM image copy. Lines 28-30 show that if there is no VM image copy which exactly satisfies the user's requirements, we can get a VM image copy that satisfies most the user's requirements, then according to user's requirements, complete the target VM image copy installation and get a new target VM image copy. Lines 32-33 store the target VM image copy on cloud platform and set the host that stored the target VM image copy as target host. Finally we generate customize VM image through the target VM image copy and create K VM nodes that satisfy the user hardware requirements.

4 Experiment analysis

To establish a minimum system requires about 1.3GB of the partition so as to have enough space to store and compile all the source packages. A larger space (2~3GB) is needed to install the application software and services that the user needs. The LFS system itself does not occupy so much space and most of the space required is used to provide adequate temporary space for the software compiler. It takes many temporary spaces to compile the package. However, these temporary spaces can be recycled after software installation done. We would better to use a small hard disk partition as swap space because memory (RAM) is not always enough during the compilation process. The kernel uses swap space to store the data in order to free up memory space for running processes. The swap partition that LFS system uses can be the same with the one that the host system uses. Therefore, we do not have to create a new one for the LFS system when the host system already has a swap partition. In this paper, we suggest getting customized VM image through LFS in order to achieve the goal of reducing system and communication consumption.

Experiment Environment: In our experimental configuration, hosts with the same type are selected. We use HP proLiant ML350 G6 (AU662A) as hosts in cloud platform. These hosts are configured with Xeon E5506 2.13GHz four core processor, 8GB DDRIII RAM, 4TB 7.2K 6Gbps hard disk and NC326i PCI Express 1000Mbit/s NIC. In order to simplify the process of customized VM image generation, we install LFS Live CD 6.2-3 with kernel 2.6.16.26 on all hosts as host system. Virtual Tool is Xen 4.1.1. We configure all VMs with single core, 40GB VM hard disk. To ensure parent-

VM boot successfully, we configure VM memory size with 512M.

Experiment: To reduce system resource occupation and to fundamentally solve the problem of too long virtual machine downtime in the deployment process, the VM image size should be as small as possible. We compare the customized VM image generated by LFS with current lite release version of Linux.

Experimental comparison results are showed in Figure 5, 6. We customize VM image as web servers. We can limit the size of customized VM image generated by LFS at 38.63M. Comparing with other lite release version of Linux, LFS visibly reduced the system consumption, simplified customize process, and it is much easier in the customize process. It also has advantage in booting speed and system consumption. We configure VM memory size with 512M. LFS takes only 6.76s to complete the booting process. The boot speed of VM that load customized VM image generated by LFS is faster than others. By reducing the VM image size, we effectively reduce the VMs communication consumption.

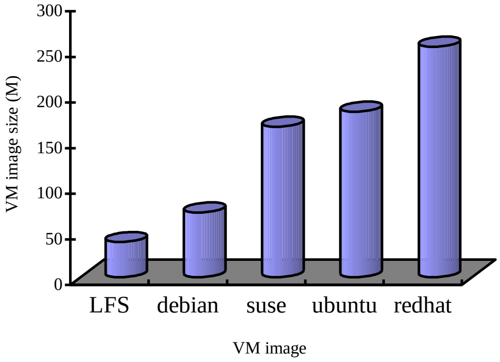


FIGURE 5 System space occupation comparisons

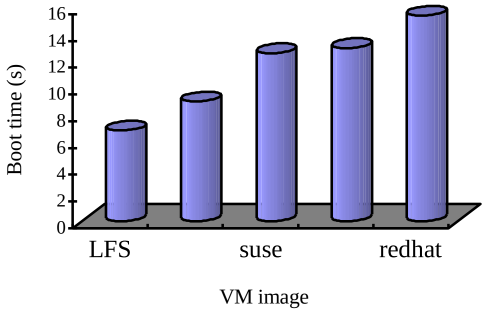


FIGURE 6 System booting time comparisons

5 Conclusion

Cloud computing has a promising development prospects and the related key technologies are growing rapidly. In this paper, we give a summary of the existing VM image customized technology and present the design, implementation and evaluation of customized VM image through LFS. We customize the VM image with LFS in order to reduce the VM image size, thus decrease the system overhead and communication overhead. The customized VM satisfies the user’s request and consume less space.

To further improve the performance of VM image customized, there are still many problems that need to be solved in the future. In customized VM image generation process, it traverses the copy to find the target host. It is not suitable for high real-time demanded cases. We plan to design a faster matching method to find the target host. We find that the target host takes only the match degree in consideration. In a cloud platform with workload or in a heterogeneous cloud platform, it is not enough and the target host we find may be not the best one. We plan to add the performance parameter for finding the target host.

References

[1] Beekmans G 2007 *Linux From Scratch* http://www.linuxfromscratch.org/lfs/view/stable/
[2] Fröhlich A A, Schröder-Preikschat W 1999 Tailor-made operating systems for embedded parallel applications *Lecture Notes in Computer Science* 1361-73
[3] Hasan M Z, Sotirios S G 2008 Customized kernel execution on reconfigurable hardware for embedded applications *Microprocessors and Microsystems* 211-20
[4] Montgomery J, Brewster G B, Yee W G 2010 A customized Linux Kernel for Providing Notification of Pending Financial Transaction Information *7th IEEE Consumer Communications and Networking Conference* 1021-2
[5] APP-V [EB/OL]. http://www.microsoft.com/app-v
[6] ThinApp [EB/OL]. http://www.vmware.com/products/thinapp/
[7] XenApp [EB/OL]. http://www.citrix.com/xenapp
[8] Zhang Hanying, Wu Qingbo, Tan Yusong 2013 On demand Customized Virtual Machine Instance System *Computer Technology and Development* **23**(4) 1-10
[9] Ert l M A, Gregg D, Krall A, et al. 2002 Vmgen: A Generat or of Efficient Virtual Machine Interpreters *Software-Practice & Experience* **32**(3) 265-94
[10] Ert l M A, Gregg D 2004 The Structure and Performance of Efficient Interpreters *Proc of the 2004 Workshop on Interpreters, Virtual Machines and Emulators*
[11] Ouyang Xingming, Zhu Jinyin 2008 An Implementation Approach to Custom-Built Virtual Machines and Their Dynamic Optimization *Computer Engineering & Science* **30**(1) 129-41
[12] GNU Binutils http://en.wikipedia.org/wiki/GNU_Binutils

Authors

Gaochao Xu, born in 1966, Xiaogan City, Hubei Province, China

Current position, grades: Professor, Doctor
University studies: Computer Science and Technology
Scientific interest: Distributed System, Grid Computing, Cloud Computing, Internet Things, etc.
Publications: 55
Experience: Professor and PhD supervisor of College of Computer Science and Technology, Jilin University, China.

	Yushuang Dong, born in 1983, Jixi City, Heilongjiang Province, China **Current position, grades:** PhD **University studies:** Computer Science and Technology **Scientific interest:** Distributed System, Cloud Computing. **Publications:** 8 **Experience:** PhD of College of Computer Science and Technology, Jilin University, China.
	Bingyi Sun, born in 1991, Changchun City, Jilin Province, China **Current position, grades:** Master Degree Candidate **University studies:** Computer Science and Technology **Scientific interest:** Distributed System, Cloud Computing. **Experience:** Master Degree Candidate of College of Computer Science and Technology, Jilin University, China.
	Xiaodong Fu, born in 1966, Changchun City, Jilin Province, China **Current position, grades:** senior engineer **University studies:** Computer Science and Technology **Scientific interest:** Distributed System, Cloud Computing. **Publications:** 14 **Experience:** senior engineer of College of Computer Science and Technology, Jilin University, China.
	Jia Zhao, born in 1982, Changchun City, Jilin Province, China **Current position, grades:** Doctor **University studies:** Computer Science and Technology **Scientific interest:** Distributed System, Cloud Computing. **Publications:** 11 **Experience:** Doctor of College of Computer Science and Technology, Jilin University, China.

Blind multi-image super resolution reconstruction with Gaussian blur and Gaussian noise

Fengqing Qin*

Institute of Computer Science and Technology, Yibin University, Wuliangye Str.8, Yibin, China

Received 1 March 2014, www.tsi.lv

Abstract

A framework of blind multi-image super resolution reconstruction method is proposed to improve the resolution of low resolution images with Gaussian blur and noise. In the low resolution imaging model, the shift motion, Gaussian blur, down-sampling, as well as Gaussian noise are all considered. Firstly, the Gaussian noise in the low resolution image is reduced through Wiener filtering method. Secondly, the Gaussian blur of the de-noised image is estimated through error-parameter analysis method. Thirdly, the motion parameters are estimated. Finally, super resolution reconstruction is performed through iterative back projection algorithm. Experimental results show that the Gaussian blur and motion parameters are estimated with high precision, and that the Gaussian noise is restrained effectively. The visual effect and peak signal to noise ratio (PSNR) of the super resolution reconstructed image are enhanced. The importance of Gaussian blur estimation and effect of Gaussian de-noising in multi-image super resolution reconstruction are tested in an experimental way.

Keywords: blind, multi-image super resolution, Gaussian blur, Gaussian noise, iterative back projection

1 Introduction

High resolution (HR) images are often required in many imaging applications. HR means that the pixel density within and image is high, and can offer more details. To improve the spatial resolution of image, the most direct way is to improve the precision and stability of the imaging system with expensive cost and some technical difficulties. Super resolution (SR) method is an efficient way with lower cost than hardware method.

In general, SR includes video SR [1] and image SR. Video SR refers to reconstructing a higher resolution video form a low resolution (LR) video by utilizing the redundancy information between the adjacent frames and the prior information of the imaging system. Image SR refers to reconstructing a higher resolution from one image or a set of images acquired from the same scene. On the basis of the number of the LR images, image SR image SR mainly includes multi-image SR [2-4] and single-image SR [5-7]. Multi-image SR is commonly researched, in which the movement with sub-pixel precision is estimated and utilized to reconstruct a HR image. Thus, image registration is very important in multi-image SR.

In many practical applications, the image restoration problem is always blind, which means that the PSF is most likely unknown or is known only to within a set of parameters [8]. In iterative back projection (IBP) SR reconstruction algorithm, the more accurate the imaging model is estimated, the better quality of the reconstructed image will be achieved. However, in most of the current algorithms, the blur is assumed to be a known Gaussian point spread function (PSF) with given parameters, or the blur is not considered at all in some algorithms, which does not meet the real imaging model of optical devices and limits the SR reconstruction quality. Thus, the blind image SR reconstruction [9-10] is one advanced issue and challenge in image restoration, which is expressed as estimating a HR image and the PSF simultaneously. The foremost difficulty of blind de-blurring is rooted in the fact that the observed image is an incomplete convolution. The convolution relationship around the boundary is destroyed by the cut-off frequency, which makes it much more difficult to identify the blurring function.

In addition, the process of noise is seldom considered in the LR imaging model, which restrained the quality of the SR reconstructed image. The noise can worsen the quality of images and bring some difficulty to image analyzing. Thus, noise should be considered in the framework of multi-image SR reconstruction.

In this paper, a framework of blind multi-image SR reconstruction method with Gaussian blurs and noise is proposed. In the LR imaging model, the processes of Gaussian blur, down-sampling, as well as noise are all considered. The Gaussian noise is reduced through Wiener filtering method. The Gaussian blur of the de-noised LR image is estimated through error parameter analysis method. The SR image is reconstructed through iterative back projection (IBP) algorithm.

**Corresponding author* e-mail: qinfengqing@163.com

2 Framework of blind multi-image SR reconstruction with Gaussian blur and noise

The framework of multi-image SR reconstruction with Gaussian blur and noise is shown in Fig.1. Firstly, the Gaussian noise in the low resolution image is reduced through Wiener filtering algorithm. Secondly, the movement parameters between the de-noised LR images are estimated. Thirdly, the Gaussian blur of the de-noised image is estimated through error-parameter analysis method. Finally, super resolution reconstruction is carried out through iterative back projection algorithm.

2.1 THE LR IMAGING MODEL

In the LR imaging model, the shift movement, Gaussian blur, down-sample, and Gaussian noise are all considered, as shown in Fig.1. The mathematical description of LR imaging model of multi-image SR reconstruction may be expressed as follows:

$$Y=EBDF+N, \qquad (1)$$

where, Y is the LR image; F is the HR image; E is the movement; B is the blur function; D is the down-sample process; N is the noise.

The real scene may be expressed by a high resolution (HR) image. Firstly, the HR is moved vertically and horizontally. The moved image is blurred by convolving with a point spread function (PSF). The blur mainly includes the Gaussian blur induced by the optical devices of the imaging system, the motion blur caused by the movement of the scene or the camera, as well as the defocus blur bringing by the false focus while imaging, etc. As Gaussian blur is the most common and is considered here. Secondly, the blurred image is down-sampled by a given integer factor. Here, the down-sampled image is gained by taking the neighborhood average gray value of the blurred image. Thirdly, the down-sampled is noised to generate the LR image. Here, the Gaussian noise is considered.

2.2 ITERATIVE BACK PROJECTION METHOD

Among the current SR reconstruction methods, the iterative back projection (IBP) method has the virtues of small computational amount, fast convergent rate, good reconstruction effect, and so on. In addition, the estimated information about the LR imaging model can be well utilized in the IBP algorithm.

In IBP algorithm, If the LR imaging model is estimated more accurately, the SR reconstructed image will achieve better quality. By back projecting the estimation error between the estimated LR image and the original image onto the HR image grid, the estimation error is gained to modify to estimated HR image. Repeating the iterative process until the iteration time is greater than a given number or the estimation error is less than a threshold, the SR image will be gained.

According to this idea, the IBP algorithm may be expressed as follows:

$$\hat{f}_i^{k+1} = \hat{f}_i^{k} - \lambda H_i^{BP}(\hat{g}_i^k - g_i^{'}) . \qquad (2)$$

Here, the initial value of the estimated HR image is taken as the interpolated image of the LR image by Bilinear interpolation algorithm. According to the LR imaging model proposed in this paper and the idea of IBP algorithm, the framework of the multi-image SR reconstruction method with noise is shown in Fig.1. Here, P is the number of LR images, i=1,…,P; k is the iteration time; $\hat{f}$ is the estimated SR image; g is the observed LR image; y' is the de-noised LR image; $\hat{g}$ is the simulated LR images of $\hat{f}$; E, B and D are the matrix forms of the motion blur and down-sampling respectively; n is the system noise; E^{-1}, B^{-1}, D^{-1} and n^{-1} denote the inverse operation of E, B, D and n; H^{BP} is the back projection operation; $\hat{g}-g'$ is the difference of simulated LR image and the de-noised LR image; λ is the gradient step.

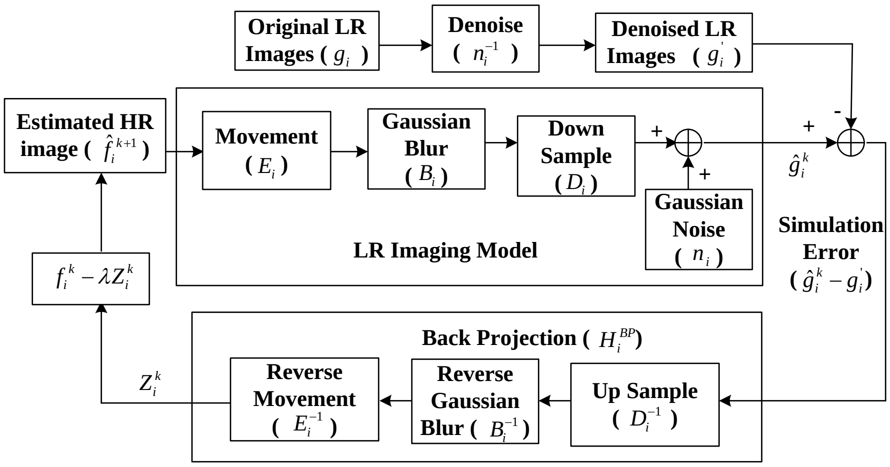

Original LR Images (g_i)
Denoise (n_i^{-1})
Denoised LR Images ($g_i^{'}$)
Estimated HR image ($\hat{f}_i^{k+1}$)
Movement (E_i)
Gaussian Blur (B_i)
Down Sample (D_i)
+
+
Gaussian Noise (n_i)
LR Imaging Model
+
-
$\hat{g}_i^k$
Simulation Error ($\hat{g}_i^k - g_i^{'}$)
$f_i^k - \lambda Z_i^k$
Back Projection (H_i^{BP})
Z_i^k
Reverse Movement (E_i^{-1})
Reverse Gaussian Blur (B_i^{-1})
Up Sample (D_i^{-1})

FIGURE 1 Framework of blind multi-image SR reconstruction with Gaussian blur and noise

2.3 WIENER FILTERING DE-NOISING

In Wiener filtering algorithm, both the blur function and the statistical character of system noise are considered. The noise is assumed to be a random process. The aim is to make the mean square error between the original image and the estimated image to be the least. According to this idea, the sketch map of Wiener filtering may be expressed in Fig.2.

The observed low resolution image may be expressed as follows:

$$y(n)=\sum_{k=-\infty}^{\infty}x(n-k)h(k)+\xi(n)=x(n)*h(n)+\xi(n), \qquad (3)$$

Where, * is the convolution operation; $x(n)$ is the original high resolution image; $\xi(n)$ is Gaussian white noise with zero mean; $h(n)$ is the blur function, which can be presented by point spread function (PSF).

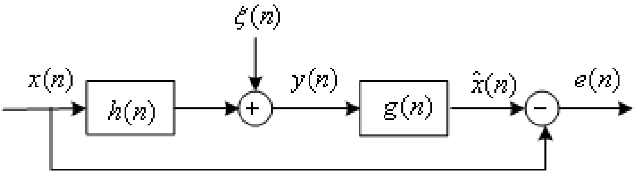


FIGURE 2 The sketch map of Wiener filtering.

When the discrete Fourier transform (DFT) method is used to estimate the restored image, the Wiener filter may be expressed as follows:

$$X=\frac{H^{*}Y}{|H|^{2}+S_{nn}/S_{xx}}, \qquad (4)$$

where, X, Y and H are the DFT of the real image (x), the blurred image (y) and the blur function (h) respectively; S_{nn} and S_{xx} denote the power spectrum of the noise and the real image. As it is usually very difficult to estimate S_{nn} and S_{xx}, the Wiener filter is usually approximated by the following formula:

$$X=\frac{H^{*}Y}{|H|^{2}+\Gamma}, \qquad (5)$$

where, Γ is a positive constant, which is often taken as an experience value.

2.4 MOVEMENT ESTIMATION

Here, the globle movement with vertical shift and horizontal shift are considered. If the reference image is $r(x',y')$, and the other image is $s(x,y)$, a and b are the horizontal shift and vertical shift respectively, the rigid transformation model between the coordinates of these two images may be denoted as:

$$\begin{bmatrix}x'\\y'\end{bmatrix}=\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}a\\b\end{bmatrix}. \qquad (6)$$

The mathematical relationship of these two images can be expressed as follows:

$$s(x,y)=r(x',y')=r(x+a,y+b). \qquad (7)$$

Two-dimensional series expansion at (x,y) is made to the right part of the preceding equation. Ignoring the high order terms, the following approximate expression will be get:

$$s(x,y)\approx r(x,y)+a\frac{\partial r}{\partial x}+b\frac{\partial r}{\partial y}. \qquad (8)$$

Thus, the object function can be written as:

$$E(a,b)=\sum\left[r(x,y)+a\frac{\partial r}{\partial x}+b\frac{\partial r}{\partial y}-s(x,y)\right], \qquad (9)$$

where, $\sum$represents the summation to the overlapped part of r and s. Monimizing the object fuction. Performing partial derivatives about a and b respectively and letting them equal to zero, the optical estimated parameters will be obtained:

$$X=A^{-1}B, \qquad (10)$$

where,

$$X=\begin{bmatrix}\hat{a}\\\hat{b}\end{bmatrix},\ A=\begin{bmatrix}\sum\frac{\partial r}{\partial x} & \sum\frac{\partial r}{\partial y}\end{bmatrix},\ B=\sum(s-r), \qquad (11)$$

2.5 GAUSSIAN BLUR ESTIMATION

Gaussian PSF is the most common blurring function of many optical measurements and imaging systems. Generally, the Gaussian blurring function may be expressed as follows:

$$h(m,n)=\begin{cases}\frac{1}{\sqrt{2\pi\sigma}}\exp\{-\frac{1}{2\sigma^{2}}(m^{2}+n^{2})(m,n)\in C, \\ 0 \qquad\qquad\qquad \text{others}\end{cases} \qquad (12)$$

where, σ is the standard deviation; C is a supporting region. Commonly, C is denoted by a matrix with size of $K\times K$, and K is often an odd number.

Thus, the size and the standard deviation need to be estimated for the Gaussian blurring function, which may be estimated according to the error-parameter curves based on Wiener filtering method [8-9].

Firstly, reflection symmetric extension is performed on the observed image (y) with size of $M\times N$, and the size of the extended image becomes $2M\times 2N$. Then, calculate the Fourier transformation of the extended image (Y). Given a size (K) of the PSF, the error-parameter curves are generated at different standard deviations (σ).

According to the error-parameter curves, the approximate size and standard deviation of the blurring function can be estimated. The size where the distance between the curves decreases greatly is assumed to be the estimated size, and the standard deviation where the corresponding curve increases obviously is assumed to be the estimated standard deviation.

In order to estimate the parameters of Gaussian PSF automatically, two thresholds T_1 and T_2 are set. Firstly, given an estimation error *e*, the curve where once the distance between curves is smaller than T_1 gives out the estimated size ($\hat{K}$) of the Gaussian PSF. The distance is defined as the absolute difference of the cycle number (*j*) of standard deviation at *e*. Then, by calculating the slop of the estimation error at different standard deviations on the estimated curve, the deviation value can be estimated. The deviation once the slop is greater than the threshold T_2 is the estimated deviation ($\hat{\sigma}$).

3 Experiments

3.1 SIMULATE LR IMAGES

Experiments are performed on multiple simulated LR images to test the algorithm objectively and subjectively. To avoid the boundary effect, a zero window with width of 16 pixels is added to the image ‘lena.bmp’ of size 256×256, and the gained simulated HR image of size 288×288 as shown in Fig.3.

The HR image is passed through the LR imaging model as shown in Fig.1. Firstly, the HR image is horizontally and vertically shifted. Here, the horizontal shift $\left(a_i\right)$ and the vertical shift $\left(b_i\right)$ are taken as shown in Tab.1. Secondly, the five moved images are convolved by a Gaussian PSF with size of 7 and standard deviation of 0.8 respectively. Secondly, the blurred image is down-sampled by a factor of 2 in horizontal and vertical direction. Finally, Gaussian noise with a density of 0.05 is added. The generated 5 LR images with size of 144×144 are shown in Fig.4. Take the first LR image as reference image.

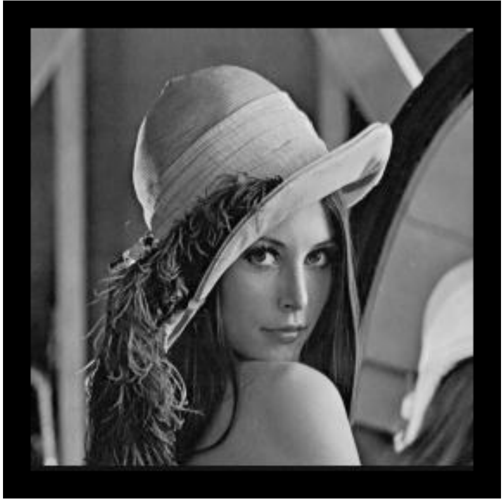


FIGURE 3 The simulated HR image

TABLE 1 The movement parameters

Sequence number (i)	Horizontal shift $\left(a_i\right)$ (in pixel)	Vertical shift $\left(b_i\right)$ (in pixel)
1	0	0
2	2.5467	2.5778
3	-2.8796	3.4356
4	2.5478	-3.4566
5	-2.5656	-3.5436

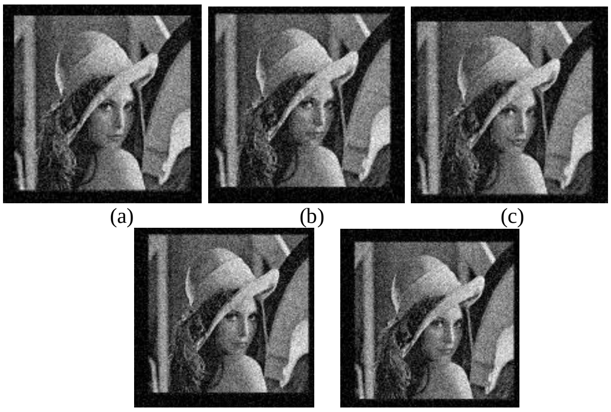


FIGURE 4 The simulated LR images

3.2 IMAGE DE-NOISE

The simulated LR image is de-noised by Wiener filtering algorithm. The de-noised LR images are generated. The bilinear interpolated reference LR image (Fig.4(a)) by 2 times is shown in Fig.5, and the PSNR is 31.6242dB. The bilinear interpolated de-noised reference LR image by 2 times is shown in Fig.6, and the PSNR is 32.8454dB. We can see that the Gaussian noise is reduced and the PSNR is improved.

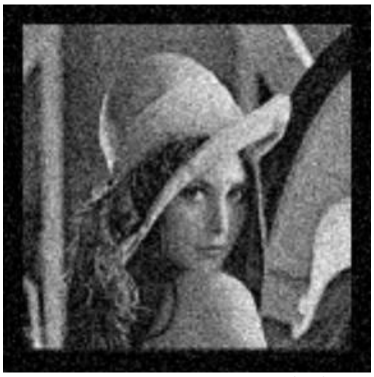


FIGURE 5 The Bilinear interpolated image of reference LR image

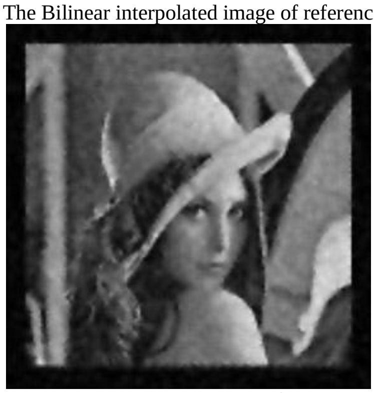


FIGURE 6 The Bilinear interpolated image of de-noised reference LR image

3.3 MOVEMENT ESTIMATION

Relative to the reference de-noised LR image, the estimated movement parameters of the *i*th LR image are denoted as $\hat{a}_i$ and $\hat{b}_i$. The absolute estimation errors are defined Eq.(13), The estimated absolute estimation errors are shown in Tab.2.

$$\Delta a_i = \left|\hat{a}_i - a_i\right|,\ \Delta b_i = \left|\hat{b}_i - b_i\right|, \tag{13}$$

TABLE 2 The estimated absolute estimation errors

Sequence number (*i*)	Horizontal estimation error (Δa_i) (in pixel)	Vertical estimation error (Δb_i) in pixel)
1	0	0
2	0.0094	0.0112
3	0.0024	0.0083
4	0.0015	0.0067
5	0.0024	0.0069

3.4 GAUSSIAN BLUR ESTIMATION

The Gaussian PSF of the de-noised reference LR image is estimated. The sizes (*K*) of the Gaussian PSF are taken as 3, 5, 7, 9 and 11 respectively. The range of the standard deviation (σ) is taken as [0.5, 2]. The searching time is taken as 100. The threshold T1 and T2 are taken as 2 and 0.5 respectively. The generated error-parameter (E-σ) curves of the LR image at different sizes are shown in Fig.7.

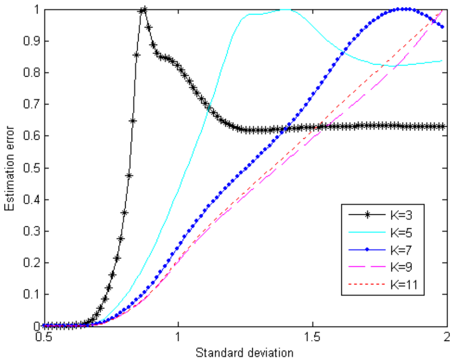


FIGURE 7 The error-parameter curves of the de-noised reference LR image

By analyzing the relationship of the multiple curves, the estimated size ($\hat{K}$) is 7, and the estimated standard deviation ($\hat{\sigma}$) is 0.755. The absolute estimation error of the size and standard deviation are as follows respectively:

$$\left|K_0 - \hat{K}\right| = \left|7-7\right| = 0,\ \left|\sigma_0 - \hat{\sigma}\right| = \left|0.8 - 0.755\right| = 0.005$$

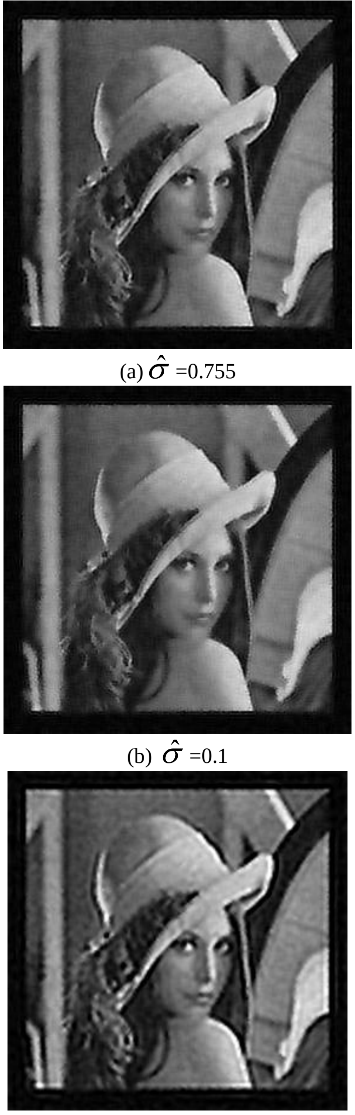
3.5 SUPER RESOLUTION RECONSTRUCTION

Utilizing the estimated movement parameters and Gaussian PSF, super resolution reconstruction is performed on the de-noised LR images through IBP algorithm. When the estimated size and standard deviation of the Gaussian PSF are 7 and 0.755, the SR reconstructed image is shown in Fig.8 (a).

In addition, in order to justify the importance of Gaussian blur estimation in multi-image SR reconstruction, in the case of the estimated size of Gaussian PSF is 7, the estimated standard deviation is taken from 0.1 to 3 with an increment of 0.1, the PSNRs of the estimated SR images are shown in Fig.9. The SR reconstructed images when the standard deviations are 0.1and 3 are shown in Fig.8 (b) and (c) respectively.

Relative to the simulated HR image, the peak signal to noise ratio (PSNR) of the reconstructed image gained by different methods are shown in Tab.3.

The experimental results show the effectiveness of the proposed method. The Gaussian noise is reduced in the SR reconstructed image. The movement parameters and Gaussian PSF are estimated with high accuracy. The SR reconstructed image has better visual effect and higher PSNR than other methods. When the Gaussian PSF is near the real value, the SR reconstructed image has better visual effect and higher PSNR. When the estimated is smaller than the real value, the SR reconstructed image is ambiguous. When the estimated is larger than the real value, the SR reconstructed image has obvious ring effect.



(a) $\hat{\sigma}$ =0.755

(b) $\hat{\sigma}$ =0.1

(c) $\hat{\sigma}$ =3

FIGURE 8 The SR reconstructed images at different estimated standard deviations

TABLE 3 The PSNRs of the images gained by different methods (in dB)

Figure	Fig.5	Fig.6	Fig.8(a)	Fig.8b)	Fig.8(c)
PSNR	31.6242	32.8454	33.7324	33.5085	32.8253

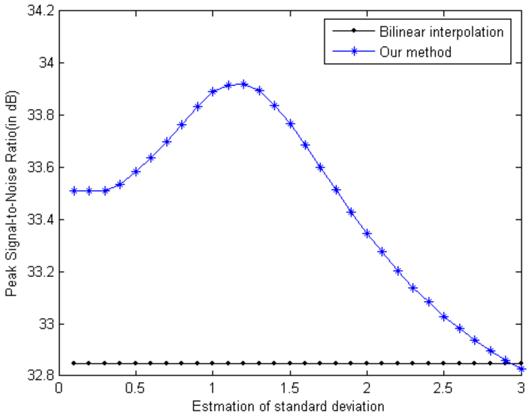


FIGURE 9 The PSNR of the SR reconstructed image at different estimated $\hat{\sigma}$

4 Conclusion

A framework of blind multi-image SR reconstruction with Gaussian blur and Gaussian noise is proposed. The degrading processes of movement, Gaussian blur, down sample and Gaussian noise are all considered in the LR imaging model. The simulated LR images are de-noised by Wiener filtering algorithm. The horizontal shift and the vertical shift between the de-noised LR images are estimated. The Gaussian blur of the de-noised LR image is estimated through error-parameter analysis method. The SR image is reconstructed through IBP algorithm. The experimental results justified the effectiveness of the proposed method. The Gaussian noise is well restrained in the SR reconstructed image. The visual effect and PSNR of the SR reconstructed image are improved. The proposed framework may be widely applied in other SR image reconstruction cases, such as different movement, different types of blur, different noise, and so on.

Acknowledgements

This work is supported in part by the National Nature Science Foundation of China (Gant no.61202195), the Sichuan Provincial Education Department project (Gant no.11ZA174), the Application Fundamental Research Project of Sichuan Provincial Scientific and Technology Department (Gant no.2011JY0139), the key project of Yibin Science and Technology Bureau (Gant no.2011SF016, 2013ZSF009).

References

[1] Zhang X H, Tang M, Tong R F 2012 Robust super resolution of compressed video *Visual Computer* **28** 1167-80

[2] Garcia D C 2013 Super resolution for multiview images using depth information *IEEE Transactions on Circuits and System for Video Technology* **22** 1249-56

[3] Zhang L, Yuan Q Q, Shen H F, Li P X 2011 Multiframe image super-resolution adapted with local spatial information *J. Opt. Soc. Am. A* **28**(3) 381-90

[4] Giannoula A 2011 Classification-Based Adaptive Filtering for Multiframe Blind Image Restoration *IEEE Transactions on Image Processing* **20** 382-90

[5] Yang S Y, Wang M, Chen Y G, Sun Y X 2012 Single-image super-resolution Reconstruction via learned geometric dictionaries and clusters sparse coding *IEEE Transactions on Image Processing* **21**(9) 4016-28

[6] Kim K I, Kwon Y 2010 Single-image super-resolution using sparse regression and natural image prior *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(3) 1127-33

[7] Rueda A, Malpica N, Romero E 2013 Single-image super resolution of brain MR images using overcomplete dictionaries *Medical Image Analysis* **17** 113-32

[8] Zou M Y 2004 Deconvolution and signal recovery *Beijing:Defense Industry Publishing*

[9] Qin F Q 2010 Blind image super-resolution reconstruction based on PSF estimation, *IEEE International Conference on Information and Automation* **II** 1200-03

[10] Qiao J P Liu J, Cheng Y W 2007 Joint blind super-resolution and shadow removing *IEICE Trans. INF. & Syst.* **E90-D**(12) 2060-69

Authors

Fengqing Qin born in 1976, in Nanchong, Sichuan, China

Current position, grades: assistant researcher and vice director in the college of computer and information engineering, Yibin University
University studies: BS degree in computer science and technology from Sichuan Technique University, Zigong, Sichuan, China, 1999, MS degree and PhD degree in communication and information system from Sichuan University, Chengdu, Sichuan, China, in 2006 and 2009
Scientific interest: is image and video processing

A 'Follow-Me' computing scheme based on virtual machine movement for QoS improvement in mobile cloud computing environments

Xu Gaochao[1], Ding Yan[1], Ou Shumao[2], Hu Liang[1], Zhao Jia[3]*

[1] *College of Computer Science and Technology, Jilin University, Qianjin Str. 2699, 130012 Changchun, China*

[2] *Department of Computing and Communication Technologies, Oxford Brookes University, Wheatley, Oxford OX33 1HX, United Kingdom*

[3] *College of Computer Science and Engineering, ChangChun University of Technology, Yan'an Str. 2055, 130012 Changchun, China*

Received 1 March 2014, www.tsi.lv

Abstract

Mobile cloud computing utilizes virtualized cloud computing technologies in the mobile Internet. To improve Quality of Service (QoS) and execution efficiency of mobile cloud applications, we propose a novel computing scheme called "Follow-Me" (FM), which is based on live wide-area virtual machine (VM) migration. In a virtualized mobile cloud environment based on the VMs of cloud side and mobile devices of user side, the purpose of the proposed FM scheme is to migrate the corresponding VM in real-time when a mobile device moves from one service area to another. FM obtains the current positions of mobile devices, estimates the next servicing areas, and finally migrates the VMs along with the mobile users' movement. The proposed FM scheme has been tested in an experimental environment by using the CloudSim platform. The experimental results demonstrate that FM evidently improves the QoS of mobile cloud computing compared with the existing approaches. FM achieves a better average service response time, a clearly smaller error rate and consumes less energy.

Keywords: Mobile Cloud Computing, Mobile Device, Virtual Machine, Area Localization, Live Wide-Area Migration

1 Introduction

Cloud computing [1, 2] is the development of parallel computing, distributed computing and grid computing. It distributes the computation tasks into a resource pool composed by great amounts of computing resources, and makes users on-demand obtainable computing power, storage space and information services. With the vigorous development of the mobile Internet, the cloud computing services based on mobile devices such as mobile phones, tablet PCs, etc. have emerged and been widely used for the fields of information sharing, mobile learning, e-commerce, home monitoring and mobile health etc. Mobile cloud computing is not only a kind of IT resources but also a delivery and utilization mode of information and services. It obtains the required infrastructures, platforms and software (applications) etc. in an on-demand and scalable manner through the underlying mobile Internet. Mobile cloud computing applies cloud computing technology into mobile Internet. Mobile devices are limited to its battery capacity and computing power, which has been always hampering the development of mobile computing. However, in mobile cloud computing system, mobile applications can be partitioned into several computation modules or components. The complicated computations of a mobile application will be performed on a VM, which is located in a cloud data center. It is envisioned that mobile cloud computing will play a crucial role in people's daily life in the near future. Figure 1 shows the differences between single-machine computing and mobile cloud computing where an application is partitioned and some parts are executed in a VM at a remote data center. The distributed execution at mobile devices and mobile clouds can be done automatically or semi-automatically.

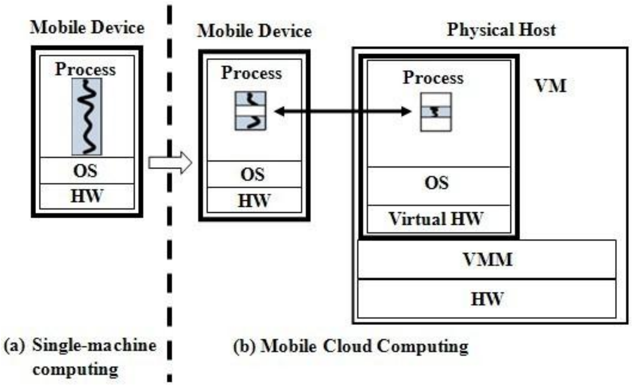


FIGURE 1 The system model of transforming a single-machine execution (mobile device computing) into a distributed execution (mobile device and cloud computing) (semi)-automatically

* ***Corresponding author*** e-mail: zhaiyj049@sina.com

The performance of mobile applications running on mobile cloud computing environments is mainly affected by three aspects: the computation in mobile device side, the computation in the corresponding VM side and the communication (mainly the wireless part) between the two sides. Since a cloud data center has large amount of resources, generally the computation cost of a mobile application in the corresponding VM is very little in comparison with it on the mobile device and normally negligible. The computation performance of mobile device side can also be improved by well-designed partition approaches, such as [3-5]. Due to the unreliability nature of the wireless communication, the mobile wireless network becomes the bottleneck of the performance of mobile cloud computing. To relieve this bottleneck, the transmission distance between a mobile device and its corresponding VM should be kept as short as possible. As shown in Figure 2, Mobile Device A in Area A is served by Cloud Data Center A, i.e. the corresponding VM of Mobile Device A in Cloud Data Center A is communicating with Mobile Device A. If Mobile Device A moves into Area B, which is far away from Area A, the wireless distance, for which it communicates with its VM located in Cloud Data Center A will become longer or even unreachable. This will seriously affect the QoS of mobile cloud applications and mobile users will take no advantage from mobile cloud computing. Assume that there are Cloud Data Center B and Access Point B in the nearby areas, and which are from the same provider. If the VM of mobile device A in Cloud Data Center A moves into Area B, it will be able to communicate with its VM locally by Access Point B and thus the reliability of the communication and to the QoS of the mobile cloud computing services can be much improved. To address this problem and achieve this goal, we propose a novel computing scheme called "Follow Me" (FM). It has abilities in efficiently moving the VM into the nearest cloud data center along with the mobile device's movement. It can minimize the communication cost and greatly improve the QoS and finally provide better and more efficient mobile cloud computing services.

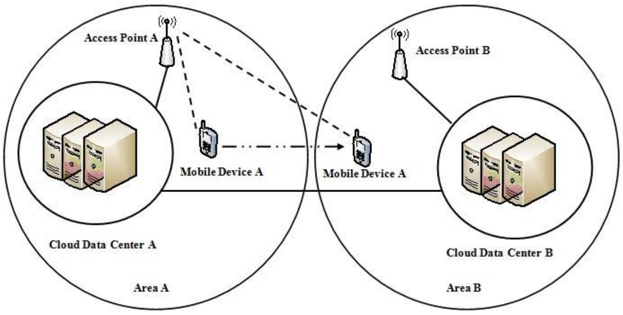


FIGURE 2 Traditional model of mobile cloud computing services

The rest of the paper is organized as follows. In Section 2, we present the related work. The reasonable prerequisites, the algorithm and implementation of the FM scheme are discussed in detail in Section 3. Section 4 reports the experimental results and analysis on CloudSim platform. Section 5 concludes the paper and lists our future work.

2 Related work

To the authors' best knowledge, very little research work on the problem of moving the VM of a mobile device from a source cloud data center to a approaching target cloud data center has been done. . Most related research is focused on some similar problems such as improving the performance and QoS of mobile cloud computing.

The applications in mobile cloud computing systems run diverse workloads under diverse device platforms, networks and clouds. Traditionally these applications are statically partitioned between less-powerful devices and powerful clouds, thus their execution may be significantly inefficient in heterogeneous environments and with different workloads. Byung-Gon and Petros in [4] proposed an approach of dynamic partitioning of applications between weak devices and clouds and argued that dynamic partitioning is the key to addressing heterogeneity problems. The authors found that most of the mobile applications running on mobile devices can be restructured so that they can be statically partitioned between the weak device and a server running in the cloud.

Jung *et al.* in [6] proposed to exploit the potential of smart phones in proximity cooperatively, using their resources to reduce the demand on the cellular infrastructure, through a decision framework called RACE (Resource Aware Collaborative Execution). RACE enables the use of other mobile devices in the proximity as mobile data relays. RACE is a Markov Decision Process (MDP) optimization framework that takes user profiles and user preferences to determine the degree of collaboration. Both centralized and decentralized policies are developed and validated through simulation using real mobile usage traces.

In [7], Klein *et al.* proposed an architecture to provide an intelligent network access strategy for mobile users to meet the application requirements. The authors proposed a so called Context Management Architecture (CMA) which is responsible for acquiring, processing, managing, and delivering context information. Finally, the authors presented a context-aware radio network simulator (CORAS) that was able to model context availability, accuracy, and delay, thus enabling an evaluation of the impact of different levels of context relevance, confidence, and quality on simulation results.

Zhang *et al.* in [8] designed and constructed a multi-hop networking system named MoNet based on WiFi, and on top of which they designed and implemented WiFace, a privacy-aware geosocial networking service. For the situation without any infrastructure, they designed a distributed content sharing protocol, which can significantly shorten the relay path, reduce conflicts and improve data persistence and availability. A role strategy was designed to encourage users to collaborate in the network. Furthermore, a key management and an authorization mechanism were developed to prevent some attacks and protect privacy.

Kovachev *et al*. in [9] proposed a Mobile Community Cloud Platform (MCCP) as a cloud computing system that can leverage the full potential of mobile community growth. An analysis of the core requirements of common mobile communities is provided before they present the design of their cloud computing architecture that supports building and evolving of mobile communities.

Liang *et al*. in [10] proposed a Security Service Admission Model (SSAM) based on Semi-Markov Decision Process to model the system reward for the cloud provider. They first define system states by a tuple represented by the numbers of cloud users and their associated security service categories, and current event type (i.e., arrival or departure). They also derived the system steady-state probability and service request blocking probability by using the proposed SSAM. The approach provided strong security protection while achieving resource management for the maximum revenue.

3 The proposed FM scheme

3.1 PREREQUISITES

In this paper, we assume that all the target areas, which a mobile device may leave for have mobile cloud data centres owned by the same provider. The storages of all cloud data centres of a provider are distributed and shared [11]. This assumption can be easily relieved for the service areas, which belong to different providers, if certain level of trust can be built between the providers.

In the proposed FM approach, the disk-image of a migrated VM will also be migrated for the reason of efficient access and better QoS performance. In another word, the migration of a VM consists of the memory of VM, its running-time status [12] (including CPU, registers, I/O states), VM disk-image and the related information for network recovery.

We further assume that the network connections between the mobile cloud data centres are wired, high-speed and reliable (e.g. by optical fibers). In addition, we assume the mobile applications can tolerant short period of inaccessibility during VM migration. Like other research work, e.g. [11, 12], these assumptions are believed not to affect to the performance and efficiency of the proposed FM scheme.

3.2 THE FM SCHEME

The proposed FM Scheme is designed to improve the performance and efficiency of mobile cloud computing. Actually, FM is a placement selection policy of mobile devices' VMs and it is based on area localization and live wide-area VM migration, which involves the migration of memory and status, the memory of VM disk-image and the redirection of wide-area network. More Specifically, FM firstly performs a decision of area location of the mobile device according to some existing method [13]; secondly, it calculates the two distances from a mobile device to the cloud data centre, which its VM is located in and to the cloud data centre, which mobile device is approaching to; thirdly, FM compares the two distances. If the former is larger, FM does not do anything. Once the latter is larger, FM will perform a live wide-area VM migration of the corresponding VM located in source cloud data centre. To achieve the efficient live wide-area VM migration proposed in FM, we have employed our previously proposed live VM migration mechanism HMDC used for live local-area VM runtime states migration [14] as well as we have combined the composed image cloning (CIC) methodology and wide-area network redirection methodology with the HMDC approach to address the problems of wide-area VM migration and thus to achieve the fast and efficient live wide-area VM migration mechanism used in the proposed FM approach. After the VM has migrated and resumed running in the target cloud data centre, the mobile device will begin to be connected to the wireless access point located in target area and communicate with the target VM located in target cloud data centre, as show in Figure 3; finally, source VM is deleted. FM keeps on monitoring the signals from all mobile devices and repeats this process.

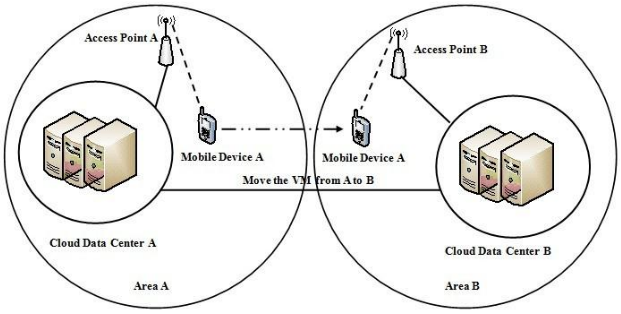


FIGURE 3 FM model of mobile cloud computing services

3.3 IMPLEMENTATION

The detailed processes of the FM scheme are described as follows:

Target area localization of mobile device A: mobile device A is moving from area A to target area B. Once the mobile device is connected to wireless access point A, the position of mobile device A will be obtained by FM. This can be done by searching a pre-defined table, for instance. The part is not the focus of this paper.

Distance Calculation: as show in Figure 4, after FM has obtained the position of mobile device A continuously, it would find that mobile device A is getting closer to cloud data centre B. FM calculates the distance l from mobile device A to cloud data centre B and the distance *h* from mobile device A to cloud data centre A which its VM is running in continuously. As long as it finds that *l* is smaller than *h*, FM will perform the follow steps. If *l* is bigger than *h*, FM will continue to monitor and calculate. Furthermore, if there are more cloud data centres available in the vicinity of mobile device A, FM will calculate the distance between the mobile device and each of the areas and once there is one distance being gradually smaller than the distance from mobile device A to current cloud data centre, FM will move the VM to that cloud data centre.

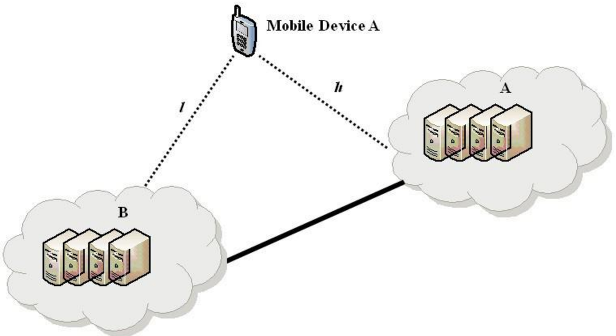


FIGURE 4 An example of mobile cloud computing model

Live wide-area VM migration: once the FM approach determines that a VM needs to be migrated, it begins to perform the live wide-area migration process of the VM. The live wide-area VM migration includes three main tasks: runtime states (memory and status) migration, VM disk-image migration and wide-area network recovery. The proposed FM approach assumes each VM is composed of two elements: the composed VM disk-image and the user work directory, both stored inside a NAS of the cloud where the VM is hosted [15]. Each cloud data centre holds its own VM components repository and NAS. The repository contains a set of VM components, whereas the NAS contains the storage of each VM (the composable and user data blocks). As mentioned above, cloud data centres are connected with each other through high-speed wired networks. When FM needs to migrate a VM, first of all it will send information to target cloud data centre in order to tell it which components the migrant VM is formed by. The target cloud checks in its own repository and notifies source cloud which components it does not hold. Thus, target cloud copies the components from the repository of source cloud, once it holds all the required VM components, it locally clones the composed VM disk-image of the migrant VM. Then, source cloud begins to migrate the memory and status with the proposed HMDC approach. At this point, we assume that target cloud has determined the target host by a specific placement selection policy.

Source host utilizes dynamic ballooning mechanism [16] to recycle idle pages of source VM to reduce total data transmitted. Source host opens up memory cache of source VM. The total memory is transmitted to target host with source VM running. During this process, if a page is to be overwritten, its original data will be copied to memory cache at first. All pages dirtied are marked to dirty_page bitmap and idle dirty pages are marked using a special mark during total memory copy. While total memory copy is completed, the source VM stops running.

After source VM stops running, source host looks up cache blocks of dirty pages marked in dirty_page bitmap. If the old version of a dirty page is cached in memory cache, the flag bit of the memory page is denoted by "1". If not, is denoted by "0". After all dirty pages have been checked, source host generates a new bitmap called cache_bitmap, which marks all dirty pages whose old versions are cached in source host. Then source host copies the two bitmaps to target VM. As illustrated in Fig.5, after target VM receives them, according to the number of dirty pages marked in cache_bitmap, target host opens up memory cache and caches the corresponding memory pages to memory cache. At the same time, an AMT (address mapping table) to maintain the cache mapping is created. Finally, it sets the lowest EPT (Extending Page Table) items of all dirty pages to "non-present" according to dirty_page bitmap. Memory cache of source and target will be recycled once its data has been used so that the approach does not generate extra space overhead. During source VM's stopping running, an IP tunnel [17] between the old IP address at the source and its new IP at the target will also be set with the help of *iproute2*. In addition, the other runtime data including CPU status and I/O states etc. will be transmitted to target host in the process.

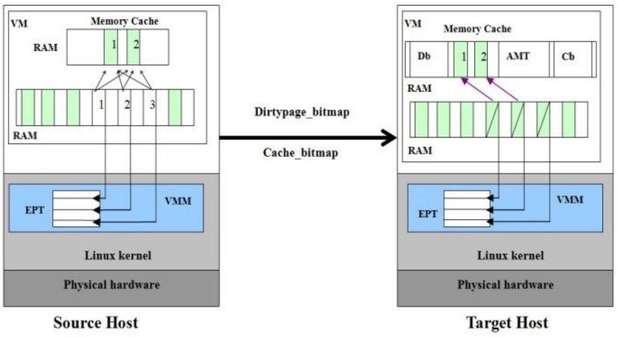


FIGURE 5 The process of bitmaps copy.

Target VM resumes running and source host begins pushing dirty pages periodically. The algorithm sets a timer. If the timer times out, according to dirtypage_bitmap, HMDC copies non-idle dirty pages to a pushing queue until it is full or dirty pages are exhausted. After a process of active push is completed, the timer restarts. If during timing receiving a page request, source host immediately suspends the timer and copies the page requested to the pushing queue. Subsequently according to the principle of locality, also copies its left and right neighbour dirty pages to the pushing queue until it is full or the dirty pages are exhausted. The timer resumes timing after sending the queue. If the dirty pages to be transmitted have the cache of old versions, the algorithm performs delta compression on the dirty pages. It firstly computes the delta page by applying XOR on the current and previous version of a page, and then get the delta compression page by compressing the delta page using XOR binary RLE algorithm [18]. Finally, a delta compression flag is set in the page header. Source host replaces dirty pages with their delta compression pages and sends the queue to target host. Finally, the corresponding cache blocks are recycled by VMM (virtual machine monitor). If the old versions of pages do not exist in memory cache, source host directly transmits dirty pages to target host. Delta compression should consume a minimum of CPU resources, both for cache hits and cache misses to not slow down the migration process or disrupt the performance of VM. Accordingly, a 2-way set associative caching scheme is employed.

At the same time when target VM resumes running, target host begins performing demand paging. While the pages marked in dirtypage_bitmap are accessed, memory access faults will occur and then be fallen into VMM kernel to be captured by the algorithm. If the page is not an idle dirty page, with target VM suspended a page request will be sent to source host. Source host transmits a set of pages, which include the requested page to target host as its response. Target host receives the response and updates memory pages correspondingly. Then target VM resumes running. If the requested page is an idle dirty page, target host does not send the request to source host but directly allocates a memory page to target VM from the local. Subsequently target VM immediately resumes running. During target VM running, target VM will receive dirty pages from source VM periodically. For target VM, both the response of a dirty page request and the active push of source host are checked whether they have delta compression flags. If yes, the algorithm firstly decompresses the pages to get delta pages and then rebuilds the new pages by applying XOR on the delta pages and the old versions cached in target host. The algorithm updates the memory pages using the new rebuilt pages and recycles the cache blocks to VMM. If no, the algorithm directly updates the pages. According to dirtypage_bitmap, while all dirty pages have been synchronized, memory migration will ends. Furthermore, after target VM resumes running, the algorithm begins forwarding all pockets, which arrive at source VM for the VM's old IP address to the target through the IP tunnel. In target host, the target VM has two IP addresses: its old one, used by existing connections through the tunnel, and its new one, used by new connections directly. The IP tunnel is torn down when no connections remain use the VM's old IP address. When the migration has completed and the VM can respond at its new network location, the dynamic DNS entry [19] for the services the VM provides will be updated in order to ensure that future connections are directed to the VM's new IP address [20]. It is notable that the packets, which arrive during VM downtime have to either be dropped or queued. The proposed FM approach will utilize *iptables* to drop them in order to avoid connections reset.

Although the proposed FM approach has exploited composed image cloning methodology to reduce the size of a VM disk-image and thus to make the algorithm need only to transmit zero or several small components through network before live migration of runtime states, the user data blocks which may be accessed at any time during VM running has not yet been migrated to target host. To address this problem of live migration of the user data blocks (the user wok directory), we have introduced a storage access mechanism into the proposed FM approach. It is composed of two main objects: a storage server and a proxy server of a block-level storage I/O protocol (e.g. iSCSI and NBD). FM has employed the NBD protocol due to its simple and plain. Source and target host nodes are connected to the storage and proxy server, respectively, by the NBD protocol, using TCP/IP. A VM accesses virtual disks via block device files (e.g. /dev/nbd0) on a host operating system. Before live migration, it works in the same manner as a normal NBD storage server, which redirects I/O requests from the VM to a disk image file. After FM starts live migration, the mechanism works together with memory migration. Once the target VM resumes running in target host during the live migration of memory, all I/O is performed at the target via the proxy server. When a user data block is accessed and it is not cached to the proxy server, it redirects the I/O requests to the storage server at the source and also caches the disk blocks to a local file at the proxy server node. This is similar to the demand paging of memory migration. The proxy server keeps on remote block copies through an NBD connection until all user data blocks are cached at the target. The background copying mechanism that copies the rest of the in-use blocks which still remain at the source is the same as the active push mechanism of memory migration. Afterwards, the proxy server terminates the NBD connection and the VM does not have to rely on the storage server at the source. It continues to work in the same way before migration [21].

After the downloading of small components, the live migration of runtime states, the redirection of wide-area network and the migration of the user data blocks have been completed, the whole live wide-area VM migration will be completed. When mobile device A needs mobile cloud computing services and communicate with its VM, it will be connected to access point B and interact with the target VM. The target VM provides mobile cloud computing services to mobile device A. The source VM is then deleted. FM continues to monitor all mobile devices and repeat the above algorithm process.

4 Evaluation

4.1 EXPERIMENTAL SCENARIOS

In this section, we will experimentally verify the proposed FM scheme. We conduct several simulations to study the performance and efficiency of the proposed FM scheme for improving QoS of mobile cloud computing services relying on effective VM migration. We implemented the proposed FM scheme and its algorithm by using CloudSim 2.0 [22]. CloudSim is an extensible simulation toolkit that enables modeling and simulation of cloud computing environment and supports modeling and creation of VMs on a simulated node of a data centre, etc [23-24]. In our experiments, the live wide-area migration process is ignored and the default migration mechanism in CloudSim is employed so it will not affect the comparison of the experimental results. The effectiveness and the efficiency of the proposed FM scheme are evaluated mainly on the average service response time (delay time), energy consumption and error rate, via comparing with the traditional models of mobile cloud computing services.

On CloudSim platform, we create two cloud data centres A and B. Each cloud data centre is composed of five hosts. There are two VMs running in each host at the beginning. Besides, we create an independent host A running Android OS to simulate our mobile device A. The host A can be regarded as mobile device A by adjusting its communication bandwidth with cloud data centres A and B. The communication bandwidth is larger and the distance from mobile device A to the cloud data centre is shorter. The mentioned bandwidth refers to available wireless bandwidth. With the distance gradually increasing, the available wireless bandwidth will decrease since the link path is longer and the loss of network signals is higher as well as the probability with which the bottleneck link is encountered is higher. In our simulation, we use the smooth curve of inverse proportion function to measure the relationship between the bandwidth and the distance.

$$Bandwidth = \frac{k}{Dis\tan ce}, \quad (Dis\tan ce > 0), \qquad (1)$$

where k is a positive degree coefficient. The cloud data centre A, cloud data centre B and host A have a coordinate attribute (i.e. (x, y)) respectively as shown in Table I to denote their positions.

TABLE 1 Position coordinates

Cloud data centre A	(0, 0)
Cloud data centre B	(x_1, y_1)
Host A	(x_2, y_2)

The position of cloud data centre A is set as the origin of a rectangular coordinate system. The position of cloud data centre B is initialized at random. The position of host A can be updated randomly and it is limited within the rectangle formed by cloud data centre A and cloud data centre B, as shown in Figure 6.

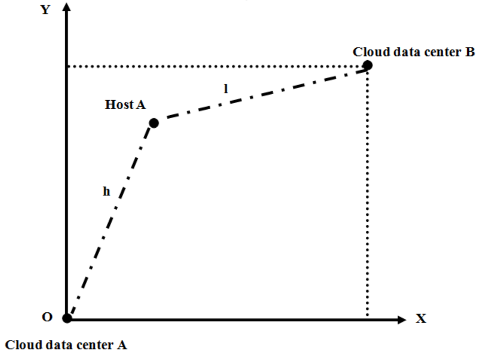


FIGURE 6 Positions in the rectangular coordinate system

In cloud data centre A, we set a VM as the source VM of mobile device A. It uses Android OS and executes a simulated mobile cloud service. The host A sends the service request to the VM periodically. After receiving a request, the VM will respond the request and send a response data to the host A. In CloudSim platform, we have implemented the simulation process of the proposed FM algorithm and processes. The host A sends simulated signals to the FM process periodically to make FM obtain its position. In addition, the traditional model is also implemented through the absence of the VM migration process.

4.2 EXPERIMENTAL RESULTS AND ANALYSIS

In the first set of experiments, we have verified the feasibility, effectiveness and efficiency of the proposed FM approach by comparing with the traditional approach on the average service response time (i.e. delay time). As shown in Figure 7, we can find that the average service response time of the proposed FM approach is shorter than that of the traditional approach all the way. The reason for the experimental result is that the proposed FM approach has abilities in making the VM providing mobile cloud computing services follow its mobile device. This will reduce intermediate links and indirectly increase the effective available bandwidth and throughput of their interaction with each other as well as to thus achieve a better average service response time finally.

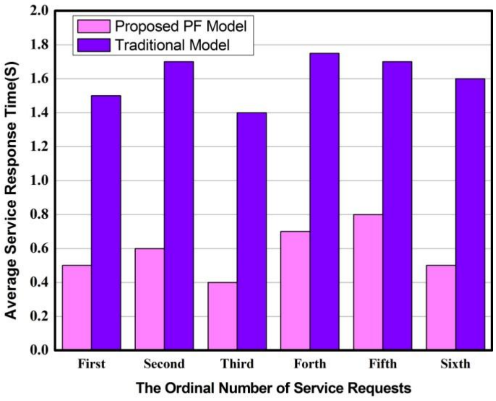


FIGURE 7 Comparison of average service response time

In the second set of experiments, we have compared the proposed FM approach with the traditional approach on the error rate of network communication. We analyze the proposed FM approach and the traditional approach by the statistics of error rate of every day on the communication data between mobile device (the host A) and its VM. Figure 8 shows the comparison of the traditional approach and FM approach in error rate. The experimental result indicates that compared with the traditional approach, FM has achieved a clearly smaller error rate. And as the time increases, the error rate of FM is more stable. This is because a better available network bandwidth, a better communications link and the absence of intermediate link will lead to a better communication quality and make the errors not easy to produce.

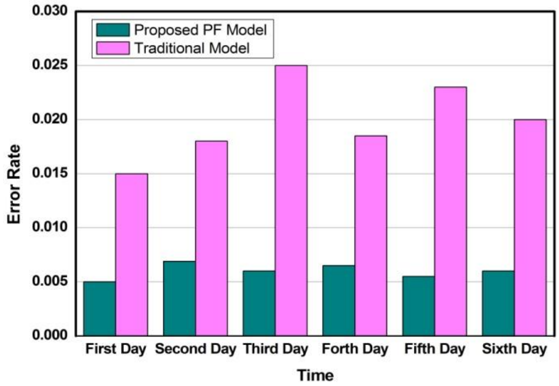


FIGURE 8 Comparison of error rate

In the third set of experiments, we have verified the proposed FM approach by evaluation of energy consumption of the whole system. The energy consumption of the two systems implementing the FM approach and traditional approach is compared to test the performance and efficiency of the FM approach. We analyze the two approaches by statistics of energy consumption of each hour in the whole system. The experimental result is as shown in Figure 9.

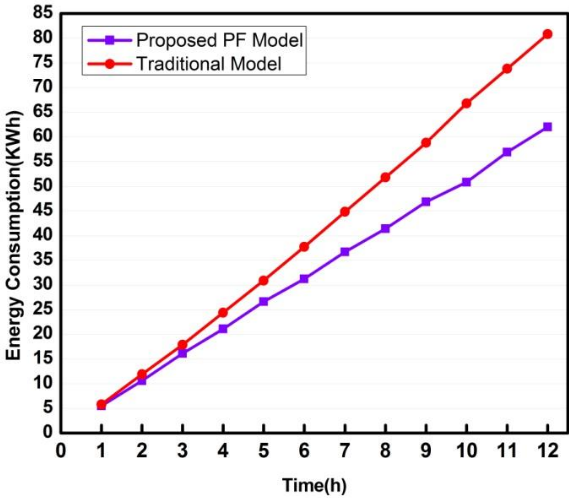


FIGURE 9 Comparison of energy consumption

It indicates that compared with the traditional approach, FM has a less energy consumption. In addition, as the time increases, the incremental energy consumption of FM is less than that of the traditional approach. The reason for the experimental result is that FM makes the interaction between mobile device and its VM not need more transmission of intermediate nodes. In other words, the FM approach has made the mobile device's communication with its VM achieved in a relatively shortest link. As a result, it consumes much less energy.

5 Conclusion and future work

In this paper, we have proposed an efficient "Follow Me" mobile computing scheme FM, which selects VM placements dynamically and migrate VMs effectively. FM scheme is designed for improving QoS and efficiency based on live wide-area VM migration in mobile cloud computing environments. Based on the mobile network features, in particular network bandwidth and network delay, the FM algorithm takes into account the distance between mobile device and its VMs. When a mobile device moves to other areas, the corresponding VM providing services to the mobile device will also be dynamically moved to that area in order to obtain better QoS. FM has achieved the live wide-area migration of VM based on obtaining the position of mobile device in real time and thus to make that the VM follows its mobile user come true. To achieve a 'Follow-Me' policy and the live wide-area VM migration based on it, FM has not only exploited the wireless positioning technology and proposed the method of comparing the distances to achieve the trigger mechanism of migration, but also combined live local-area migration mechanism HMDC with the composed image cloning (CIC) methodology and wide-area network redirection methodology to achieve the efficient live wide-area VM migration mechanism. The final experimental results demonstrate that FM evidently improves the performance, efficiency and QoS of mobile cloud computing with the user moving compared with the traditional approach. FM has the better average response time and error rate as well as has a less energy consumption. It makes the result of mobile cloud computing higher effective and more meaningful.

Aiming to further improve the performance of FM, we plan to study the robustness of FM in the future. In the cases of power outage or server crash, etc., FM should have the abilities of recovering the original VM and run-time data. As a future work, we will implement a prototype of the FM scheme to further study the energy efficiency aspect and the throughput of a FM scheme implemented mobile cloud network.

Acknowledgments

This work was supported by European Commission the 7th Framework Programme MONICA project under Grant No. PIRSES-GA-2011-295222.

References

[1] Gaochao Xu, Yan Ding, Liang Hu, Xiaodong Fu, Jia Zhao, Hao Yan, Jianfeng Chu 2013 *Journal of Convergence Information Technology* **8**(8) 341-8

[2] Armbrust M, Fox A, Griffith R, Joseph A, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, Zaharia M 2009 *Communications of the ACM* **53**(4) 50-8

[3] Byung-Gon Chun, Sunghwan Ihm, Petros Maniatis, Mayur Naik, Ashwin Patti 2011 *Proc. of the sixth conference on Computer systems* ACM New York: Salzburg 301-4

[4] Byung-Gon Chun, Petros Maniatis 2010 Dynamically partitioning applications between weak devices and clouds *Proc. of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond* ACM New York: San Francisco

[5] Lei Yang, Jiannong Cao, Yin Yuan, Tao Li, Andy Han, Alvin Chan 2013 *Performance Evaluation Review* **40**(4) 23-32

[6] Eric Jung, Yichuan Wang, Iuri Prilepov, Frank Maker, Xin Liu, Venkatesh Akella 2010 User-profile-driven collaborative bandwidth sharing on mobile phones *Proc. of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond* ACM New York: San Francisco

[7] Klein A, Mannweiler C, Schneider J, Hans D 2010 Access schemes for mobile cloud computing *Proc. of the 11th International Conference on Mobile Data Management (MDM)* IEEE Computer society: Kansas City 387

[8] Lan Zhang, Xuan Ding, Zhiguo Wan, Ming Gu, Xiang-Yang Li 2010 WiFace: a secure geosocial networking system using WiFi-based multihop MANET *Proc. of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond* ACM New York: San Francisco

[9] Kovachev D, Renzel D, Klamma R, Yiwei Cao 2010 Mobile Community Cloud Computing: Emerges and Evolves *Proc. 2010 Eleventh International Conference on Mobile Data Management (MDM)* IEEE Computer Society: Kansas City 393-5

[10] Liang H, Huang D, Cai L X, Shen X, Peng D 2011 Resource allocation for security services in mobile cloud computing *Proc. of Computer Communications Workshops (IEEE Conference on INFOCOM WKSHPS)* IEEE: Shanghai 191-5

[11] Jia Zhao, Liang Hu, Gaochao Xu, Yan Ding, Jianfeng Chu 2013 A survey on green computing based on cloud environment *International Journal of Online Engineering* **9**(3) 27-33

[12] Liang Hu, Jia Zhao, Gaochao Xu, Yan Ding, Jianfeng Chu 2012 A Survey on Data Migration Management in Cloud Environment *Journal of Digital Information Management* **10**(5) 324-31

[13] Cheng C, Jain R, van den Berg E 2003 Location prediction algorithms for mobile wireless systems *Wireless internet handbook* ed Borko Furht and Mohammad Ilyas: Boca Raton 245-63

[14] Liang Hu, Jia Zhao, Gaochao Xu, Yan Ding, Jianfeng Chu 2013 HMDC: Live Virtual Machine Migration Based on Hybrid Memory Copy and Delta Compression *Applied Mathematics & Information Sciences* **7**(2L) 639-46

[15] Celesti A, Tusa F, Villari M, Puliafito A 2010 Improving Virtual Machine Migration in Federated Cloud Environments *Proc. of 2010 Second International Conference on Evolving Internet* CPS: Valencia 61-7

[16] Hines M R, Gopalan K 2009 Post-Copy Based Live Virtual Machine Migration Using Adaptive Pre-Paging and Dynamic Self-Ballooning *Proc. of the 2009 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments* ACM New York: Washington 51-60

[17] Perkins C 2003 IP encapsulation within IP *RFC 2003*

[18] Pountain D 1987 Run-length encoding *Byte*, 317-319

[19] Wellington B Secure DNS Dynamic Update *RFC 3007*

[20] Bradford R, Kotsovinos E, Feldmann A, Schioberg H 2007 Live wide-area migration of virtual machines including local persistent state *Proc. of the 3rd international conference on Virtual execution environments* ACM: San Diego 169-79

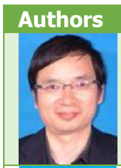
[21] Takahiro Hirofuchi, Hirotaka Ogawa, Hidemoto Nakada, Satoshi Itoh and Satoshi Sekiguchi 2009 A Live Storage Migration Mechanism over WAN for Relocatable Virtual Machine Services on Clouds *Proc. of 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid* IEEE Computer Society: Washington 460-5

[22] CloudSim 2.0 http://www.cloudbus.org/cloudsim/ 2013

[23] Calheiros R N, Ranjan R, Beloglazov A, De Rose C A F, Buyya R 2011 CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms *Software: Practice and Experience* **41**(1) 23 – 50

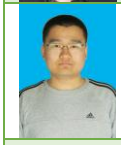
[24] Buyya R, Ranjan R, Calheiros R N 2009 Modeling and simulation of scalable cloud computing environments and the CloudSim toolkit: challenges and opportunities *Proc. International Conference on High Performance Computing & Simulation* IEEE: Leipzig 1 – 11

Authors



Gaochao Xu, born in 1966, Xiaogan City, Hubei Province, China

Current position, grades: Professor and PhD supervisor of College of Computer Science and Technology, Jilin University, China.
University studies: Computer Science and Technology
Scientific interest: Distributed System, Grid Computing, Cloud Computing, Internet Things, etc.
Publications: 55



Yan Ding, born in 1988, China

Current position, grades: Changchun, Master
University studies: bachelor degree at Jilin University in 2011
Scientific interest: Virtualization, Cloud Computing, Mobile Cloud Computing
Publications: SCI 1

Shumao Ou, Changsha

Current position, grades: a senior lecturer in the Department of Mechanical Engineering and Mathematical Sciences at Oxford Brookes University, Professor
University studies: PhD degree in Electronic Systems Engineering and MSc degree (with distinction) in Computer Information Networks from the Department of Electronic Systems Engineering, University of Essex, Colchester, United Kingdom, in 2004 and 2007 respectively
Scientific interest: Wireless Networks, Vehicular Communications networks
Publications: SCI 12
Experience: Dr Shumao Ou is a member of IEEE. From the end of 2006, he worked in the School of Computer Science and Electronic Engineering, University of Essex as a Senior Research Officer. From June 2009, he took a lectureship in the School of Technology at Oxford Brookes University.

Liang Hu, born in 1968, Changchun

Current position, grades: Professor PhD supervisor of College of Computer Science and Technology, Jilin University, China
University studies: BS degree on Computer Systems Harbin Institute of Technology in 1993 and his PhD on Computer Software and Theory in 1999
Scientific interest: Network Security, Computer Network, Cloud Computing
Publications: SCI 12
Experience: As a person in charge or a principal participant, Dr Liang Hu has finished more than 20 national, provincial and ministerial level research projects of China.

Jia Zhao, born in 1982, Changchun City, Jilin Province, China

Current position, grades: Doctor
University studies: Computer Science and Technology
Scientific interest: Distributed System, Cloud Computing.
Publications: 11
Experience: Doctor of College of Computer Science and Technology, Jilin University, China.

A cloud-removal method based on image fusion using local indexes

Xiao Xiaohong[1, 2], Wu Yonggang[1*]

[1] *School of Hydropower and Information Engineering, Huazhong University of Science & Technology, Wuhan 430074, Hubei, China*

[2] *School of Computer Science, Huanggang Normal University, Huanggang 436000, Hubei, China*

Received 1 March 2014, www.tsi.lv

Abstract

For optical images, cloud and cloud shadow is always a problem during image processing and interpretation. Landsat ETM+ images, as a kind of optical images, are affected by cloud too. On the other hand, microwave images such as ALOS PALSAR images, which depend on microwave, is not affected by cloud, thus they are cloud-free. The aim of this study is to develop a semi-automatic method for removing cloud and cloud shadow in Landsat ETM+ images based on fusion of Landsat ETM+ image and ALOS PALSAR image. The key point of this method is to develop a cloud and cloud shadow mask based on which Landsat ETM+ and ALOS PALSAR images can be fused. To accurately define cloud and cloud shadow area, we first approximately draw the area of interest containing cloud and cloud shadow manually, and the resulted AOI image greatly reduce the number of ground objects and the confusion between objects as well. By analysing the spectral and the grey value of the AOI image, we then define LCI (local cloud index), LSI (local shadow index), and LGI (local ground index) to accurately identify cloud and cloud shadow area in Landsat ETM+ images. Finally, a combination mask of cloud and cloud shadow is developed. Based on this mask, Landsat ETM+ image and ALOS PALSAR image are merged. The fused image is cloud free, at the same time; it keeps the spectral feature and the integrity of Landsat ETM+ image.

Keywords: Cloud-removal, AOI, Landsat ETM+, ALOS PALSAR, LCI, LSI, LGI

1 Introduction

The Landsat series have long provided users high resolution and multi-spectral remote sensing data for scientific research and earth observation for almost 40 years since its first launch. It’s one of the most widely used satellites [1]. However, its imaging greatly depends on sunlight. Clouds are common features of images collected from many tropical, humid, mountainous, and coastal regions of the world [2]. The existence of cloud and cloud shadow will influence the processing and observation of ground objects. Therefore, during the pre-processing stage, both clouds and cloud shadows should be removed. Mainly, there are the following three kinds of methods for cloud-removal appearing in documents.

(1) Cloud-removal methods based on multi-spectral images. Tasselled cap transformation [3] is an often used method. Ritcher found that the fourth component of tasselled cap transformation is related to noise (cloud). It can be removed first, and then apply reverse tasselled cap transformation on the remaining components to produce a cloud-free image [4]. Tasselled cap transformation is highly dependent on sensors. It is only fit to MSS and TM (ETM) images. Besides, it will result in band information loss [5].

(2) Cloud-removal methods based on single image by using homomorphic filtering [6] or wavelet transformation [7]. Zhao and Zhu used homomorphic filtering method to remove thin clouds [8]. They first classified the cloud area, and then used interpolation method to restore the cloud area. Homomorphic filtering may acquire a good cloud removal result, but it can remove some useful information of the images. Moreover, once the area is covered by clouds and cloud shadows, we can hardly guess the true objects underneath the clouds and shadows through only one single image.

(3) Cloud-removal methods based on multi images. Use other images to complement the images to be processed. One method is time averaging. The idea is taking the pixel with smallest value from several images for the same area to produce a final image. Its precondition is that the shooting time of these images is near and the cloud area of each image is not overlapped, otherwise, the result is not ideal [9]. Another way is image fusion [10], which takes the corresponding part of a cloud-free image to replace the cloud and cloud shadow area of the image with clouds. Compared to multispectral and single image-based cloud removal methods, multi image-based cloud removal methods can achieve higher quality cloud-free images. On the other hand, as this method involves several images, the pre-processing of images is more complicated. Accurate registration and histogram matching is necessary.

When using multi images for cloud-removal, the key is to define cloud and cloud shadow area. In present, the

[*] ***Corresponding author*** e-mail: ygangwu@163.com

general way is to define some kinds of cloud index(CI) to determine cloud, and then according to the shape of cloud and the distance between shadow and cloud to determine shadow area. Combination of Total Reflectance Radiance Index (TRRI) and Cloud-Soil Index (CSI) is used to define cloud by Nguyen Thanh Hoan when removing clouds of optical images [11]. However, various CIs proposed are defined globally, that's, they are based on the overall images. As we know, for a remote sensing image, there are a great number of ground objects. Due to the phenomenon of "same objects having different spectrum or same spectrum corresponding to different objects", CI is very difficult to be defined. Even though it is determined, some non-cloud area may be extracted, while some cloud area may be omitted. To alleviate this problem, based on human vision [12], we first approximately draws the area of interest (AOI) including cloud and cloud shadow so as to reduce the number of ground objects at a minimum level. The resulted AOI image is mainly composed of cloud, cloud shadow and ground area. Then we define Local Cloud Index (LCI), Local Cloud Shadow Index (LSI), and Local Ground Index (LGI) by analysing the spectral feature of the AOI image. As there are fewer objects appearing in the AOI image, there is less confusion between objects, and it is much easier to define LCI, LSI, and LGI. Finally, we use LCI, LSI, and LGI to accurately produce the combination mask of cloud and cloud shadow area, based on which Landsat ETM+ image with cloud and cloud-free ALOS PALSAR HH image are merged together to produce cloud-free Landsat ETM+ image. This method can help to make a series of free cloud multi-temporal images for change detection studies, land cover classification studies, environmental monitoring studies and so on.

2 Data

2.1 LANDSAT ETM+ DATA

In the present study, Landsat ETM+ image (Path 122, Row 39) dated on August 19, 2008 has been used (see table 1). Landsat ETM+ image data consist of eight spectral bands, with a spatial resolution of 30 meters for bands 1 to 5 and band 7. The resolution for band 6H/6L (thermal infrared) is 60 meters or 30 meters. The resolution for band 8 (panchromatic) is 15 meters. The approximate scene size is 170 km north-south by 183 km east-west (106 mi by 114 mi). Figure 1 shows the false colour composite image of band4, band3 and band2 of image1.

TABLE 1 Landsat ETM+ data specifications

Acquired time	Path/row	Landsat sensor	Cloud cover (%)	Used band
2008-08-19	122/39	ETM+(SLC-off)	5.35	1,2,3,4,5,6,7

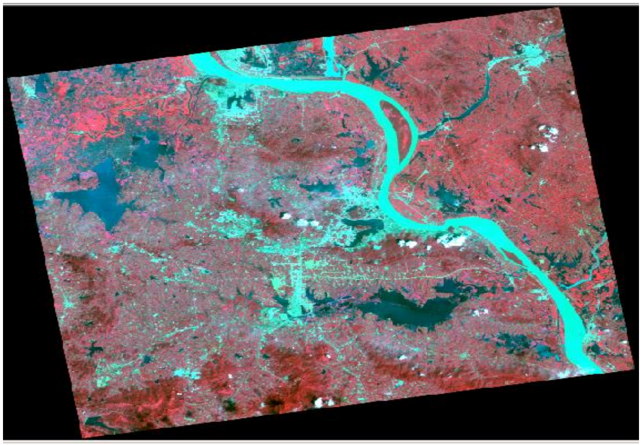


FIGURE 1 Landsat ETM+ 432 false colour composite image (20080819)

2.2 ALOS PALSAR DATA

ALOS PALSAR data is a Japanese Earth observation satellite carrying a cloud-piercing L-band radar, which is designed to acquire fully polarimetric images. The resolution is 12.5 meters. In the present study, the HH polarization image dated on July 3, 2008 is mainly used (table 2). As shown in figure 2, there is no cloud at all.

TABLE 2 ALOS PALSAR Data Specifications

Track/ Frame	Date	Mode (Polarization)	Incidence angle	orbit
454/590	2008-07-03	FBD(HH)	34.3	ascending

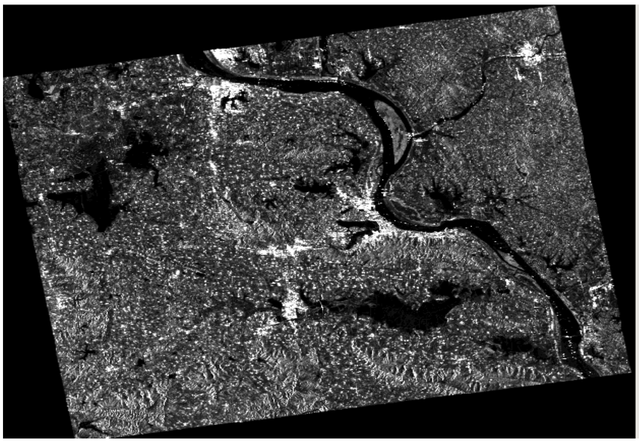


FIGURE 2 ALOS PALSAR HH image (20080703)

3 Methodologies

Figure 3 shows the whole processing framework.

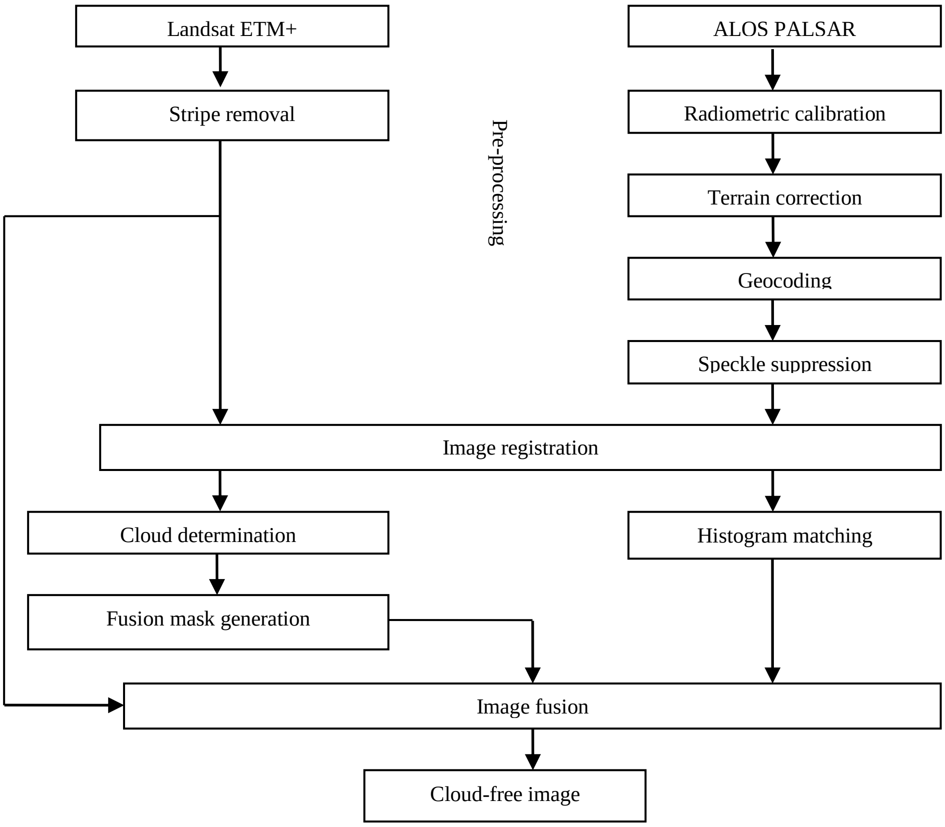


FIGURE 3 The whole processing framework of cloud removal

3.1 PREPROCESSING OF LANDSAT ETM+ DATA

The Landsat ETM+ data were acquired from http://datamirror.csdb.cn/ with stripes. For each band, the strip was successfully removed by multi image adaptive local regression method (RGF) provided on http://datamirror.csdb.cn/.

3.2 PREPROCESSING OF ALOS PALSAR DATA

The ALOS PALSAR images were provided by Alaska Satellite Facility at 1.5 level format. It was pre-processed through ASF MapReady 3.0 software package developed by the Engineering group at the Alaska Satellite Facility. Pre-processing included radiometric calibration using Sigma calibration coefficients, terrain correction based on DEM information, topographic normalization as well as geocoding to 30m pixel resolution (WGS84, UTM 50N).

An obvious disadvantage of ALOS PALSAR image is its speckles which greatly degrades image quality and influences land cover classification and interpretation. Here, a 3*3 local region adapter was adopted to reduce speckles of ALOS PALSAR images.

As Landsat ETM+ and ALOS PALSAR are taken from different sensors, and have different sizes, co-registration is necessary. ALOS PALSAR image was rectified to the coordinates of the Landsat ETM+ image using 12 ground control points (GCPs) defined from a topographic map of the study area. For the transformation, a second-order transformation and nearest-neighbour resampling approach were applied and the related root mean square error was 0.5 pixels.

3.3 FUSION MASK DETERMINATION

Generally one image often contains a great number of objects, due to the phenomenon of "same objects having different spectrum and same spectrum corresponding to different objects"; they tend to be confused with each other. As figure 4 shows, the spectrum of cloud shadow and lake can hardly be distinguished for band 1 to 7, and the spectrum of thin cloud and building can be easily confused as showing in figure 5. Therefore, when extracting a particular object based on a predefined rule, some parts not belonging to the object may be extracted, while some parts belonging to it may be omitted.

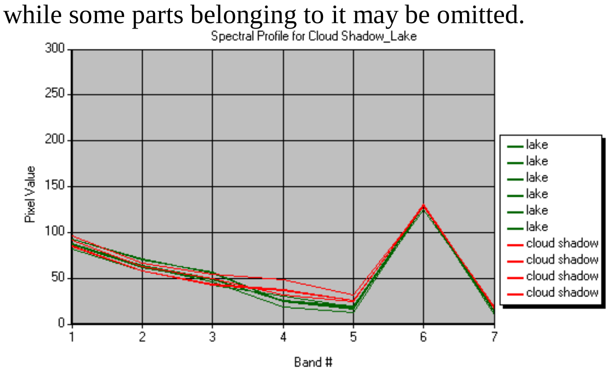


FIGURE 4 The spectral profile of cloud shadow and lake

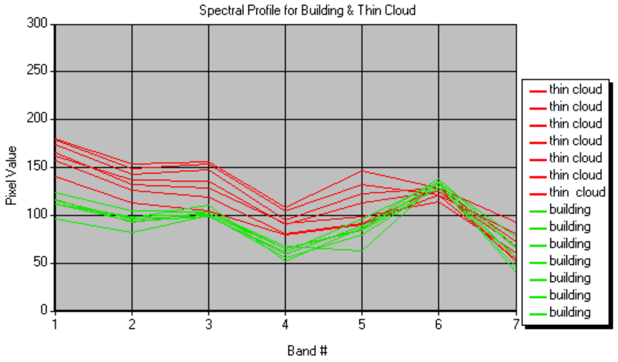


FIGURE 5 The spectral profile of building and thin cloud

In practical application, we often focus on one particular object instead of all objects. As shown in figure 1(432 false colour composite image), there are three objects relevant to cloud: (1) thick cloud, with white colour; (2) thin cloud, with light blue colour, easy to be confused with buildings; (3) cloud shadow, with black colour, easy to be confused with water(lake). We can first approximately separate the cloud area manually so that there are fewer objects involved and thus reducing the confusion at a great scale.

Figure 6 shows the processing flow.

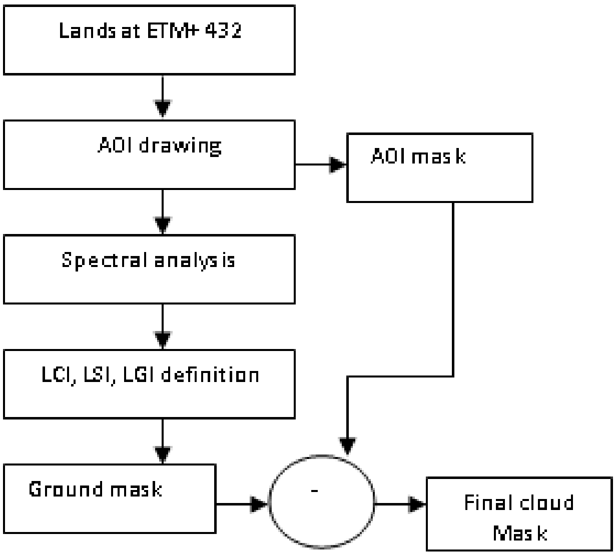


FIGURE 6 The processing flow of cloud and cloud shadow mask generation

(1) AOI drawing

In "Towards cognitive image fusion" [12], based on human visual system, dozens of observers are invited to draw the outline of ground objects. As human vision is sensitive to different objects, and has good ability of macro control over objects, we can first draw the approximate outline of cloud and cloud shadow area based on eyes' observation so as to greatly reduce the number of objects in the image. Here, we first draw a block of cloud area, as shown in figure 7. In figure 7, there are only three classes of objects: cloud (white thick cloud, light blue thin cloud), cloud shadow (dark black), and ground (red, in 432, it is vegetation).

(2) LCI, LGI, LSI definition

For each class, we take 10-12 samples from different parts to draw their spectral curves. Figure 8 gives the spectrum of thick cloud, thin cloud, cloud shadow, and ground. We found that for each class, the spectral curves of the sampled points are almost the same; for different classes, there is great difference between them. This feature can help us more easily to classify them.

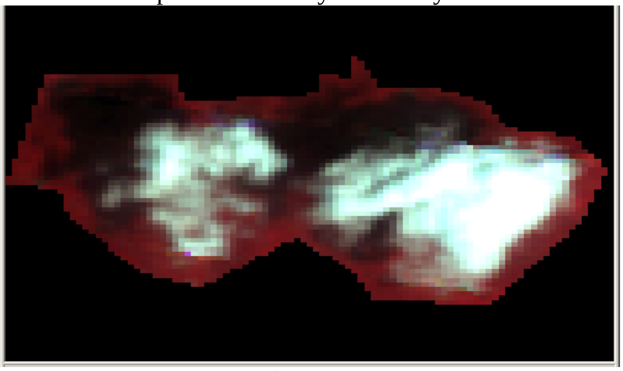


FIGURE 7 A block of cloud and cloud shadow area

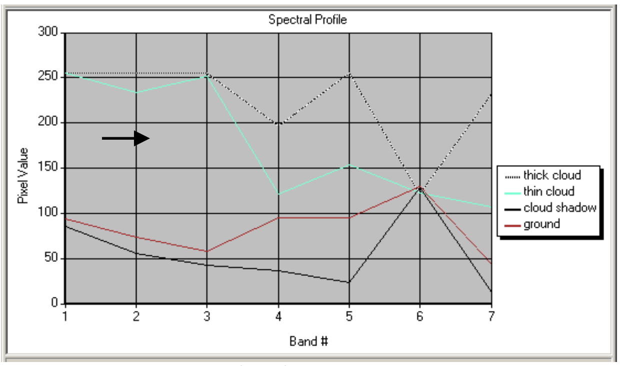


FIGURE 8 the spectral profile of thick cloud, thin cloud, cloud shadow and ground

Table 3 gives the grey value range of sampled pixels for each class.

TABLE 3 the grey value range of sampled pixels for each class

Class band	Thick cloud	Thin cloud	Cloud shadow	Ground
Band1	255-255	175-242	78-84	**87-110**
Band2	227-255	147-212	50-56	**65-83**
Band3	243-255	154-234	37-44	**47-72**
Band4	144-200	103-133	28-38	**64-89**
Band5	211-255	120-201	16-26	**61-101**
Band6	121-123	124-126	128-131	**129-132**
Band7	**73-136**	**73-136**	**9-16**	**28-50**

1) LCI definition

Based on the above spectral curves and the gray value range of sampled pixels, we concluded that the gray value of cloud(no matter thick cloud or thin cloud) in band1, band2, and band3 is much higher than that of cloud shadow and ground, thus the rule to extract cloud can be written as:

$$band1 > 130\ And\ band2 > 100\ And\ band3 > 100\,. \qquad (1)$$

2) LSI and LGI definition

The grey value of both cloud shadow and ground is almost less than 100 in band2, band3, band4, band5, and band7. However, from the spectral curve and gray range of sampled pixels, we can see the difference between cloud shadow and ground. The gray value of ground in each band is greater than that of cloud shadow, besides, for cloud shadow, the gray value displays such a relation: band5<band4<band3; while for ground, the relation is like this: band5>band4>band3. After dozens of times experiments, we built the following rules for cloud shadow and ground:

LSI for cloud shadow:

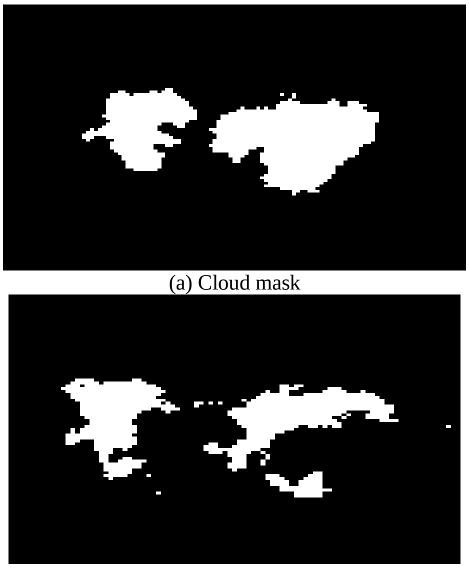
$$\frac{band\,5-band\,3}{band\,5+band\,3}<0\ (band\,2<100\ And\ band\,3<100)\,,(2)$$

LGI for Ground:

$$\frac{band\,5-band\,3}{band\,5+band\,3}>0\ (band\,2<110\ And\ band\,3<110)\,.(3)$$

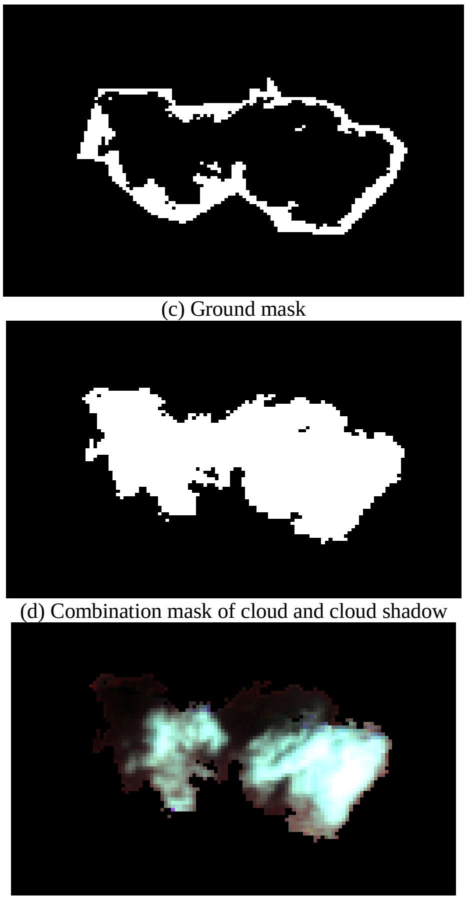
(3) Cloud mask determination

Based on the above LCI, LSI, and LGI, the following cloud mask (figure 9(a)), cloud shadow mask (figure 9(b)), and ground mask (figure 9(c)) are obtained. As considering, only the spectral feature of ground is much simpler, and it does not involve the transitional area between cloud and cloud shadow, we can first obtain the ground mask, and then use the AOI to subtract the ground mask to get the final mask. Doing this is less time-consuming, and the resulted final mask is intact, as shown in figure 9 (d) and (e).



(a) Cloud mask

(b) Cloud shadow mask



(c) Ground mask

(d) Combination mask of cloud and cloud shadow

(e) Cloud and cloud shadow area extraction

FIGURE 9 Cloud, cloud shadow, ground, and combination mask

4 Cloud-free image generation

4.1 CLOUD AND CLOUD SHADOW MASK GENERATION

Based on the above method, we first draw out AOI to produce the approximate outline of cloud and cloud shadow. When drawing AOI, the shooting time, the direction of cloud shadow, and corresponding relation between the shape of cloud and cloud shadow should be taken into full consideration. At the same time, it would be better to use colour composite image instead of single band image because single band image is a kind of greyscale image, cloud, cloud shadow and other objects can be easily confused with each other. In this paper, the 432 false colour composite image is used, in which, the thick cloud (white colour), cloud shadow (dark black), and ground (dark red) can be clearly distinguished. Moreover, during this process, some image processing experts' advices should be adopted.

The combination mask of cloud and cloud shadow is shown as figure 10.

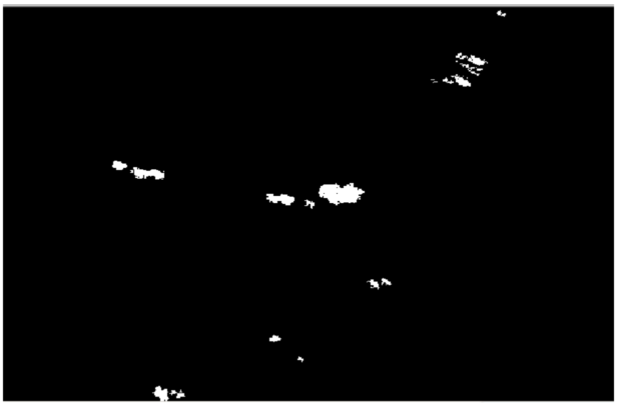


FIGURE 10 The combination mask of cloud and cloud shadow (cloud_mask)

4.2 CLOUD-FREE IMAGE GENERATION BASED ON IMAGE FUSION

Here, we take Landsat ETM+ band5 image (image1) and ALOS PALSAR HH image (image2) as an example. The former one has clouds, while the latter one has no cloud at all. Figure 11 shows the former one, and figure 2 shows the latter one. Both of them are well pre-processed for later image fusion.

As these two images are taken at different time and under different environment, there colour is not uniform. Before fusion, histogram matching is necessary. Based on the combination mask and the following rule, the above two images are merged. The final fused image is shown as figure 12. There is no cloud in the fused image at all.

$$F(x,y)=\begin{cases} image1(x,y) & cloud\_mask(x,y)=0 \\ image2(x,y) & cloud\_mask(x,y)=1 \end{cases}, \tag{4}$$

where F(x,y) refers to the final fused image.

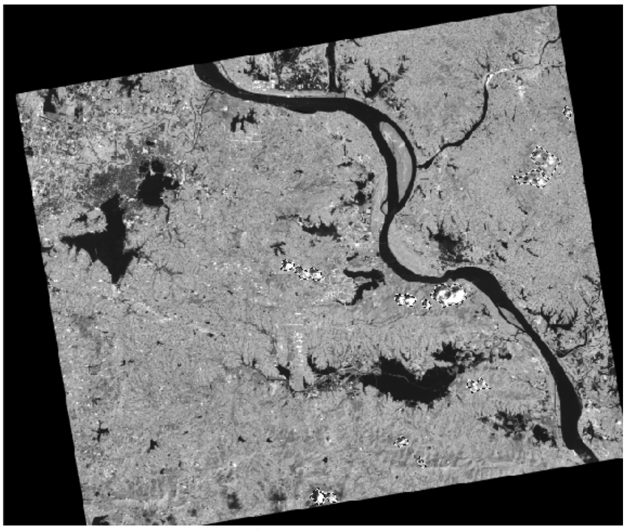


FIGURE 11 band5 of 2008-08-19 with clouds (image1)

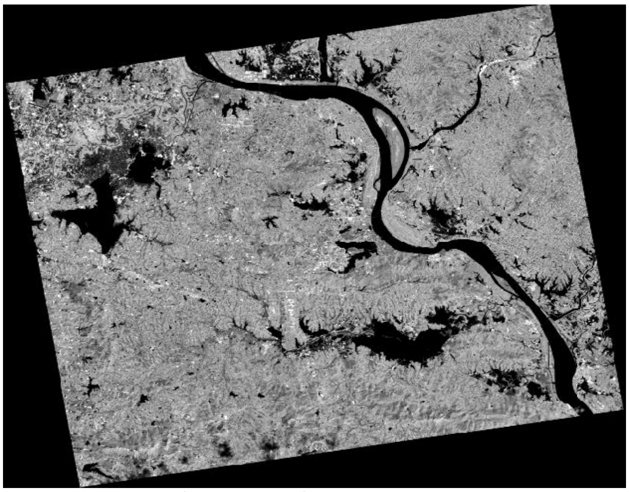


FIGURE 12 The fused image of Landsat ETM+ band5 and ALOS PALSAR HH

5 Conclusions

Due to the phenomenon of “same objects having different spectrum and same spectrum corresponding to different objects”, if there are a great number of ground objects in an image, they tend to be confused with each other. When classifying land types, according to the actual situation, we can do it locally instead of globally so that the number of land types can be greatly reduced, thus decreasing confusion between different land types.

For cloud-removal, the focus is cloud and cloud shadow area classification, therefore, taking human vision into consideration, we first approximately drew the interest area related to cloud to reduce the number of ground objects at a minimum scale. The AOI mainly contains three land type classes: cloud, cloud shadow, and ground. Based on the AOI, by analysing the spectral feature and grey value of these three classes, we then defined LCI, LSI, and LGI, which can correctly extract cloud, cloud shadow and ground area without confusion or omission.

When generating the combination mask of cloud and cloud shadow, we first obtained the ground mask, and then used the AOI mask to subtract the ground mask to produce the final combination mask of cloud and cloud shadow. This method is simple and not time consuming, and since the transitional zone between cloud and cloud shadow is not involved in the processing, there are no holes left in the final mask.

For image fusion, a very important point is not to destroy the original images, but to combine the good points of the original images. Landsat ETM+ images can better reflect land cover types, but they are affected by whether and clouds, while ALOS PALSAR images are not affected by whether and clouds, however, they have speckles, and are not good for interpretation. Therefore, we can use ALOS PALSAR images to complement Landsat ETM+ images to remove the cloud and cloud shadow area of Landsat ETM+ images, and at the same time retain the multispectral feature and the integrity of

Landsat ETM+ at a maximum extent.

After pre-processing of cloud-free ALOS PALSAR HH image, we merged Landsat ETM+ band5 image with ALOS PALSAR HH image, and got a cloud-free image to a good effect.

The method proposed in this paper is suitable for removing closely tied big clouds but not scattered small clouds. Removing scattered small clouds would be more time-consuming.

Acknowledgements

This project was funded by Hubei Provincial Education Bureau and Huanggang Normal University jointly. Landsat ETM+ images are provided by http://datamirror.csdb.cn/. We thank to Dr. Don Atwood of Alaska Satellite Faculty for providing two sets of ALOS PALSAR data and ASF MapReady 3.0 software for ALOS PALSAR data pre-processing. We also thank to RSMAS of University of Miami for providing a good experimental environment.

References

[1] Song Xiaoyu, Liu Liangyun, Li Cunjun, Wang Jihua, Zhao Chunjiang, 2006. Cloud Removing Based on Single Remote Sensing Image. *Optical Technique* **32**(2) 299-303

[2] Martinuzzi S, Gould W A, Ramos González O M 2007 Creating Cloud-Free Landsat ETM+ Data Sets in Tropical Landscapes: Cloud and Cloud-Shadow Removal *General Technical Report IITF-GTR-32* 1-12

[3] Horne J H 2003 A Tasselled Cap Transformation for IKONOS Images *ASPRS 2003 Annual Conference Proceedings, Anchorage, Alaska*

[4] Richter R 1996 Atmospheric Correction of Satellite Data with Haze Removal Including a Haze/Clear transition Region *Comput. Geosci*. **22** 675-81

[5] Cao Shuang 2006 *Cloud-removal Methods of High Resolution Remote sensing Images* Master Dissertation of Hehai University

[6] Li Gang, Yang Wunian, Weng Tao 2007 A Method of Removing Thin Cloud in Remote Sensing Image Based on the Homomorphic Filter Algorithm *Science of Surveying and Mapping* **32**(3) 47-8

[7] Chen Fen, Yan Dongmei, Zhao Zhongming 2007 Haze Detection and Removal in Remote Sensing Images Based on Undecimated Wavelet Transform *Geomatics and Information Science of Wuhan University* **32**(1) 71-4

[8] Zhao Zhongming, Zhu Chongguang 1996 Thin Cloud Removal Methods of Remote Sensing Images *Journal of Remote Sensing* **11**(3) 195-9

[9] Liu Yang, Bai Junwu 2008 Research on the Cloud Removal Method of Remote Sensing Images *Geomatics & Spatial Information Technology* **3** 120-2

[10] Guo Tongying, You Hongjian 2007 Cloud Reduction of Milts-temporal Space-borne Remote Sensing Image Based on Wavelet Fusion *Bulletin of Surveying and Mapping* (3) 40-2

[11] Nguyen Thanh Hoan, Ryutaro Tateishi 2008 Cloud Removal of Optical Image Using SAR Data for ALOS Applications. Experimenting on Simulated ALOS Data *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **37**(Part B4) Beijing 379-84

[12] Toet A, Hogervorst M A, Nikolov S G, Lewis J J, Dixon T D, Bull D R, Canagarajah C N 2009 Towards Cognitive Image Fusion *Information Fusion* 1-19

Authors

	Xiaohong Xiao, born on July 14, 1975, Hubei, China **Current position, grades:** she is pursuing the Ph.D. degree at School of Hydropower and Information Engineering at Huazhong University of Science & Technology **University studies:** M.Ei. Degree in Computer technology in 2004 from Wuhan University of Science and Technology in Hubei, China **Scientific interest:** remote sensing image processing and image fusion for earth observation **Experience:** In 2013, she had been in RSMAS (Rosenstiel School of Marine and Atmospheric Science) at University of Miami as an exchange scholar for half a year, where she mainly engaged in SAR image processing and image fusion of optical images and Radar Images.
	Yonggang Wu, born on October 27, 1963, Hunan, China **Current position, grades:** professor at School of Hydropower and Information Engineering at Huazhong University of Science & Technology **University studies:** Ph.D. degree in 1997 from Huazhong University of Science and Technology **Scientific interest:** Generic Algorithm, AGC(Automatic Gain Control), Reservoir Optimized Operation and Control, Water Resources, and Image Recognition **Experience:** He had been involved in several projects related to the optimization scheduling and AGC of the Three Gorges cascade hydropower station, the Gezhou Dam Er Jiang Power Station, Fujian Reservoirs, and so on. Right now he is holding a Natural Science Foundation of China project related to the AGC of power grid.

An efficient method for acquiring and processing signals based on compressed sensing

Song Xiaoxia[*]

School of Mathematics and Computer Science, Shanxi Datong University, Datong 037009, Shanxi, China

Received 1 March 2014, www.tsi.lv

Abstract

Compressed sensing (CS) theory provides a novel sensing/sampling and processing paradigm that breaks through the limitation of Nyquist rate to some applications. However, it is usually happened to the instability and redundancy of the acquired CS measurements. In view of this, we propose an efficient method to achieve adaptive minimal measurements with fewer measurements and good reconstruction performance by adding the pre-processing block into CS data acquiring and processing paradigm. In the proposed method, we firstly obtain the measurements to perfectly reconstruct the signal, and then design the optimization method to obtain adaptive minimal measurements by eliminating the redundant measurements. Experimental results show that the proposed method can obtain fewer measurements to perfectly reconstruct the signal than that of classical CS and sequential compressed sensing frameworks.

Keywords: compressed sensing, sequential compressed sensing, signal reconstruction, homotopy method

1 Introduction

In the conventional approach to sampling signal, the sampling rate must satisfy the Shannon/Nyquist sample theorem to not lose information [1-3]. Then the signal must be compressed to transmit or store since the high Nyquist rate results in too more redundant samples. FIGURE 1 illustrates the procedure for acquiring and processing signals in the conventional approach, including sampling and compressing the signal, transmitting/storing the data, and decompressing from the received data. While too high Nyquist rate limits some applications [2, 4], such as medical scanners, radar imaging and high-speed analogy-to-digital converters (AIC). Fortunately, emerging compressed sensing (CS) theory [1-7] provides a novel sensing/sampling paradigm that breaks through the limitation of the traditional approach. The CS theory claims that far fewer samples or measurements than the conventional approach can be used to perfectly recovery the signals when restricted isometry property (RIP) is satisfied and the underlying signal is sparse [1-8]. FIGURE 2 demonstrates data acquirement by CS method and then data processing, including obtaining the measurements at the sender, transmitting/storing, and decoding at the receiver. However, such classical CS framework cannot ensure that the acquired measurements can certainly be used to perfectly reconstruct the signal [9]. In addition, the measurements transmitted/stored are usually redundant [10], which will result in the waste of transmission/storage resources.

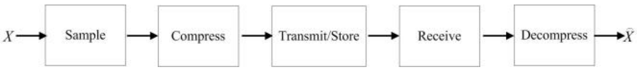


FIGURE 1 The conventional approach to sampling and processing signals

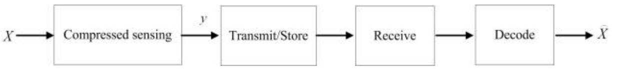


FIGURE 2 Compressed sensing to sampling and processing signals

To address the above two problems of classical CS framework, this paper proposes a new efficient method shown in FIGURE 3 to acquire and process the data based on CS. In the proposed method, we adds the pre-processing block B to ensure that the measurement results can be used to perfectly reconstruct the signal and obtain adaptive minimal measurements, whose any proper subset cannot be used to perfectly reconstructed the signal. In the proposed method, we firstly judge whether initial measurements are excess or not via sequential compressed sensing (SCS) [11]. Then for two cases of non-excess and excess initial measurements, we design the optimization method to obtain adaptive minimal measurements. Experimental results show that the measurement of a certain measurement set has indeed different important degree to signal reconstruction, and the proposed method can obtain fewer measurements to perfectly reconstruct the signal than that of classical CS and sequential compressed sensing frameworks.

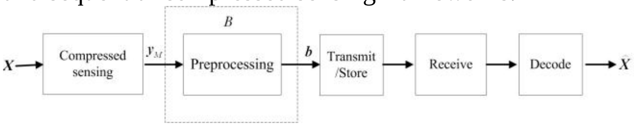


FIGURE 3 The proposed method to sampling and processing signals

[*] ***Corresponding author*** e-mail: sxxly2002@163.com

Before proceeding, we define some denotations. Let $\boldsymbol{X}^*$ be the signal with the length N and the scarcity level k . Let $\boldsymbol{y}^M = [y_1, y_2, \cdots, y_M]^T$ represent initial measurements, where M is the number of initial measurements. Then $\boldsymbol{y}^M = \boldsymbol{\Phi X}^*$, where $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \cdots, \boldsymbol{\Phi}_M]^T$ is $M \times N$ Gaussian measurement matrix. The measurements $\boldsymbol{s} = [s_1, s_2, \cdots, s_L]^T$ are called as SCS or MSCS measurements since it is obtained by SCS or MSCS method, and its corresponding measurement matrix denoted by $\boldsymbol{S}$. Let $\boldsymbol{b} = [b_1, b_2, \cdots, b_K]^T$ $(K \le L)$ be adaptive minimal measurements to transmit/store, whose corresponding measurement matrix denoted by $\boldsymbol{B}$. Let $\widehat{\boldsymbol{X}}_M$ represent the reconstruction signal by using initial measurements $\boldsymbol{y}^M$. Let $\widehat{\boldsymbol{X}}_i$ represent the reconstruction signal by using $\boldsymbol{y}^i = [y_1, y_2, \cdots, y_i]^T$, where i is a positive integer.

2 The proposed method for acquiring and processing the signal

In this section, we propose an efficient method to acquire and process signals based on compressed sensing, recall FIGURE 3. In the proposed method, adaptive minimal measurements with fewer measurements and good reconstruction performance can be achieved by adding the pre-processing block. In the block, we firstly achieve the measurements, which can be used to perfectly reconstruct the signal with probability 1 for the Gaussian measurement ensemble. Then the optimization method is used to reduce redundant measurements to obtain adaptive minimal measurements.

2.1 THE FLOW CHART OF THE PRE-PROCESSING BLOCK

To understand easily the function of the block B , we give the flow-chart of B in FIGURE 4. After acquiring initial measurements $\boldsymbol{y}^M$, we firstly judge that $\boldsymbol{y}^M$ is excess or non-excess. If $\boldsymbol{y}^M$ is non-excess, SCS method is used to achieve SCS measurements $\boldsymbol{s}$. Otherwise, we provide MSCS method to obtain MSCS measurements $\boldsymbol{s}$ by removing some redundant measurements from $\boldsymbol{y}^M$. So SCS or MSCS measurements can be used to perfectly reconstruct the signal. And we find that the signal will not can be perfectly reconstructed if the last measurement of $\boldsymbol{s}$ is removed. It demonstrates that the last measurement of is important.

Inspired by the finding that the last measurement of $\boldsymbol{s}$ is important, we study each measurement of $\boldsymbol{s}$ and conclude that some measurements are important since the signal cannot be perfectly reconstructed if any of them is removed from $\boldsymbol{s}$, and some measurements are unimportant since the signal can still be perfectly reconstructed if any of them is removed from $\boldsymbol{s}$. Therefore, each measurement of a certain measurement set has different important degree to signal reconstruction. Based on this, the optimization method is used to achieve adaptive minimal measurements.

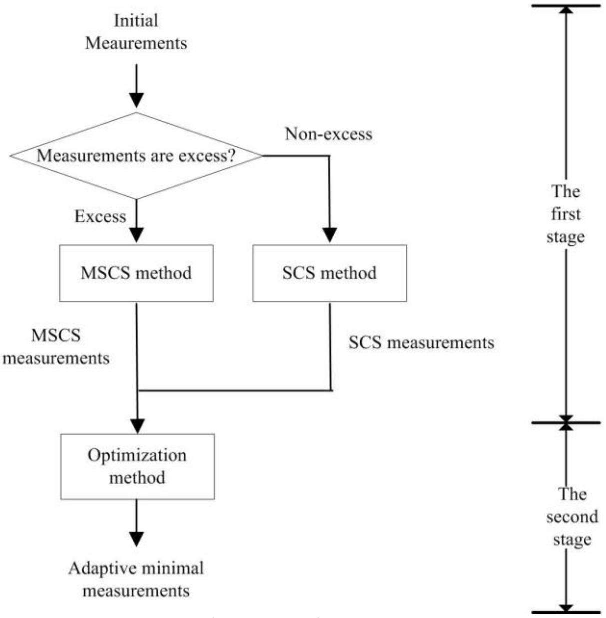


FIGURE 4 The flow-chart of the proposed method

2.2 THE PRE-PROCESSING BLOCK

In the first stage of the processing block B, we need to judge whether initial measurements $\boldsymbol{y}^M$ are excess or not. To this aim, MSCS method is given in FIGURE 5 and Proposition 2 is provided as judgment criterion based on Proposition 1.

Step 1: $\widehat{\boldsymbol{X}}_M$, $\widehat{\boldsymbol{X}}_{M-1}$, and $\widehat{\boldsymbol{X}}_M = \widehat{\boldsymbol{X}}_{M-1}$ are known , and let $i = M-2$.

Step 2: $\boldsymbol{y}^i$ is used to reconstruct $\widehat{\boldsymbol{X}}_i$.

Step 3: If $\widehat{\boldsymbol{X}}_M = \widehat{\boldsymbol{X}}_i$, then let $i = i-1$, and repeat step 2. Otherwise go to step 4.

Step 4: $\boldsymbol{y}^{i+1}$ is MSCS measurements $\boldsymbol{s}$.

FIGURE 5 MSCS method

In FIGURE 5, we know $\widehat{\boldsymbol{X}}_M = \widehat{\boldsymbol{X}}_{M-1}$ since MSCS method deals with the case that initial measurements are excess, and let $i = M-2$, see step 1. The first i measurements $\boldsymbol{y}^i$ of initial measurements $\boldsymbol{y}^M$, whose corresponding measurement matrix is $[\boldsymbol{\Phi}_1, \cdots, \boldsymbol{\Phi}_{i-1}, \boldsymbol{\Phi}_i]^T$, is used to reconstruct the signal $\widehat{\boldsymbol{X}}_i$, see step 2. If $\widehat{\boldsymbol{X}}_M \ne \widehat{\boldsymbol{X}}_i$, then $\boldsymbol{y}^{i+1}$ is just MSCS measurements. Otherwise let $i = i-1$ and repeat steps 2-3, see steps 3-4. MSCS measurements $\boldsymbol{y}^{i+1} = [y_1, \cdots, y_i, y_{i+1}]^T$ are rewritten as $\boldsymbol{s} = [s_1, s_2, \cdots, s_L]^T$, where $L = i+1$. And the corresponding measurement matrix is $\boldsymbol{S}=[\boldsymbol{\Phi}_1,\cdots,\boldsymbol{\Phi}_i,\boldsymbol{\Phi}_{i+1}]^T$.

MSCS method is so named because it adopts such a procedure in which the measurement is eliminated one-by-one from initial measurements until the stop condition is satisfied, while SCS method increases the measurement one-by-one until the agreement rule is satisfied. According to the procedure shown in FIGURE 5, we know that the signal will not can be perfecty reconstructed if the last measurement of MSCS measurements is removed. It demonstrates that the last measurement of MSCS measurements is also important, which is the same as that of SCS method.

Proposition 1 [11] In the Gaussian (generic continuous) measurement ensemble, if $\widehat{\boldsymbol{X}}_{M+1}=\widehat{\boldsymbol{X}}_M$ holds, then $\widehat{\boldsymbol{X}}_M=\boldsymbol{X}^*$ with probability 1.

Proposition 2 In the Gaussian (generic continuous) measurement ensemble, if $\widehat{\boldsymbol{X}}_M=\widehat{\boldsymbol{X}}_{M-1}$, then $\boldsymbol{y}^M$ is excess. Otherwise $\boldsymbol{y}^M$ is non-excess.

According to the principles of SCS and MSCS methods, the measurements $\boldsymbol{s}$ can be used to perfectly reconstruct the signal with probability 1 for the Gaussian measurement ensemble. To reduce the number of the measurements to store/transmit, the optimization algorithm shown in FIGURE 6 can be used to achieve adaptive minimal measurements according to the different important degree of the measurement.

In FIGURE 6, we take SCS or MSCS measurements $\boldsymbol{s}$ as the input of the optimization method, see steps 1-2. In steps 3-6, $\boldsymbol{s}$ are divided into two sets: the key set T_1 and the non-key set T_2. In step 4, the important degree of each measurement is illustrated by the reconstruction error denoted by $E_j=\left\|\widehat{\boldsymbol{X}}_p-\widehat{\boldsymbol{X}}_{p-1,j}\right\|/\left\|\widehat{\boldsymbol{X}}_p\right\|$, which can be solved since $\widehat{\boldsymbol{X}}_p$ can be regarded as the original signal. If T_2 is empty or only has a measurement, then T_1 is just adaptive minimal measurements $\boldsymbol{b}$, see step 7. Otherwise, $\boldsymbol{b}$ should be composed of T_1 and some measurements of T_2. Since T_2 is sorted by the descending order of E_j, we may consider that the important degree of the measurement in T_2 is also decreasing. Next we try to remove the measurements as many as possible from the back of T_2 so that the remainder of T_2 together with T_1 can be used to perfectly reconstruct the signal, see step 8. The measurements w are updated, then repeat the above procedure until the condition in step 7 is satisfied, see step 9. So we obtain adaptive minimal measurements $\boldsymbol{b}$, which is a subset of $\boldsymbol{s}$. And the corresponding measurement matrix $\boldsymbol{B}$ is a submatrix of $\boldsymbol{S}$.

Step 1: Let $\boldsymbol{w}=\boldsymbol{s}$, and w contains p components. Then $p=L$.

Step 2: $\widehat{\boldsymbol{X}}_p$ is reconstructed by using $\boldsymbol{w}$.

Step 3: $\widehat{\boldsymbol{X}}_{p-1,j}$ is reconstructed by taking $p-1$ measurements obtained by removing the *j-th* measurement from $\boldsymbol{w}$ ($1\le j\le p$).

Step 4: Compute reconstruction error $E_j=\left\|\widehat{\boldsymbol{X}}_p-\widehat{\boldsymbol{X}}_{p-1,j}\right\|/\left\|\widehat{\boldsymbol{X}}_p\right\|$.

Step 5: Sort $\boldsymbol{w}$ into $\boldsymbol{z}$ according to the descending order of E_j.

Step 6: The measurements of $\boldsymbol{z}$ are divided into two sets T_1 with m_1 components and T_2 with m_2 components ($m_1+m_2=p$ according to *Theorem 1*), where the corresponding $E_j=0$ for each measurement of T_2, and the corresponding $E_j\neq 0$ for each measurement of T_1.

Step 7: If T_2 contains 0 or 1 measurement, then T_1 can be used to perfectly reconstruct the signal. So T_1 is just low-redundancy measurements $\boldsymbol{b}$. Otherwise go to step 8.

Step 8: Let $l=m_2-1$. T_1 together with the first l measurements of T_2 are taken to reconstruct the signal. If the signal can be perfectly reconstructed, then $l=l-1$ and repeat step 8. Otherwise go to step 9.

Step 9: Update $\boldsymbol{w}$ with T_1 and the first $l+1$ measurements of T_2, and $p=m_1+l+1$. Repeat the above steps 2-9.

FIGURE 6 The optimization algorithm to achieve adaptive minimal measurements

3 Experimental results

According to the principles of SCS and MSCS methods, the measurements $\boldsymbol{s}$ can be used to perfectly reconstruct the signal with probability 1 for the Gaussian measurement ensemble. So in this section, we need design some experiments to verify the following issues. (i) The measurement has different important degree to signal reconstruction for the measurements $\boldsymbol{s}$ and $\boldsymbol{b}$. (ii) The proposed method can obtain fewer measurements than that of the classical CS and SCS with good reconstruction performance. In the experiments, sparse signals are used as test signals, and homotopy method [12-14] is selected as the reconstruction algorithm since it is suitable to the recovery of sparse signals.

Experiment 1

In this experiment, a random signal with the length $N=200$ and the sparsity level $k=10$ is generated. And initial measurement numbers $M=30$ are adopted. The proposed method firstly obtains SCS measurements $\boldsymbol{s}$ with $L=36$ measurements. Then the reconstruction error $E_j=\left\|\widehat{\boldsymbol{X}}_L-\widehat{\boldsymbol{X}}_{L-1,j}\right\|/\left\|\widehat{\boldsymbol{X}}_L\right\|$ ($\widehat{\boldsymbol{X}}_L$, $\widehat{\boldsymbol{X}}_{L-1,j}$ refer FIGURE 6, $1\le j\le L$) is used to illustrate the important degree of each measurement in $\boldsymbol{s}$, the results are shown in FIGURE 7.

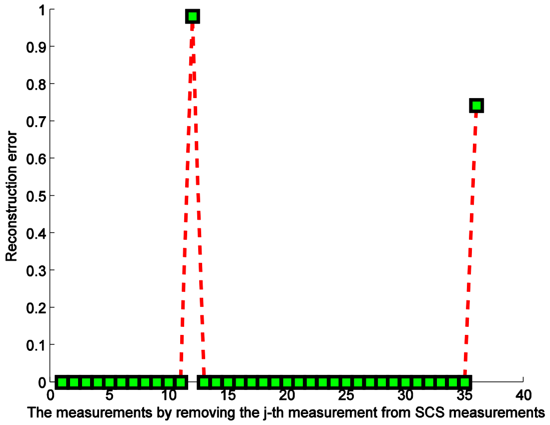


FIGURE 7 Reconstruction error E_j for the measurements obtained by removing the *j-th* measurement from SCS measurements $\boldsymbol{s}$

From FIGURE 7, we see that the signal cannot be perfectly reconstructed when the 12-th or 36-th measurement is removed from $\boldsymbol{s}$, while the signal can still be perfectly reconstructed when any other measurement is removed from $\boldsymbol{s}$. Apparently, the 12-th and 36-th measurements are more important than other measurements in $\boldsymbol{s}$, i.e., they are key measurements. So we can consider that the measurement in $\boldsymbol{s}$ has the different important degree to signal reconstruction. From FIGURE 7, it is easy to see $E_{12} > E_{36}$, which illustrates that the 12-th measurement is more important than the 36-th measurement. This viewpoint can also be verified by the next experiment.

For SCS measurements above, each key measurement is randomly replaced 1000 times to reconstruct the signal, respectively. The results show that, among 1000-time replacements, for the 12-th measurement, the signal can be perfectly reconstructed 19-time and the signal cannot be perfectly reconstructed 981-time. For the 36-th measurement, the signal can be perfectly reconstructed 446-time and the signal cannot be perfectly reconstructed 554-time. So the 12-th measurement is more important than the 36-th measurement from the perspective of probability. It illustrates that the measurement in $\boldsymbol{s}$ has different important degree.

For the above SCS measurements $\boldsymbol{s}$, the optimization method in FIGURE 6 is used to achieve adaptive minimal measurements $\boldsymbol{b}$ with $K = 23$ measurements. Each measurement of $\boldsymbol{b}$ is randomly replaced 1000 times to reconstruct the signal, respectively. The reconstruction probability for each measurement is shown in FIGURE 8. From FIGURE 8, we find that the signal can be perfectly reconstructed with a small probability ($0 \sim 0.257$) when the i-th measurement of $\boldsymbol{b}$ is randomly replaced 1000 times. Different reconstruction probability illustrates that each measurement of $\boldsymbol{b}$ has the different important degree to signal reconstruction. Especially for the 1-st measurement, the reconstruction probability is 0 among 1000-time replacements. It demonstrates that the measurement in $\boldsymbol{b}$ has different important degree.

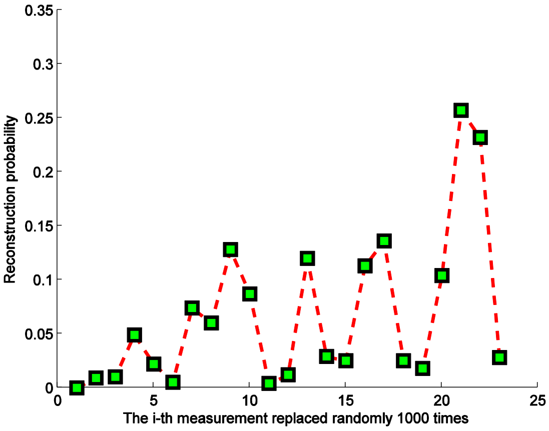


FIGURE 8 Reconstruction probability for the *i-th* measurement of $\boldsymbol{b}$ replaced randomly 1000 times

Experiment 2

Next experiments are used to verify that the proposed method can obtain fewer measurements with good reconstruction performance than that of the classical CS and SCS.

A signal with the length $N = 128$ and the sparsity level $k = 10$ is generated. We choose uniformly initial measurement numbers M from 10 to 60 with the interval 10. The experimental results are shown in TABLE 1.

In TABLE 1, the 1-st column M, the 3-rd column L and the 5-th column K represent the measurement numbers of $\boldsymbol{y}^M$, $\boldsymbol{s}$, $\boldsymbol{b}$, respectively. The 2-nd column E_1, the 4-th column E_2 and the 6-th column E_3 give the reconstruction errors when $\boldsymbol{y}^M$, $\boldsymbol{s}$, $\boldsymbol{b}$ are used to reconstruct the signal, respectively. When the sparsity level k is known, it is well known the fact that $3k \sim 5k$ measurements can be taken to perfectly reconstruct the signal with high probability. The results of the first two columns in TABLE 1 are consistent with the above conclusion. From the 3-rd column, we can see that the number of $\boldsymbol{s}$ lies between $3k \sim 4k$ which is a smaller range than that of the classical CS framework $3k \sim 5k$. The 5-th column shows that measurement numbers of $\boldsymbol{b}$ lie between $1k \sim 3k$, which is fewer than that of $\boldsymbol{s}$. According to CS theory, the minimal measurement numbers of this signal are $k + 1 = 11$ which is obtained by solving the NP hard problem of l_0 model. For $M = 40$, the number of $\boldsymbol{b}$ is $K = 14$, which is very near to the minimal value 11. Compared to initial measurements $\boldsymbol{y}^M$, $\boldsymbol{b}$ decrease 26 measurements, but it can be used to obtain almost the same reconstruction quality as $\boldsymbol{y}^M$.

TABLE 1 Reconstruction errors for different initial measurement numbers

M	E_1	L	E_2	K	E_3
10	1.1580	39	1.6999e-10	30	3.1438e-10
20	0.8543	34	1.9212e-10	21	7.8366e-11
30	0.8125	32	3.7478e-10	27	3.3689e-10
40	3.5374e-11	38	3.6561e-11	14	1.2733e-10
50	3.0281e-11	36	1.3316e-10	26	4.5560e-10
60	2.9799e-11	40	1.3754e-10	28	4.7159e-10

FIGURE 9 is used to intuitively illustrate the reconstruction performance of SCS measurements and adaptive minimal measurements sequences for $M=20$ of TABLE 1. The dashed line shows that the signal can be perfectly reconstructed when the initial measurements $M=20$ is increased adaptively to SCS measurements ($L=34$). The solid line shows that adaptive minimal measurements ($K=21$), which is a subset of SCS measurements, can also be used to perfectly reconstructed the signal with almost the same reconstruction error, see TABLE 1.

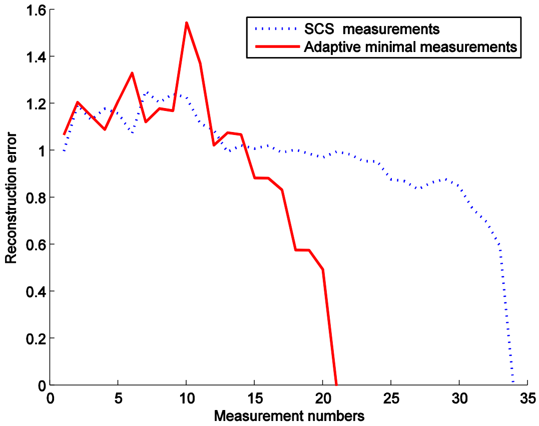


FIGURE 9 Reconstruction errors of SCS measurements and adaptive minimal measurements sequences for $M=20$

To verify the generality of the conclusion reflected by TABLE 1, we do the following statistical experiment. The above signal is still adopted, and 1000 experiments are run for $M=35$. The detailed results are shown in FIGURE 10. In this figure, the horizontal-axis represents the measurement times, the vertical-axis shows the measurement numbers for each measurement task. The solid line represents adaptive minimal measurement numbers, and the dashed line shows SCS or MSCS measurement numbers. From FIGURE 10, all adaptive minimal measurement numbers are fewer than that of SCS or MSCS measurements in these 1000 random measurements, while they can obtain almost the same reconstruction performance. Among 1000 random measurements, 827 SCS or MSCS measurement numbers are not more than $4k$, 743 adaptive minimal measurement numbers are not more than $3k$.

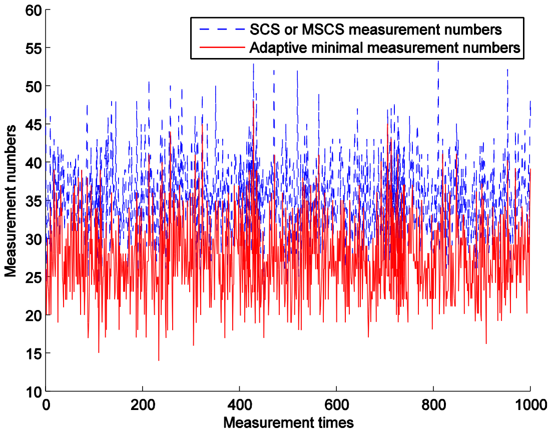


FIGURE 10 Comparisons with SCS or MSCS measurements, and adaptive minimal measurements. ($M=35$)

4 Discussions and conclusions

For the instability and redundancy of the acquired CS measurements, we propose an efficient method to achieve adaptive minimal measurements with fewer measurements and good reconstruction performance by adding the pre-processing block into CS data processing paradigm. In the proposed method, we firstly obtain the measurements to perfectly reconstruct the signal, and then design the optimization method to obtain adaptive minimal measurements by eliminating the redundant measurements. Experimental results show that the proposed method can obtain fewer measurements to perfectly reconstruct the signal than that of classical CS and SCS framework. Therefore, the proposed method provides the whole clue to sampling, pre-processing, storing, transmitting and decompressing, which is helpful to improve CS data processing framework.

Acknowledgments

This work was supported by the NSFC of China under grant 61307121.

References

[1] Candès 2006 Compressive sampling *In Proceeding of the Int. Congress of Mathematics* 1433-52

[2] Baraniuk R 2007 Compressive sensing *IEEE Signal Processing Magazine* **52**(4) 118-21

[3] Candès E, Wakin M 2008 An introduction to compressive sampling *IEEE Signal Processing Magazine* **25**(2) 21-30

[4] Lustig M, Donoho D, Santos J, Pauly J 2008 Compressed sensing MRI *IEEE Signal Processing Magazine* **25**(2) 72-82

[5] Donoho D 2006 Compressed sensing *IEEE Trans. on Information Theory* **52**(4) 1289-306

[6] Candes E, Romberg J, Tao T 2006 Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. on Information Theory* **52**(12) 5406-25

[7] Ward R 2009 Compressed sensing with cross validation *IEEE Trans. on Information Theory* **55**(12) 5773-82

[8] Haupt J, Nowak R 2006 Signal reconstruction from noisy random projections *IEEE Trans. on Information Theory* **52**(9) 4036-48

[9] Song X X, Shi G M 2013 Fewer Bernoulli measurements satisfying the constraint of reconstruction probability *Aata Automatica Sinica* **39**(1) 53-6 *(In Chinese)*

[10] Song X X, Shi G M 2012 The low-redundancy compressed sensing measurements *Journal of Xidian University* **39**(4) 144-8 *(In Chinese)*

[11] Malioutov D M, Sanghavi S R, Willsky A S 2010 Sequential compressed sensing *IEEE J. Sel. Top. Sig. Proc* **4**(2) 435-44

[12] Malioutov D M, Çetin M, Willsky A S 2005 Homotopy continuation for sparse signal representation I*n Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing* 733-6

[13] Donoho D, Tsaig Y 2008 Fast solution of l1-norm minimization problems when the solution may be sparse *IEEE Trans. on Information Theory* **54**(11) 4789-812

[14] Asif M S, Romberg J 2010 Dynamic updating for l1 minimization *IEEE J. Sel. Top. Sig. Proc* **4**(2) 435-44

Authors

Song Xiaoxia, born in May, 1975, Datong, China

Current position, grades: Associate professor
University studies: Shanxi Datong University
Scientific interest: Compressed sensing, signal processing, wireless sensor networks
Publications: more than 10.
Experience: She is an associate professor in Shanxi Datong University. She received her B.S. degree in computer science in 1998 from Yanshan University. And she earned her M.S. degree in computer software and theory in 2005 from Guangxi Normal University. She has earned her Ph.D. degree in intelligent information processing at Xidian University in 2013.

Detection of WCDMA uplink signal with combination between sliding match and power spectrum

Xiaoping Wang*, **Jin Yao, Wangang Wang**

Chongqing City Management College, Chongqing, China

Received 25 March 2014, www.tsi.lv

Abstract

Aiming at problem of WCDMA uplink signal being difficult to be detected under low SNR, this paper proposes a type of algorithm in which sliding match combines with power spectrum to detect WCDMA signal. Firstly, this algorithm estimates desynchronizing point of signal using Frobenius norm. According to desynchronizing point, a whole cycle of information sequence is intercepted. Correlation of OVSF code sequence is utilized in which residual carrier or DC component of signal would come into being while the received OVSF code sequence completely matches with local OVSF code sequence. Then its power spectrum is calculated and sharp spectral peak would appear in the frequency position of power spectrum. Through detecting amplitude and position of spectral peak, frequencies of OVSF code sequence and residual carrier utilized in WCDMA signal could be accurately estimated. Simulation results show that this algorithm rapidly realizes the estimation on OVSF code sequence and desynchronizing point keeps good detection effect and may effectively overcome the influences of residual frequency offset on it.

Keywords: WCDMA signal, OVSF code sequence, Frobenius norm, sliding match, power spectrum

1 Introduction

In WCDMA [1] System, OVSF code is applied to do spread spectrum on data which belongs to one type of pseudo code sequences. Therefore, detection and estimation of OVSF code sequence are studied according to properties of pseudo code. In spread spectrum communication system, many estimation methods have been proposed at home and abroad nowadays in allusion to do detection and estimation on pseudo code under low SNR including energy detection [2], frequency doubler [3], cyclic spectrum [4], fourth-order cumulant [5], delay multiplication [6], etc. Literature [7] raises a type of method to do tracking on binary carrier modulation signal with residual frequency offset. Literature [8] puts forward a kind of stable symbol period of spread spectrum signal and an algorithm of desynchronizing point blind estimation. However, all of the above-mentioned blind estimation literatures of spread spectrum sequence on record are aimed at general DS system. In terms of blind estimation of WCDMA signal, Literature [9] comes up with signal model of WCDMA physical layer. Literature [10] does channel estimation under the circumstance of OVSF code sequence being known. Literature [11] proposes a WCDMA power control algorithm of low complexity. If estimation and blind estimation could be done on subtle features of WCDMA signal with residual frequency offset under low SNR, this type of weak signal would present great significance on administration and reconnaissance in aspects of civil use, military, software radio, intelligence communication, etc.

Aiming at characteristics like favourable autocorrelation of spreading code OVSF code sequence used by WCDMA signal, this paper utilizes a method in which sliding match combines with power spectrum. This method firstly takes advantage of Frobenius norm to estimate desynchronizing point of the received signal thus intercepting a whole cycle of data after orderly moving it in length of desynchronizing point and then leads it to do sliding match and FFT operation with local OVSF code sequence. When they are completely matched, sharp power spectral peak and sine and cosine signal reflecting residual frequency offset would appear thus rapidly catching and estimating OVSF code sequence. Meanwhile estimated value of residual frequency offset could be acquired, which provides solid foundation for subsequent dispreading processing.

2 Signal Model of WCDMA Uplink

In WCDMA, dedicated physical channels for uplink include DPDCH (Dedicated Physical Data Channel) and DPCCH (Dedicated Physical Control Channel) in which DPDCH is used to support data and DPCCH is used to bear control information. In WCDMA channel, transmission of data takes slot as cell in which Tslot is 0.625ms and each 15 slots form one frame. Each slot of DPCCH is 10 bit. Bit number of each slot of DPDCH has relationship with SF (spreading factor). Expression of sending and receiving model of WCDMA signal is:

* ***Corresponding author*** e-mail: workmail73@126.com

$$y(t)=A\left[d_I(t)c_I(t)s_I(t)-d_Q(t)c_Q(t)s_Q(t)\right] \\ h(t)\cos(\omega t)+j*A\left[d_I(t)c_I(t)s_Q(t)+\right. \; . \\ \left. d_Q(t)c_Q(t)s_I(t)\right]h(t)\sin(\omega t) \tag{1}$$

Here A is amplitude of signal. $d_I(t)$ and $d_Q(t)$ are respective input signals of DPDCH and DPCCH. $c_I(t)$ and $c_Q(t)$ are their corresponding OVSF codes. Channelization codes used in each channel differ from each other and are orthogonal independent. $s_I(t)$ and $s_Q(t)$ are respectively real part and imaginary part of complex-scrambled code which are orthogonal independent.

$h(t)=\dfrac{\sin\left[\pi(1-\alpha)t/T\right]+(4\alpha t/T)\cos\left[\pi(1+\alpha)t/T\right]}{(\pi t/T)\left[1-(4\alpha t/T)^2\right]}$ is root-raised cosine filter whose roll-off factor is $\alpha=0.22$; $\sin(\omega t)$ and $\cos(\omega t)$ are modulated carriers; $n(t)$ is white Guassian noise whose mean value is zero and variance is σ_n^2. The above-mentioned OVSF code sequence is used to keep orthogonality among different physical channels of user thus assuring that different operations could be better transmitted.

3 Detection and Estimation of OVSF Code Sequence of WCDMA Signal

3.1 DESYNCHRONIZING POINT ESTIMATION OF OVSF CODE SEQUENCE

Utilize Frobenius norm to do blind estimation on desynchronizing point in which desynchronizing point receiving signal is effectively estimated which provides good security for subsequent detection and estimation of OVSF code. Relevant matrix of real data $r_I(n)=d_I(n)+n_0(n)$ receiving WCDMA signal is:

$$R_{r_I}(\tau)=E\left\{r_I(n)r_I(n+\tau)^H\right\}=\sigma_I^2\sum_{k=1}^{K}\left[d_{Ik}(\tau_k)d_{Ik}{}^H(\tau_k)\right]+\sigma_v^2 I\,. \tag{2}$$

Among them σ_I^2 is variance of information code, σ_v^2 is variance of noise and H stands for conjugate transpose.

Do eigenvalue decomposition on Formula (2) above and acquire that:

$$R_x(\tau)=\frac{\sigma_b^2 N_c}{T_s}\sum_{k=1}^{K}\left[(T_s-\tau_k)u_k^r u_k^{rH}+\tau_k u_k^l u_k^{lH}\right]+\sigma_v^2 I\,. \tag{3}$$

In the formula $U=\left[u_1^r u_1^l \ldots u_k^r u_k^l\right]$ is normalized unit matrix. Then eigenvalue of $R_x(\tau)$ is:

$$\begin{cases}\lambda_{2i-1}=\sigma_b^2 N_c/T_s(T_s-\tau_i)+\sigma_v^2 \\ \lambda_{2i}=\sigma_b^2 N_c/T_s\tau_i+\sigma_v^2\end{cases}. \tag{4}$$

Here comes $i=1,2,\cdots,K$. When $i>2K$, $\lambda_i=\sigma_v^2$. Therefore, desynchronizing point $\hat{\tau}_k$ could be estimated based on relatively larger eigenvalue of relevant matrix receiving signal.

Frobenius norm square of relevant matrix actually corresponds to the sum of square of function eigenvalue namely:

$$\left\|R_x(\tau)\right\|_F^2=\sum_{i=1}^{N_c}\lambda_i^2(\tau). \tag{5}$$

$\{\lambda_i, i\geq 3\}$ does not depends on desynchronizing point. It is obvious that $\left\|R_x(\tau)\right\|^2$ reaches the maximum when $(\lambda_{2i-1})^2+(\lambda_{2i})^2$ acquires its maximum value. It is obtained through Formula (4) that:

$$\lambda_{2i-1}+\lambda_{2i}=c\,. \tag{6}$$

Here c is a constant, which does not depends on desynchronizing point. Therefore:

$$(\lambda_{2i-1})^2+(\lambda_{2i})^2=2(\lambda_{2i-1})^2-2c\lambda_{2i-1}+c^2 \\ =(4\lambda_{2i-1}-2c)^2+c^2/2 \; . \tag{7}$$

Suppose that eigenvalues sort according to descending order. As $\lambda_{2i-1}\geq c/2$, the formula above is always positive value (otherwise $\lambda_{2i}>\lambda_{2i-1}$ in Formula (6)). Therefore, $\left\|R_x(\tau)\right\|_F^2$ is increasing function of λ_{2i-1} meaning that maximizing λ_{2i-1} is just maximizing $\left\|R_x(\tau)\right\|_F^2$.

It is known from the definition of Frobenius norm that Frobenius norm of one matrix is equal to arithmetic square root of quadratic sum of all elements in this matrix. Therefore, it is much easier to calculate $\left\|R_x(\tau)\right\|_F^2$ than calculating eigenvalue. Desynchronizing point is estimated to be:

$$\hat{\tau}_k=\arg\max_{\tau}\left(\left\|R_x(\tau)\right\|_F^2\right). \tag{8}$$

3.2 DETECTION AND ESTIMATION OF OVSF CODE SEQUENCE

In WCDMA system, the received baseband signal would keep certain residual frequency offset because of the frequency deviation and Doppler shift among transceivers. On the receiving end, do down-conversion simulation on the received WCDMA signal and then analog-to-digital conversion in which sample rate equals to chip rate namely $f_s = 1/T_c$. Here T_c is the duration of each chip. The received WCDMA signal changes to number with residual frequency offset. Complex sinusoidal signal is presented as:

$$r(i) = A\left(d_I(i) + j*d_Q(i)\right)\exp\left[j\left(2\pi\triangle f\, iT_c + \varphi\right)\right] + n_0(i). \quad (9)$$

Here A stands for amplitude attenuation value of signal. $d_I(i)$ and $d_Q(i)$ are discrete amplitude values of real part and imaginary part of the received WCDMA signal. $\triangle f$ is the size of residual frequency offset. φ is the initial phase. $n_0(i)$ is white Guassian noise whose average value is zero and variance is σ_0^2.

Move the received signal $r(i)$ backward $\hat{\tau}_k$ in order according to the estimated size of desynchronizing point and intercept a whole cycle of WCDMA signal data to be:

$$\begin{aligned} r_s(k) &= A\left(d_I(k) + j*d_Q(k)\right) \\ &\exp\left[j\left(2\pi\triangle f\, kT_c + \varphi\right)\right] + n_0(k)\left(1 \le k \le R\right) \end{aligned}. \quad (10)$$

Here R is the size of OVSF code sequence cycle. When cycle of OVSF code is known, utilize correlator to do correlation operation on $r_s(k)$ and local OVSF code sequence $s^{'}\left(i+i^{'}\right)$ among which:

$$s^{'}(i) = \sum_{i=1}^{R*R}\left(d_s(i) + j*d_s(i)\right). \quad (11)$$

Both lengths of correlator and the intercepted WCDMA data are R. $i^{'}$ is duration between local OVSF sequence and the intercepted data in the received signal. In order to simplify the analysis, make it to be $A = 1$, in which useful signal part is $s_l(k)$ and noise part may be presented as $n_s(k)$ in $x_l(k)$ if output of the l th sliding match is $x_l(k)$.

$$\begin{aligned} x_l(k) &= r_s(k)s^{'}(l*R)^{*} \\ &= \left(d_I(k) + j*d_Q(k)\right)\exp\left[j\left(2\pi\triangle fkT_c + \varphi\right)\right] \\ &\left(d_{si}(l*R) - j*d_{sq}(l*R)\right) + \\ &n_0(k)\left(d_{si}(l*R) + j*d_{sq}(l*R)\right) \\ &= R_l(k)\exp\left[j\left(2\pi\triangle fk + \varphi\right)\right] + n_l(k) \end{aligned}. \quad (12)$$

Among those $s^{'}(l*R)^{*}$ stands for conjugation of local OVSF code sequence $s^{'}(l*R)$ and $R_l(i)$ stands for correlation value of the intercepted data after correlation operation. Make that:

$$s_l(k) = R_l(k)\exp\left[j\left(2\pi\triangle fk + \varphi\right)\right], \quad (13)$$

$$n_l(k) = n_0(k)\left(d_{si}(l*R) - j*d_{sq}(l*R)\right). \quad (14)$$

Among them:

$$\begin{aligned} R_l(k) &= \left(d_I(k) + j*d_Q(k)\right)\left(d_{si}(l*R) - j*d_{sq}(l*R)\right) \\ &= d_I(k)*d_s(l*R) + j*\left(d_I(k)*d_{sq}(l*R)\right. \\ &\left.-d_Q(k)*d_{si}(l*R)\right) + d_Q(k)*d_{sq}(l*R) \end{aligned}. \quad (15)$$

As OVSF code sequence keeps orthogonality, then:

$$\frac{1}{R}\sum_{\text{k=1}}^{R} d_i(k)*d_j(k) = \begin{cases} 1 & \text{i=j} \\ 0 & \text{i} \ne \text{j} \end{cases}. \quad (16)$$

So when the intercepted data and local OVSF code sequence just match well, $R_l(k) = d_I^2(k) + d_Q^2(k)$. Then comes:

$$\begin{aligned} &x_l(k) \\ &= \left(d_I^2(k) + d_Q^2(k)\right)*\exp\left[j\left(2\pi\triangle fk + \varphi\right)\right] + n_l(k) \end{aligned}. \quad (17)$$

It is seen from Formula (17) that only complex signal with residual frequency offset remains in this formula when they match well.

If the way of vector is utilized, output signal calculated through correlation operation could be expressed as:

$$\mathbf{x}_l = \mathbf{s}_l + \mathbf{n}_l. \quad (18)$$

As noise in the received signal conforms to Gaussian distribution whose mean value is zero and variance is σ_0^2, it is known from central-limit theorem that output noises passing correlator also keep Gaussian distribution which are unrelated if length of correlator is large enough. There exists complex sinusoidal random signal of ω_D ($\omega_D = 2\pi\triangle f$) in data vector $\mathbf{x}_l$ whose probability density function can be presented as:

$$\begin{aligned} &p\left(\mathbf{x}_l - \mathbf{s}_l\right) \\ &= \frac{1}{\left(\pi l\sigma^2\right)^l}\exp\left[-\frac{1}{l\sigma^2}\left(\mathbf{x}_l - \mathbf{s}_l\right)^H\left(\mathbf{x}_l - \mathbf{s}_l\right)\right] \end{aligned}. \quad (19)$$

In the formula H stands for conjugate dispose, Directly does Fourier transform on observation data $\mathbf{x}_l$ and act square of Fourier transform output value as estimated value $\hat{S}(\omega)=\frac{1}{l}\left|\sum_{n=0}^{l-1}\mathbf{s}_l(i)\exp(-j\omega_D i)\right|^2$ of power spectrum density. In terms of complex sinusoidal signal $\mathbf{s}_l(i)$ with complex white Guassian noise, its maximum likelihood solution $\hat{\omega}_D$ is the corresponding frequency to the maximum value of power spectrum. When the intercepted data in WCDMA signal and local OVSF code sequence just align, maximum value of output amplitude would appear on frequency point $\hat{\omega}_D$ after FFT. Therefore, estimated value $\triangle\hat{f}$ of residual frequency offset would be acquired while intercepting OVSF sequence through Formula $\omega_D = 2\pi l \triangle f T_c$ and according to the maximum value of power spectrum peak.

3.3 ANALYSIA ON COMPUTATIONAL COMPLEXITY

Aiming at the proposed algorithm of OVSF code blind estimation in this paper, multiplication serves as the index to measure complexity. Suppose that length of the received signal is M. Desynchronizing point estimation through F- norm needs 2M times of multiplication. Sliding match and power spectrum calculation mainly include corresponding bit multiplication and FFT transform in which corresponding bit multiplication needs M times of multiplication, FFT transform needs R*Mlb(100M) times of multiplication and module-square on its result needs 2M times of multiplication whose total number is 5M+R*Mlb(100M).

4 Algorithm Flow

Figure 1 shows the structure diagram realizing the above-mentioned algorithm:

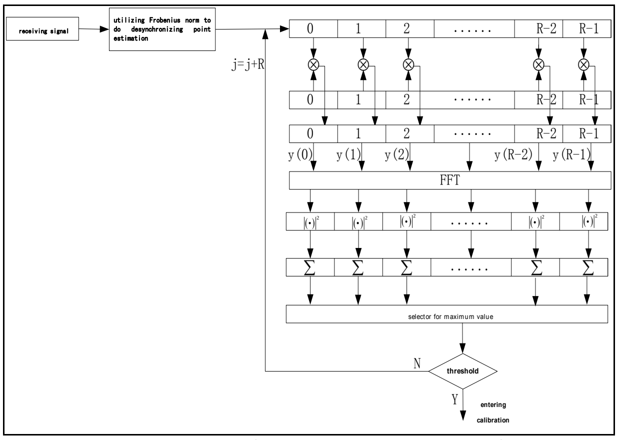


FIGURE 1 Block Diagram of Detection and Estimation of OVSF Code Sequence

This algorithm could be summarized as follows:

(1) After doing down-conversion simulation on the received WCDMA signal, do sampling on it in the rate of chip in which data of real part is chosen to do desynchronizing point estimation utilizing Frobenius norm.

(2) Orderly move the received WCDMA data backward desynchronizing-point data and intercept part of data in the length of a whole cycle of OVSF code sequence.

(3) Do sliding match calculation on the intercepted data and local OVSF code sequence. Each time choose R point from local OVSF code sequence to do its multiplication with the intercepted data.

(4) Do R point FFT calculation on the result of correlation operation, choose square of its module and orderly move local OVSF code sequence backward R chips namely $j = j + R$.

(5) Compare the maximum value with threshold value, which are chosen by maximum selector. If it surpasses the presupposed threshold value, the intercepted OVSF code sequence successfully match with local OVSF code sequence. Detect OVSF code sequence of WCDMA signal. Through detecting the position of maximum peak, OVSF code sequence used at transmitting end would be estimated.

5 Simulation Results

Experiment one: Utilize algorithm in this paper to do simulation experiment on OVSF code sequence of WCDMA signal in which SNR is -5dB, length of OVSF code SF is 256 and chip rate is 3.84Mchips/s. Apply RRC filter whose roll-off factor is $\alpha = 0.22$. Delay point of receiving signal is 66 in which eight times of sampling is done on the received signal and F- norm is used to do desynchronizing point estimation on it as shown in the following Figure 2:

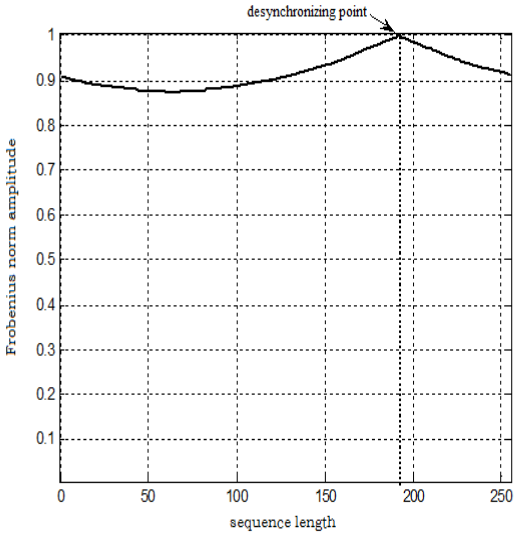


FIGURE 2 Blind Estimation of Desynchronizing Point

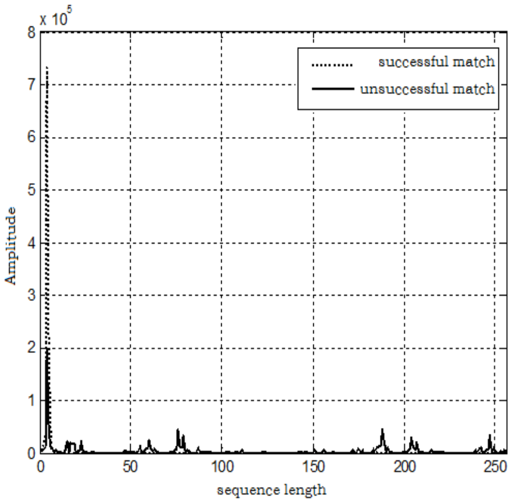
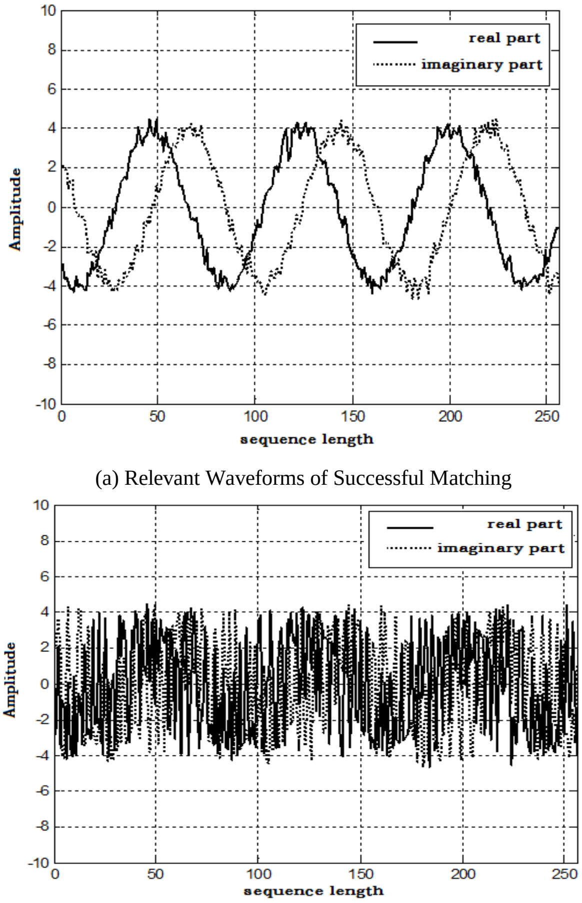


FIGURE 3 Power Spectrums of Different Matching Stages

Figure 2 is the simulation diagram in which desynchronizing point is estimated on the basis of Frobenius norm. It is seen from Figure 2 that F- norm amplitude reaches the maximum in the position of desynchronizing point. Through the position of maximum amplitude value, number of desynchronizing point of WCDMA signal could be estimated to be 192, which is 256-66+2 thus being well prepared for subsequent detection and estimation of OVSF code.

Figure 3 is power spectrum diagram to determine whether matching is successful when length of OVSF code is 256. It is seen from Fig.3 that sharp power spectral peak would appear in power spectrum when correlation matching is successful between the intercepted WCDMA signal and local OVSF code sequence. While during sliding match processes of other positions meaning that matching is unsuccessful, power spectral peak is relatively lower. OVSF code sequence utilized at transmitting end would be estimated through detecting the position of sharp power spectral peak.

Figure 4 are simulation diagrams of relevant waveforms of successful matching and unsuccessful matching. Because of influences of residual frequency offset, related complex sine and cosine signal would be acquired as presented in Figure 4(a) at the time intercepted WCDMA data and local OVSF code sequence successfully match. However, related complex sine and cosine signal would not be acquired as seen in Figure 4(b) if they do not match well. Therefore according to whether phenomenon of relevant waveforms comes into being, it would be determined that whether intercepted WCDMA data successfully matches with local OVSF code sequence. Also size of relevant residual frequency offset value $\triangle \hat{f}$ could be estimated.



(a) Relevant Waveforms of Successful Matching

(b) Relevant Waveforms of Unsuccessful Matching

FIGURE 4 Relevant Waveforms of Each Matching Stage

Experiment two: The following one is performance curve simulation experiment of detection and estimation on OVSF code.

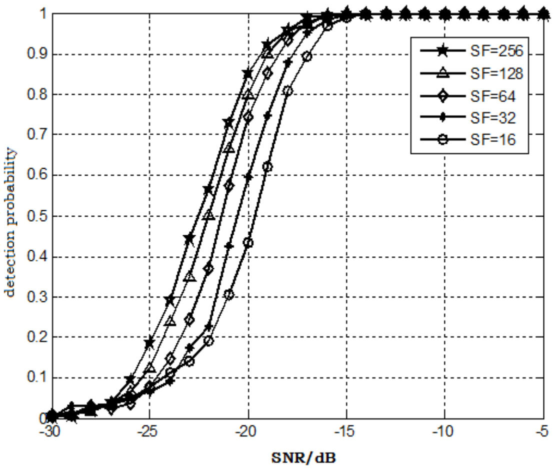


FIGURE 5 Influences of OVSF Code Sequences with Different Lengths on Desynchronizing Point Estimation

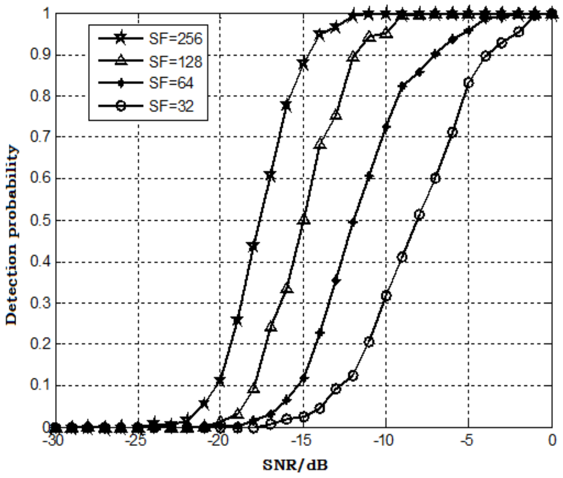


Figure 6 Influences of Different Lengths of OVSF Code Sequences on Detection Performances

Figure 5 shows performance curves of desynchronizing point estimation of different lengths of OVSF code sequences under different types of SNR. It is seen from Figure 5 that probability of detection on desynchronizing point increases with the enlarging of OVSF code length under the same SNR, in which detection probability would present corresponding improvement with increasing of SNR under any type of OVSF code length.

Figure 6 presents simulation curves of performance detection on OVSF code sequences of different lengths under different types of SNR. It is expressed in Fig.6 that capability of detection on different lengths of OVSF code sequences increases with the improving of SNR. Meanwhile probability of detection on OVSF code sequence also enlarges with the elongation of OVSF code under the same SNR.

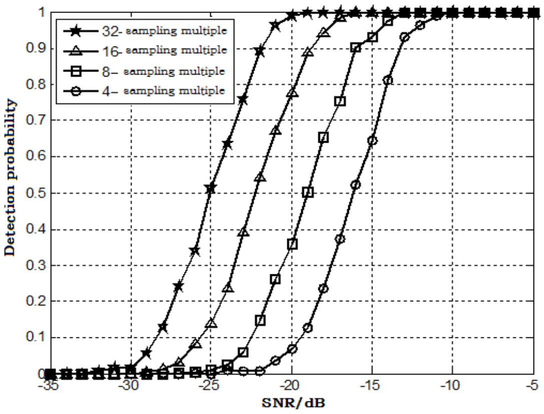


FIGURE 7 Influences of Different Sampling Numbers on Detection Performances

Figure 7 shows the influences on detection probability of OVSF code sequence under the condition of OVSF code length being 128 and under different sampling multiples of 4, 8, 16 and 32. It is known from Figure 7 that detection probability gradually raises with the increasing of sampling multiple under the same SNR. Under the same detection probability, SNR reduces about 3dB when sampling multiple changes from 8 to 16.

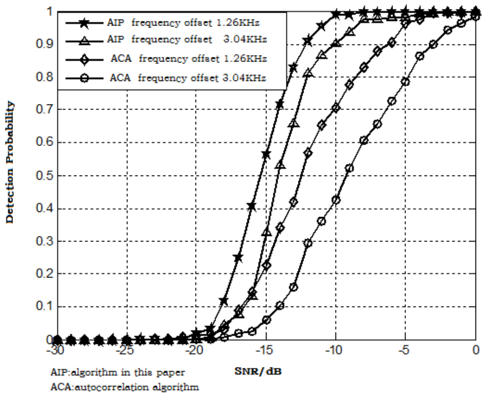


FIGURE 8 Influences of Residual Frequency Offset on Detection Performances

Figure 8 shows the influences of residual frequency offset on detection probability under different types of SNR. It is known from Figure 8 that related loss of matching operation increases and detection probability gradually decreases with the enlargement of residual frequency offset. Under the same type of condition, compare the methods of autocorrelation algorithm to detect OVSF code sequence. Frequency offset keeps relatively higher influences on performance detection utilizing autocorrelation algorithm. In addition it is presented in the figure that algorithm in this paper is better in performance detection than autocorrelation algorithm.

6 Conclusions

Under low SNR, what should be firstly determined are blind synchronization and spreading code of receiving signal namely blind estimation of OVSF code sequence so that detection on WCDMA signal would be realized. Through analysing relevant matrix structure of F- norm and OVSF code sequence in WCDMA signal, this paper firstly acquires desynchronizing point of WCDMA signal according to F- norm estimation and secondly does sliding match and power spectrum calculation. As OVSF code sequence keeps good orthogonal independence, sharp power spectrum peak would be acquired when the intercepted data matches well with local OVSF code sequence. Through size of spectral peak and position of it, OVSF code sequence used at transmitting end could be detected and estimated thus realizing estimation on different operations in WCDMA uplink signal. Simulation results present that this algorithm keeps favourable detection and estimation effects, which could be better when OVSF code sequence is longer.

Acknowledgements

This work was supported by the Projects of Scientific and Technological Research Program of Chongqing Municipal Education Commission (Grant No. No.KJ131705, No.KJ121706).

References

[1] Zhang Chuanfu, Lu Huibin, et al. 2009 *The third generation mobile communication* Beijing: Publishing House of Electronics Industry Press *(In Chinese)*

[2] Urkowitz H 1967 Energy detection of unknown deterministic signals *Proceeding of IEEE* **55**(4) 523-31

[3] Hill D, Bodie J B 2001 Carrier detection of PSK dignals *IEEE Trans on Commu* **49**(3) 487-96

[4] Guo Xiaofang, Li Yi, Li Feng 2010 A research on dection of DSSS signal based on improved cyclic spectrum *Science Technology and Engineering* **10**(32) 7937-41 *(In Chinese)*

[5] Zhao Zhijin, Pu Junjie 2009 A detection method of DS-CDMA signal based on the quadratic fourth-order moment chip *NSWCTC* **2** 759-62 *(In Chinese)*

[6] Yuan Liang, Liu Jinan, Wen Zhijin 2005 A detection method for DS/SS signal in the low SNR condition *Modern Electronics Technique* **196**(5) 50-1 *(In Chinese)*

[7] Yao Zheng, Lu Mingquan, Feng Zhenming 2009 Unambiguous technique for multiplexed binary offset carrier modulated signals tracking *Signal Processing Letters* **16**(7) 608-11 *(In Chinese)*

[8] Bouder C, Azou S, Burel G 2002 A robust synchronization procedure for blind estimation of the symbol period and the timing offset in spread spectrum transmissions *Proc of the 7th International Symposium on Spread- Spectrum Techniques and Applications* 233-41

[9] Zhang Qiang, Wan Min, Liu Kewei 2008 Study and Simulation of WCDMA Uplink Channel's Spread Spectrum and Modulation *Research and Exploration in Laboratory* **27**(11) 49-50 *(In Chinese)*

[10] Shi Jingjing, Du Shuanyi, Yao Pei 2008 Study and simulation of channel's estimation in WCDMA *Electronic Science and Technology* **21**(9) 43-5 *(In Chinese)*

[11] Xu Wenjun, Liu Dechang, He Zhiqiang, et al. 2007 A low computational complexity simulation algorithm for power control in WCDMA system *Acta Metallurgica Sinica* **30**(4) 93-7 *(In Chinese)*

[12] Yu Cheng, Zhan Fei 2003 *WCDMA system physical layer design* Beijing: Posts&Telecom Press *(In Chinese)*

[13] Bazil Taha Ahmed, Miguel Calvo Ramon 2011 WCDMA multiservice uplink capacity of highways cigar-shaped microcells with adjacent channel interference *European transactions on telecommunications* **22**(6) 322-31

[14] Bazil Taha Ahmed, Miguel Calvo Ramon 2012 Multiservice capacity and interference statistics of the uplink of high altitude platforms (HAPs) for asynchronous and synchronous WCDMA system *Annals of telecommunications* **67**(9/10) 503-9

[15] 3GPP TS 25.213, Spreading and modulation (FDD) 3Gpp.org, 2004

Authors

	Xiaoping Wang, born on February 13,1973, Langzhong, Sichuan, China **Current position, grades:** South 2th Road NO.151, Huxi University Town, Chongqing, China, associate professor **University studies:** postgraduate **Scientific interest:** engaged in researches on wireless communication, wireless sensor network, embedded development, etc. **Publications:** 8 **Experience:** 2005-2007, Engineer of ZTE Corporation. 2007-2008,teacher of Chongqing city management college
	Yao Jin **Current position, grades:** Teacher of Chongqing City Management College since 2012.07 **University studies:** Major in Communication and Information system, Master degree and graduated from Southwest Jiaotong University. **Scientific interest:** Have rice experience of Modelling & Simulation of Communication Systems and familiar with Wireless Sensor Networks. **Publications:** 5 **Experience:** 2011.01-2012.07, Engineer of ZTE Corporation. 2008.08-2010.12, Engineer of Huawei Technologies Corporation.
	WanGang Wang, born on May 16, 1977, Kaixian, Chongqing, China **Current position, grades:** South 2th Road NO.151, Huxi University Town, Chongqing, China, associate professor **University studies:** postgraduate **Scientific interest:** Electronic Information Technology **Publications:** 11 **Experience:** 1999-,teacher of Chongqing city management college

Multidisciplinary design optimization of complex products based on data fusion and agent model

Lei Li[1, 2*], **Jianrun Zhang**[2]

[1] *Jiangsu University of Science and Technology, Zhenjiang 212000, China*

[2] *Southeast University, Nanjing 211189, China*

Received 1 March 2014, www.tsi.lv

Abstract

Multidisciplinary design optimization (MDO) of complex products is discussed in this article. For the characteristics of higher order, high-dimensional, multi-input and multi-output in design of complex products, application of MDO in design and optimization of complex products is difficult. An effective MDO framework combined with the method of data fusion and agent model is proposed. Firstly, date fusion is applied to deal with the process with a large number of incomplete, vague and uncertain in complex product's evaluation and optimization; secondly, agent model is used to reduce the complexity of the MDO model; and finally, MDO is applied to complex products design and optimization according to the collaborative design and optimization method. In order to identify the feasibility of this method, the design of diesel engine motion mechanism is discussed and shown a good result. The current study provides a powerful tool for complex products designing and optimization and owns great theory and practical values.

Keywords: Complex Products, MDO, Date Fusion, Agent model

1 Introduction

Complex products are composed by a number of associated or interacted components (factors). Design of complex products involves different coupled disciplines and large numbers of design variables, which lead to multi-inputs and multi-outputs in the design process. The design freedom of complex products is greatly reduced while the field of knowledge involved in the design process increase [1]. As shown in Figure 1, the stage of conceptual design owns the highest degree of freedom for there is less knowledge involved; along with the design stage increased, the design freedom will rapid decreased for there are more and more design knowledge and involved knowledge.

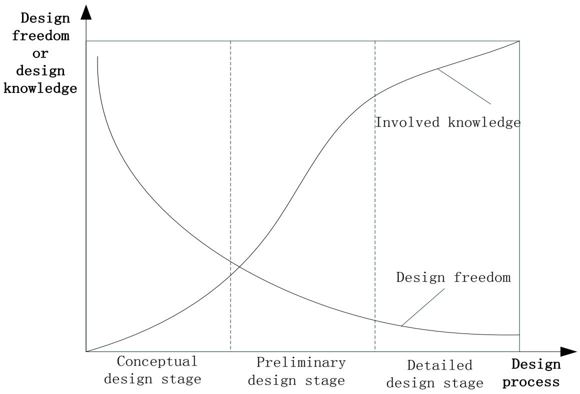


FIGURE 1 Design process of complex products

In order to reduce the difficulty of product development and obtain the best performance of complex products, it is important to consider the coupled performance in various disciplines and discover the potential information in the design process. Traditional design optimization method neglects the coupling of different disciplines and cannot deal with large number of design variables and often leads to the failure of design process, and then multidisciplinary design optimization (MDO) method is developed [2].

MDO takes advantage of the interactions between disciplines as well as to improve the product development time and has emerged as a new technology dealing with the design of complex systems. MDO were applied primarily to design of aeronautics, astronautics and automobile, and the techniques have been in development over the last decade [3, 4], including: (1) MDO theory and algorithms research [5, 6], including system modelling and decomposition, optimization algorithms and the methods of space design searching; (2) Methods of multidisciplinary analysing [7, 8], including mathematic model, sensitivity analysis, design of experiment, agent model and so on; (3) Research of software integration framework for MDO based on the in-depth study of MDO theory and methods.

However, for the characteristics of higher order, high dimensional, multi-input and multi-output in complex systems, application of MDO in design and optimization of complex products is difficult. The purpose of this article is presenting an effective framework for the design of complex products combined with the method of data

[*] ***Corresponding author*** e-mail: lilei0064@sina.com

fusion, agent model and MDO. Firstly, date fusion is applied to deal with the process with a large number of incomplete, vague and uncertain in complex product's evaluation and optimization; secondly, agent model is used to reduce the complexity of the MDO model; and finally, MDO is applied to complex product design and optimization according to the collaborative design and optimization method. In order to identify the feasibility of this method, the design of diesel engine motion mechanism is discussed and shown a good result.

2 Principles of data fusion for complex systems

Data fusion was firstly applied in the field of military. According to the statistics, 54 data fusion systems in the military electronic systems have been applied in the United States up to 1991, and 87% have been used for experimental prototype, which proved to be usefulness. Oregon State Science and Technology Research Institute have carried out for research and discussion of the theory and application of a wide range of data fusion; New York State University has set up a multi-source information fusion centre for the research of fusion framework; and British BAE System company has developed a new technology which is called distributed data fusion and integration (Decentralized Data Fusion, DDF). In recent years, the technology of data fusion is greatly developed, and the research has been applied in automatic control, target identification, traffic control, process monitoring, navigation, repair of complex machine and robot, and so on.

2.1 DEFINITION AND THEORY OF DATA FUSION

Data fusion is an information process used for decision-making and estimation according to certain criterion and the information obtained by several sensors. The objective of data fusion is to derive more information by data combination and synergism and improve the effectiveness of the sensor system. Currently, data fusion has been a combination of many traditional disciplines and emerging multi-disciplinary engineering. Because of the highly development and mutual penetration of these disciplines and fields, data fusion methods exhibit the characteristics of diversity and pluralism. The general data fusion technologies include association analysis, judgment or detection theory and estimation theory. Association analysis is a method can be used for mining hidden implicit relationship between the data appear to be unrelated. Data fusion is a great technology dealing with the MDO problem of complex products, which involves the characteristics of higher order, high dimensional, multi-input and multi-output.

2.2 DATA FUSION METHOD BASED ON ASSOCIATION ANALYSIS

The process is described as follow:

(1) Questionnaire module: gathering the feedbacks of design parameters from the expert and providing data for optimal design. The experts evaluate the design parameters with scores ranging from 1 to 10 which represent the importance of the parameters;

(2) Database or data warehouse: saving feedback information collected from the questionnaires;

(3) Normalize of the design parameters: The normalized of the design parameters can be expressed as follows:

$$y=(x\text{-Min Value})/ (\text{Max Value-Min Value}), \quad (1)$$

where: x, y are the values before and after conversion respectively; Max Value, Min Value are the maximum and minimum of the samples respectively.

(4) Evaluation of the weight of the parameters: determine the weight of every design parameters.

(5) Information fusion: checking and correcting the reduction's parameters according to the expert's evaluation. In order to verify the validity of the model, the correlation coefficient and the target residuals need to be obtained. Correlation coefficient, also known as linear correlation coefficient, is an indicator measuring the linear correlation between the variables. The target residual is the difference between the observed and predicted values, that is, the difference between the actual observations and regression estimates.

(6) User Interface: show the optimized results to the designer.

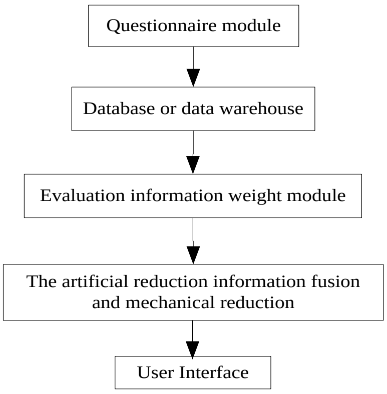

Questionnaire module
Database or data warehouse
Evaluation information weight module
The artificial reduction information fusion and mechanical reduction
User Interface

FIGURE 2 Data fusion-based evaluation and optimization architecture

3 Agent Model for complex products

The efficiency of MDO for complex products is seriously affected by large number of optimization variables and large scale of analysis model. Agent model is useful in reducing the analysis time. As shown in Figure 3, the agent model constructs math function in design space to express the relationship between the design variables and the system response. In the other word, agent model uses math function taking the place of simulation analysis. Because the math function is much simpler than simulation model, it can sharply cut down the design and analysis time.

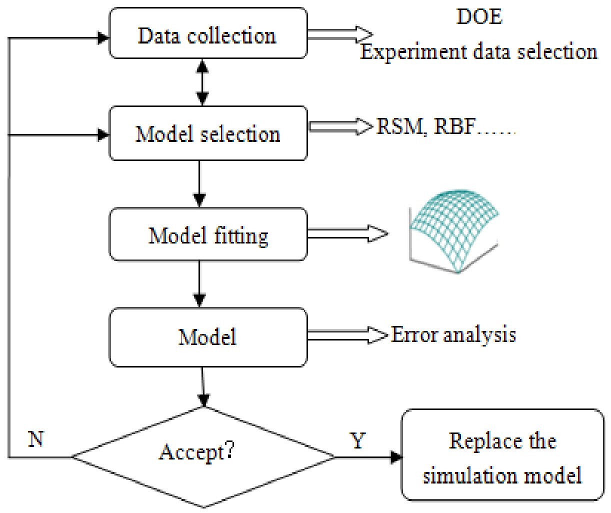


FIGURE 3 Design flow of MDO based agent model [10]

The common agent models are response surface model (RSM), Radial basis function neural network model (RBS) and so on. RSM is based on the knowledge of statistics and mathematic and using simple mathematical expressions (commonly the lower level polynomial) take the place of actual analysis model. Two-level polynomial model is the most commonly used agent model and can be expressed as follow [11]:

$$\tilde{F}(X) = a_0 + \sum_{i=1}^{N} b_i x_i + \sum_{i=1}^{N} c_{ii} x_i^2 + \sum_{ij(i<j)} c_{ij} x_i x_j \ . \qquad (2)$$

Here N is the number of input various, x_i is the ith input various, a, b and c are the polynomial coefficients.

As an example, agent model for the function of $Y = X_1^3 + 4X_2$ is shown in Figure 4. Firstly, get samples by orthogonal experiment; secondly, establish the function of second-order response surface model.

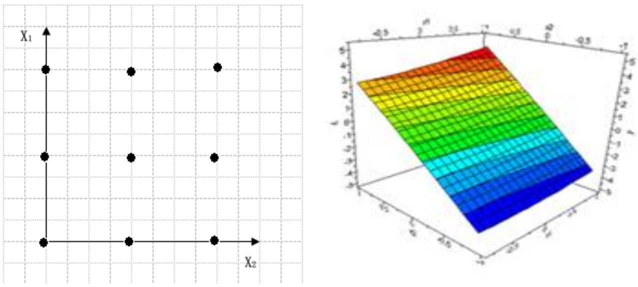


FIGURE 4 Construction Process of agent model formulation

And then, accuracy of agent model can be evaluated by the residual sum of squares:

$$R = \sum_{i=1}^{n} (F(X_i) \text{-} \tilde{F}(X_i))^2 \cdot \qquad (3)$$

4 MDO of complex products based on data fusion and agent model

In response to the large number of incomplete and uncertain reasoning process presented in the design and optimization process of the complex products, a theoretical framework for the multi-disciplinary design optimization of complex products is proposed this paper. Detail of the MDO framework combined with the method of data fusion and agent model is shown in Figure 5.

Basing the analysis of the MDO problem of complex products, the optimization objectives, constrains and variables can be extracted and large number of samples would be obtained based on the design of experiment (DOE). Then, data fusion algorithm is used to analyse the relationship between the objectives, constraints and design variables, and then the weight of each parameter can be obtained. In order to reduce the difficult of the MDO problem, the smaller weight parameters would be ignored. At the same time, data mining algorithm is often used for finding the potential information between the design parameters and objectives. At last, the agent model can be obtained using the method of response surface or neural networks to improve the MDO efficiency.

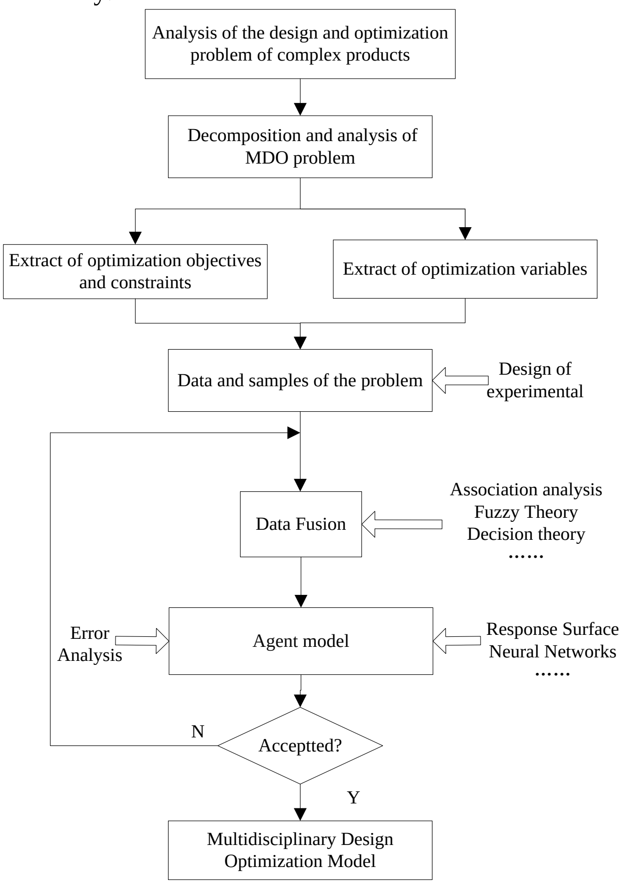

Analysis of the design and optimization problem of complex products
Decomposition and analysis of MDO problem
Extract of optimization objectives and constraints
Extract of optimization variables
Data and samples of the problem
Design of experimental
Data Fusion
Association analysis
Fuzzy Theory
Decision theory
......
Error Analysis
Agent model
Response Surface
Neural Networks
......
N
Acceptted?
Y
Multidisciplinary Design Optimization Model

FIGURE 5 Framework of Multidisciplinary design optimization for complex products

5 Case studies

The diesel engine is a typical complex product. Designing of various parts of the diesel engine affects its overall performance. However, because of the large number of design parameters and their potential relationship, the best performance of the diesel engine is hard to obtain. The diesel engine motion mechanism of the diesel engine is selected as the complex product to identify the feasibility of MDO framework proposed above.

5.1 DESCRIPTION OF THE PROBLEM [12]

It is a typical multidisciplinary design optimization problem, which comes to different disciplines such as lightweight, thermal, vibration, kinematic and et.al. In this paper, three-dimensional parametric model of crankshaft-connecting rod-piston is founded. The dynamic characteristics of crankshaft-connecting rod-piston system is simulated in the software of ADAMS(Figure 6), getting the largest load of the small head of the connecting rod, which will be set as the boundary conditions for further analysis. The performance of modal and thermal of the piston is analysed by software of ANSYS. At the same time, the structural strength of the connecting rod can be got. In the end, the MDO platform is founded in the soft of ISIGHT for getting the best comprehensive performance of the diesel engine system.

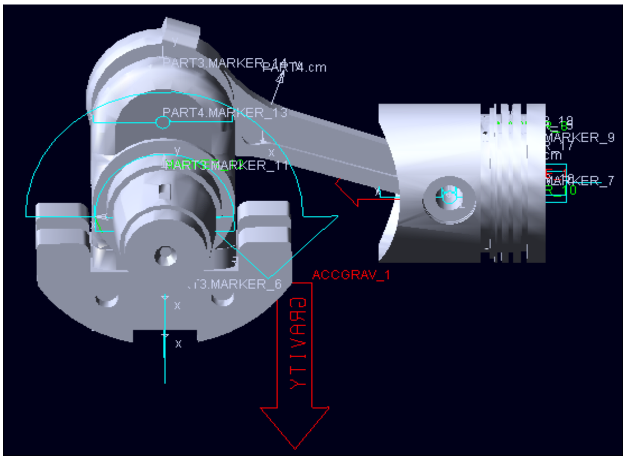


FIGURE 6 Diesel sports agency model

Table 1 lists the design parameters in the optimization of the diesel motion mechanism. D1, TK1…TK9, Hs_h and Hs_cs are design parameter. FEEQ1, FREQ1, FREQ3, STMAX and UMAX are design constraints and V is the objective function.

TABLE 1 the meaning of the link parameters (Unit: mm)

Variables	Initial value	Constraint	Remarks
TK1	27	25<TK1<29	Width of big ending of connecting rod
TK2	21.6	18<TK2<22	Width of small ending of connecting rod
TK3	15	13<TK3<17	Height of connecting rod side
TK4	8	6<TK4<10	Height of connecting rod notch
TK5	10	8<TK5<12	Transition radius of small ending of connecting rod
TK6	60	55<TK6<65	Transition radius of big ending of connecting rod
TK7	5	3<TK7<7	Transition radius of small end of connecting rod notch
TK8	7	5<TK8<9	Arc centre distance of big end of connecting rod notch
TK9	26	24<TK9<28	Centre distance of small end connecting rod notch to the centre of small end
Hs_h	8	6<Hs_h<10	Height of piston top shore
Hs_cs	5	3<Hs_cs<7	Depth of piston ring groove
D1	23	20<D1<26	Aperture of small ending of connecting rod
FREQ1			The first frequency of link
FREQ2			The second frequency of link
FREQ3			The third frequency of link
STMAX			Maximum stress of link in the compressed condition
UMAX			Maximum displacement of the lower link maximum load
V			Connecting rod volume

5.2 ANALYSIS OF MDO PROBLEM BASED ON DATA FUSION

Here, connecting rod, which involves many variables is selected to do data fusion analyses. The number of design parameters which affects the designing of connecting rod is fifteen. Based the design of experiment (DOE), it forms 400 effective historical data. Because of great difference between the design parameters, these parameters are normalized by the normalized linear conversion function. At the same time, the weight values of the design parameters are obtained by the method of least squares. It can be seen in Table 2 that the weights of nine parameters are below 0.05. These parameters can be removed from the design parameters because they own less impact on the optimization results. In the last, six parameters (D1, TK1, TK2, TK4, FREQ1, FREQ2) are selected as the design parameters. And then the complexity of the MDO process greatly reduced.

TABLE 2 Weight of the design parameters

Parameter	**D1**	**TK1**	**TK2**	**TK3**	**TK4**	**TK5**	**TK6**	**TK7**
Weight	0.1053	0.0613	0.1186	0.0127	0.0721	0.0094	0.0108	0.0062
Parameter	TK8	TK9	FREQ1	FREQ2	FREQ3	STMAX	UMAX	Co-coefficient
Weight	0.0007	0.0118	0.3407	0.1984	0.0383	0.0051	0.0086	0.9878

Furthermore, the correlation coefficient is calculated at 0.9878, which means the high degree of correlation between the optimization parameters, constraints and optimization goals. Compared to the mathematical models established by fifteen parameters (see Table 1) whose correlation coefficient is calculated at 0.9823, the mathematical models established by six parameters shown a great predictive ability. This fully illustrated that data fusion algorithm can reduce the dimension of MDO problem and greatly improve the efficiency.

5.3 MULTIDISCIPLINARY DESIGN OPTIMIZATION OF THE PROBLEM [12]

Collaborative optimization (CO) [13-15] is a two-level MDO algorithm proposed by Kroo basing the consistency constraints algorithm. The top level is called system optimizer, which optimizing the system variables in order to satisfy the compatibility and minimize the system objective. Every subsystem level optimizes the design variables in the subspace for minimizing the minimum mean square. The collaborative optimization solves the system design variables under the condition of meeting the subsystem constrains. At the same time, the system variables keep unchanged when optimizing the subsystem optimization. This algorithm avoids complex system analysis and makes every subsystem do the analysis and optimize simultaneously, which owns great convergence and reliability.

According to CO, the crankshaft-connecting rod-piston system can be divided into system level and subsystem level. The objective of the system level is to minimal the system mass, and the subsystem level includes four sub-disciplines which completes the analysis respectively under different constraints shown in Figure 7. For instance, connecting rod does the vibration and structural strength analysis, the piston does the vibration and thermal analysis and the whole system do the dynamics analysis.

Considering the objectives and constraints, the system optimization model can be founded:

$$\begin{cases} find\ \ x=\left(TK1...TK7,Hs_h,Hs_{cs},D1\right) \\ \min\ \mathrm{W}=\mathrm{W1}+\mathrm{W2} \\ s.t.\ \sigma \le \sigma_{\max} \\ \quad T \le T_{\max} \\ \quad f \ge f_{\min} \\ \quad x^l \le x \le x^u \end{cases} \quad . \qquad (4)$$

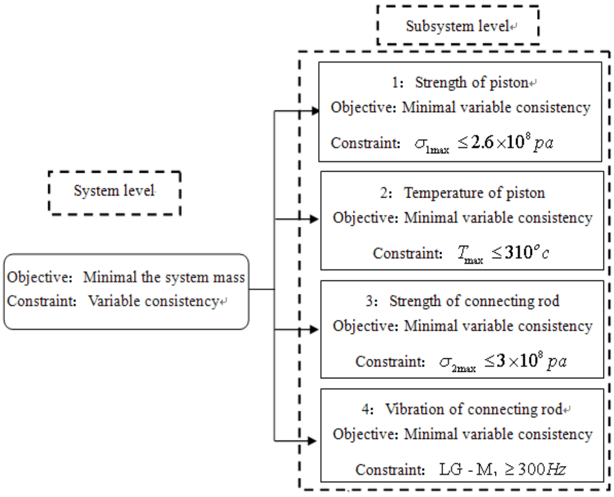

Subsystem level
System level
Objective: Minimal the system mass
Constraint: Variable consistency
1: Strength of piston
Objective: Minimal variable consistency
Constraint: $\sigma_{1\max} \le 2.6\times10^8\,pa$
2: Temperature of piston
Objective: Minimal variable consistency
Constraint: $T_{\max} \le 310^o c$
3: Strength of connecting rod
Objective: Minimal variable consistency
Constraint: $\sigma_{2\max} \le 3\times10^8\,pa$
4: Vibration of connecting rod
Objective: Minimal variable consistency
Constraint: LG - M, ≥ 300Hz

FIGURE 7 Collaborative multidisciplinary design optimization model of the diesel engine

Here x is the design variables, W is system mass, σ is the constraint of strength analysis, T is the constraint of thermal analysis, f is the constraint of vibration analysis, x^l and x^u are the lower and up limit of the design variables listed in Table 1.

According to the collaborative optimization algorithm and the MDO framework proposed above, the MDO collaborative model of crankshaft-connecting rod-piston is established in the soft of ISIGHT (see Figure 8):

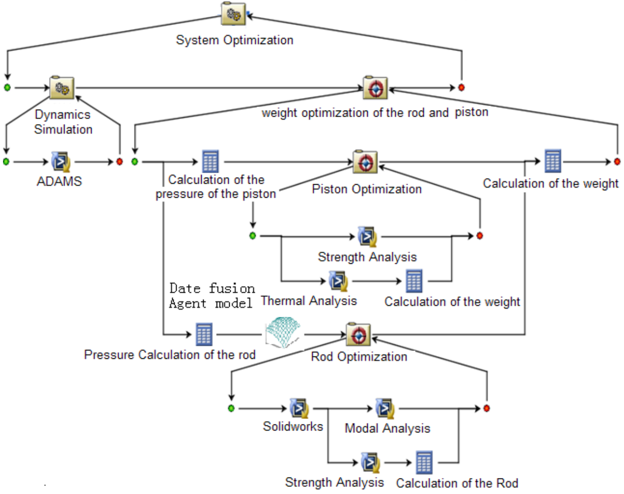


FIGURE 8 MDO collaborative model of crankshaft-connecting rod-piston

5.4 ANALYSIS RESULT

Selecting the nonlinear quadratic programming algorithm as the optimization algorithm, we obtain good results after the crankshaft-connecting rod-piston did 25 iterations, the piston subsystems did 699 iterations and connecting rod did 1961 iterations. Because agent model technology is used, the optimization time is reducing from more than 20 hours to less than 6 hours.

The results show that: mass of the system reduced from 1.02kg to 0.94kg, reduced about 7.51%, which is acceptable. The results of the design variables are listed in Table 3.

TABLE 3 Result of the design variables (Unit: mm)

Variables	TK1	TK2	TK4	Hs_h	Hs_cs	D1
Before optimization	27	21.6	8	8	5	23
After optimization	25.9	18	6	6	6.9	25.5

As shown in Table 4, the objective (system mass) gets the best result after several iterations and satisfies the constraint conditions. It also can be seen that the objective of diesel engine motion mechanism will affect the other disciplines (just like the subject of stress and natural frequent of the connection rod) because of the conflict and coupling of different disciplines. However, MDO along with agent model the second-order response surface model (RSM) and orthogonal array experimental method solved the problem and shown a good result.

TABLE 4 Comparison results of constraint and objective functions

Variables	Max stress of the piston (Pa)	Max temperature of the piston (ºC)	Max stress of the connection rod (Pa)	1st frequent of the connection rod (Hz)	System mass (Kg)
Before optimization	1.55e8	295	1.51e8	393	1.02
After optimization	1.31e8	300	2.1e8	300	0.94

6 Conclusions

In order to deal with the difficulty of the higher order, high-dimensional, multi-input and multi-output in MDO of complex products, an effective MDO framework combined with the method of data fusion and agent model is proposed in this paper. According to the MDO framework, date fusion is applied to deal with the process with a large number of incomplete, vague and uncertain in complex product's evaluation and optimization; agent model is used to reduce the complexity of the MDO model and improve the optimize efficiency. In order to identify the feasibility of the MDO framework, collaborative optimization algorithm and the MDO framework is applied to design of diesel engine motion mechanism and shown a good result. The current study provides a powerful tool for complex products designing and optimization and owns great theory and practical values.

Acknowledgments

This work was supported in part by National Defense Basic Scientific Research program of China (A3820110003).

References

[1] Xu Yong, Zou Huijun 2006 *Conceptual design of mechatronic systems* **31** 661-9

[2] Zhao Qian, Chen Xiao-kai, Lin Yi 2010 *Journal of Jilin University (Engineering and Technology Edition)* **40** 1487-91

[3] Hu Wen-jie, Chen Liang 2010 *Transactions of the Chinese Society for Agricultural Machinery* **41** 17-20

[4] ZHU Ya-tao, CHEN Fang, LI Gao-hua, et al 2011 *Journal of Astronautics* **32** 721-6

[5] Md. Saddam Hossain Mukta, T M Rezwanul Islam, Sadat Maruf Hasnayen. 2012 *International Journal of Emerging Trends and Technology in Computer Science* **1**(3) 255-60

[6] Sobieszczanski-Sobieski J 1997 *Structural 0ptimization* **14**(1) 1-23

[7] Meckesheimer M, Booker A J, Barton R R, Simpson T W 2002 *AIAA Journal* **40**(10) 2053-60

[8] Meckesheimer M 2001 [Ph. D. Dissertation Thesis] *Industrial Engineering* The Pennsylvania State University, University Park

[9] Lee H D, McMullen S A H 2004 Artech House

[10] Forsberg J, Nilsson L 2005 *Struct. Multidise. Optim.* **29**(3) 232-43

[11] Li lei, Zhang Jianrun 2013 *Applied Mathematics & Information Sciences* **7**(5) 1957-62

[12] Li Lei, Zhang Jianrun, Chen Lin 2013 *Transactions of the Chinese Society for Agricultural Machinery* **44**(3) 33-7

[13] Kroo I, Maning V 2000 *Collaborative Optimization: Status and Directions* Stanford: AIAA

[14] SU Ruiyi, GUI Liangjin, WU Zhangbin, et al 2010 *Journal of Mechanical Engineering* **46** 128-33

[15] SONG Baowei, DU Wei, GAO Zhiyong, et al 2009 *Torpedo Technology* **17** 7-11

Authors

Li Lei, born in August 24, 1981, China

Current position, grades: a lecturer in Jiangsu University of Science and Technology
University studies: MS degree in Mechanical Engineering from Nanjing University of Technology in 2007, and studies for PhD degree in Mechanical Engineering in Southeast University
Scientific interest: mechanical design, structure optimization, and information technology.

Zhang Jianrun, born in August 2, 1962, China

Current position, grades: professor in Southeast University
University studies: PhD degree in Mechanical Engineering from Southeast University in 1997, and then engaged in the postdoctoral research in Technical University of Berlin, Germany
Scientific interest: structural dynamic design and optimization, control of vibration and noise

Research on the influence of the different logistics demand structures of the city in regional logistics planning

Si Chen[1], Gan Mi[2*], Dingqi Shuai[1]

[1] *Department of Transportation, Emei Branch, Southwest Jiaotong University, Emei, Sichuan Province, China 614202*

[2] *School of Transportation and Logistics, Southwest Jiao tong University, Chengdu, Sichuan Province, China 610031*

Received 1 March 2014, www.tsi.lv

Abstract

This research analysis the different parts of regional logistics demand at first. There are three parts of the logistics demand considered in this paper; they are logistics demand in the city, logistics demand between cities and the logistics demand from or outside the area. The relationships have been studied by the Grey Theory, and a numerical example has been made to show the way how to analysis the logistics demand structure in the regional logistics planning. In the regional logistics, planning the difference of logistics demand structures of the cities should be fully considered. Then the logistics planning with different regional logistics planning purposes have been programmed. Based on the numerical example, different plans and different influence scopes have been got at last.

Keywords: Logistics Demand, Grey Theory, Logistics planning, Demand Structure

1 Introduction

Regional logistics planning is more about the location of the logistics distribution centres and the optimizing of the transportation network. Moreover, the location of logistics distribution centres are usual in the cities, which have strong attractive to the flow of logistics. The reason of this situation is that the most development economic centres, high-density population centres and the concentrated industry centres are all located in the cities. Since different cities with different structure of economic and industry, the logistics planning to different cities is changed with the situations of cities [1].In a city which is located in a given area, the logistics planning of this city has strong connection with the structure of the logistics demand. In addition, logistics demand is kind of derivative demand of the developing of economic and society. The regional logistics demands perform in one of cities in this area can be classified into three parts showed in Figure. 1. The first part is logistics demand for city itself, which is used to ensure the regular operation of the city. The second part of logistics demand is between cities in the area, which is. The last part is the logistics demand of communication with the outside world [2].When the three parts has been considered as three influence factors to the whole logistics demand in the city; we can analyse the relationship between the logistics demand and the influence factors by Grey Theory [3].

The grey theory, one of the methods that are used to study uncertainty, is superior in theoretical analysis of systems with imprecise information and incomplete samples [4]. Grey related degree analysis is based on the grey theory. Grey theory has become an important ingredient in the development of information processing. Grey systems-based techniques are powerful tools in addressing those systems in which information is partially known and partially unknown [5]. So based on the analyses of the influence of the different structures of the logistics demand we can program the logistics planning by the particular needs of the area. We can build local distribution centres for the logistics demand of one city or between cities in the area. And we can build logistics centre for the whole area to communicate to the outside the area. The process of logistics planning is follow the method of Analytic Hierarchy Process (AHP) [6], which is the approach to relative measurement, a scale of priorities is derived from pair wise comparison measurements only after the elements to be measured are known [7]. The method of Analytic Hierarchy Process (AHP) was developed by Thomas. Saaty [8] in the 1970s. Moreover, it is kind of a structured technique for analysing the complex decisions, which is built up based on mathematics and psychology.

In addition, in this paper we assume that the whole area is made up by cities, so we just can build logistics hubs in the selected cities but the place outside the cities. Regional logistics planning is more about the location of the logistics distribution centres and the optimizing of the transportation network. Moreover, the location of logistics distribution centres are usual in the cities, which have strong attractive to the flow of logistics. The reason of this situation is that the most development economic centres, high-density population centres and the concentrated industry centres are all located in the cities. Since different cities with different structure of economic and industry, the logistics planning to different cities is

* ***Corresponding author*** e-mail: migan@swjtu.cn

changed with the situations of cities. In a city, which is located in a given area, the logistics planning of this city has strong connection with the structure of the logistics demand. And logistics demand is kind of derivative demand of the developing of economic and society. The regional logistics demands perform in one of cities in this area can be classified into three parts showed in Figure. 1. The first part is logistics demand for city itself, which is used to ensure the regular operation of the city. The second part of logistics demand is between cities in the area, which is. The last part is the logistics demand of communication with the outside world. When we consider the three parts as three influence factors to the whole logistics demand in the city, we can analyse the relationship between the logistics demand and the influence factors by Grey Theory.

The grey theory, one of the methods that are used to study uncertainty, is superior in theoretical analysis of systems with imprecise information and incomplete samples. Grey related degree analysis is based on the grey theory. Grey theory has become an important ingredient in the development of information processing. Grey systems-based techniques are powerful tools in addressing those systems in which information is partially known and partially unknown. So based on the analyses of the influence of the different structures of the logistics demand we can program the logistics planning by the particular needs of the area. We can build local distribution centres for the logistics demand of one city or between cities in the area. In addition, we can build logistics centre for the whole area to communicate to the outside the area. The process of logistics planning is follow the method of Analytic Hierarchy Process (AHP), which is the approach to relative measurement, a scale of priorities is derived from pair wise comparison measurements only after the elements to be measured are known. The method of Analytic Hierarchy Process (AHP) was developed by Thomas. Saaty [9] in the 1970s. And it is kind of a structured technique for analysing the complex decisions, which is built up based on mathematics and psychology. Moreover, in this paper we assume that the whole area is made up by cities, so we just can build logistics hubs in the selected cities but the place outside the cities.

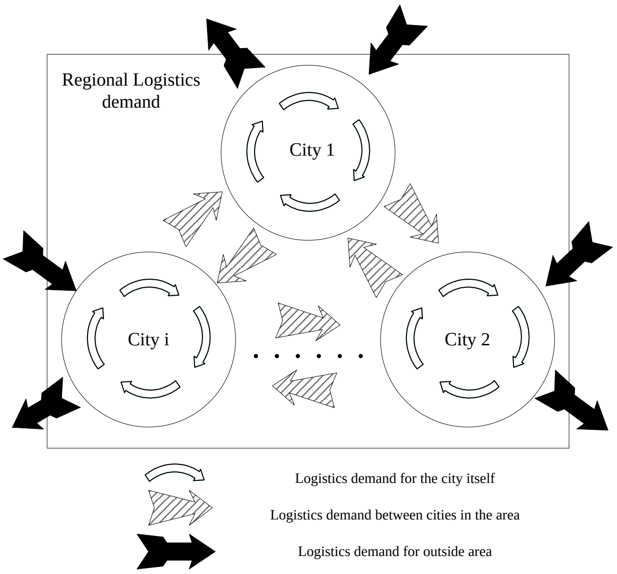

Regional Logistics demand
City 1
City i
City 2
· · · · · ·
Logistics demand for the city itself
Logistics demand between cities in the area
Logistics demand for outside area

FIGURE 1 The components of Regional Logistics Demand

2 Formulation of grey related degree

In this paper, we try to calculate the grey related degree of the logistics demand and the influence factors. The characteristic sequence of system is expressed as:

$$X_0 = \left(x_0(1), x_0(2), \cdots, x_0(n)\right). \tag{1}$$

The characteristic sequence of related factors is shown as:

$$X_i = \left(x_i(1), x_i(2), \cdots, x_i(n)\right), \tag{2}$$

$$\begin{aligned} X_i D &= \left(x_i(1)d, x_i(2)d, \cdots, x_i(n)d\right) \\ &= \left(\left(x_i(1) - x_i(1)\right), \left(x_i(2) - x_i(1)\right), \cdots\left(x_i(n) - x_i(1)\right)\right) \end{aligned}$$ is remarked as:

$$X_i^0 = \left(x_i^0(1), x_i^0(2), \cdots, x_i^0(n)\right), \tag{3}$$

φ_{0i} is the related degree of the i^{th} influence factor. And $0 < \varphi_{0i} \le 1$, the bigger φ_{0i} the indication of geometric similarity closer between X_0 and X_i, or otherwise.

$$\varphi_{0i} = \frac{1 + |S_0| + |S_i|}{1 + |S_0| + |S_i| + |S_i - S_0|}. \tag{4}$$

The $|S_0|$, $|S_i|$, $|S_i - S_0|$ in the Eq. 4 are showed in the following: $|S_0| = \left|\sum_{k=2}^{n-1} x_0^0(k) + \frac{1}{2} x_0^0(n)\right|$;

$$\left|S_i\right| = \left|\sum_{k=2}^{n-1} x_i^0(k) + \frac{1}{2} x_i^0(n)\right| ;$$

$$\left|S_i - S_0\right| = \left|\sum_{k=2}^{n-1}\left(x_i^0(k) - x_0^0(k)\right) + \frac{1}{2}\left(x_i^0(n) - x_0^0(n)\right)\right| .$$

φ_{0i} will be seem as the relationship between the whole logistics demand and the three parts of logistics demand. When the one of φ_{0i} is bigger than the others, the logistics demand of the city is tend to have much closer relationship with that part of logistics demand. In addition, the influence scale of logistics demand of the city will change with it too. Following this, the logistics planning will adapt the police with different ways.

In this paper we can classified the logistics demand Structures of the cities in regional logistics planning into three groups. The first structure is that the outside logistics demand is stronger than other demands. We can name it as Structure 1 (S1). The second structure is that the logistics demand between cities is much stronger than other types of logistics demands. Moreover, we name it as Structure 2 (S2). The last structure is that the logistics demand in the city is stronger than others. It is named as Structure 3 (S3). Then we can chose suitable logistics planning based on the different purposes by the different structures of cities.

3 Numerical example

In this section, we suppose there are three main cities in this area, which we studied. And we choose the data of quantity of logistics demand in ten years. We assume that the three parts of logistics demand are already known. The data are shown in following Table 1, Table 2 and Table 3.

TABLE 1 The quantity of logistics demand and influence factors of city 1

Years	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
The whole logistics demand	20	23	24	28	31	33	35	40	44	50
Logistics demand in city 1	10	13	14	16	18	20	22	26	30	33
Logistics demand between cities	6	3	4	3	4	10	3	6	9	10
Logistics demand outside area	4	7	6	9	9	3	10	8	5	7

TABLE 2 The quantity of logistics demand and influence factors of city 2

Years	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
The whole logistics demand	100	110	120	140	150	155	168	170	175	180
Logistics demand in city 2	60	61	50	40	70	40	68	58	70	70
Logistics demand between cities	20	25	35	50	55	66	80	72	85	90
Logistics demand outside area	20	24	35	50	25	49	20	10	20	20

TABLE 3 The quantity of logistics demand and influence factors of city 3

Years	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
The whole logistics demand	60	70	80	95	102	105	106	120	130	135
Logistics demand in city 3	25	20	21	19	15	30	23	28	15	27
Logistics demand between cities	15	20	15	16	28	5	10	12	30	8
Logistics demand outside area	20	30	44	60	60	70	73	80	95	100

According to the analysis in section 1 and 2, we can calculate the related degrees by Eq. 4, and we also can make use of the software Matlab to achieve the results. And the results are shown in Table 4.

TABLE 4 The Related Degrees of three cities

Related degrees **Cities**	**Related degree of Logistics demand in city i (R1)**	**Related degree of Logistics demand between cities (R2)**	**Related degree of Logistics demand outside area (R3)**
City 1	0.92	0.73	0.62
City 2	0.80	0.74	0.65
City 3	0.76	0.67	0.78

From the data of Table 4, we can find that the logistics demand of City 1 has much stronger connection with the logistics demand in the city than other parts of logistics demand. Which means when we consider the logistics planning of City 1, we'd better treat City 1 as a self supply city. The logistics distribution centres should be built to meet the needs of the logistics demand in the city. And comparing the data of related degrees of logistics demand between the cities, the data of City 2 is larger than City 1 and City 3 (0.74>0.73>0.67), the logistics canters for cities in the area may be sit in or near City 2. That will be much more reasonable and good for the whole regional logistics planning.

The related degree of logistics demand in city (R1) is the maximum degree in the data of City 2, so in the three parts of logistics demand the R1 is the most important to City 2. However R1, R2 and R3 are very similar (0.8, 0.74 and 0.65), City 2 should be consider to be balance point in the area, when we try to do a regional logistics planning.

Since City 3 has the biggest related degree of logistics demand outside area between the three cities, it may be seem as the most important hub to communicate with outside area. The influence scale of City 3 may be wider than City 1 and City 2. In the regional logistics, planning the difference of logistics demand structures of the cities should be fully considered. In one word, in this paper the structure of City 1 is S3, the structure of City 2 is S2, the structure of City 3is S1.

4 Regional logistics planning

In the process of Analytic Hierarchy Process, the project should be decomposed in to different parts and elements at first. Then the elements will be classified into different levels and different groups. The elements in the same level will be treated as the standard to the elements of next level. And in the same time the elements will be controlled by the upper level [10]. The whole hierarchy can be divided into three classes: the top level, the middle level and the lowest level. The relationship of dominance between different levels is not completed. There is an element, which not controls all the elements in the next level but part of them. The structure of the relationship of dominance is called as hierarchical structure [11].

In this paper, we consider the purpose of the Logistics planning as the top level. In this level there are three purposes, one is the logistics planning is to meet the logistics needs of outside the area, the other is to meet the logistics demand between cities in the area. And the last purpose is build logistics centres to satisfy the most needed city [12]. The middle level of this project is the demand structures of the cities.

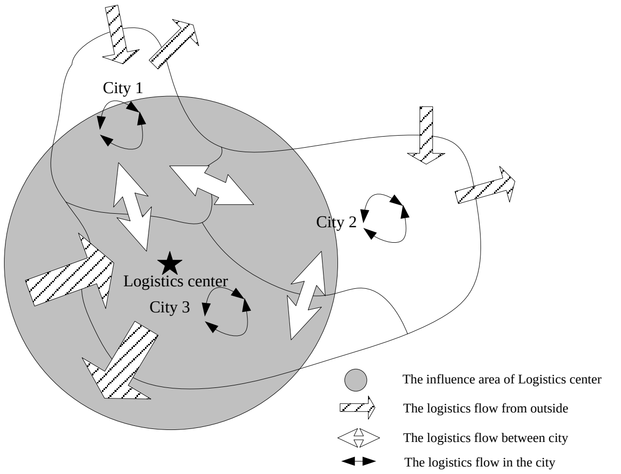

City 1
City 2
Logistics center
City 3
The influence area of Logistics center
The logistics flow from outside
The logistics flow between city
The logistics flow in the city

FIGURE 2 The Regional Logistics Planning for Logistics demand of outside the area

Since we consider the whole area is made up by cities, we just can build logistics hubs in the selected cities. So based on the conclusion from the section 3, we can find that if we try to optimize the regional logistics to meet the needs of logistics demand outside of the area, we can find that City 3 is the best choice of the three cities. As shown in the Figure. 2.

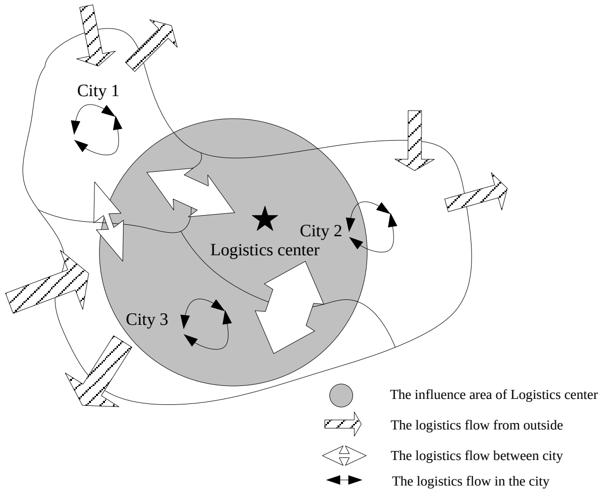

City 1
City 2
Logistics center
City 3
The influence area of Logistics center
The logistics flow from outside
The logistics flow between city
The logistics flow in the city

FIGURE 3 The Regional Logistics Planning for Logistics demand between cities

Because of the new Logistics Centre, which is built for the logistics demand from outside the area, the whole logistics demand flow will be attracted to the location of new logistics centre. All the logistics flows of City 3 are increased with the new logistics centre. On other hand, the flows of City 1 and City 2 decreased.

When we chose the purpose, which is to meet the logistics demand between cities in the area. We can find that City 2 has the strongest connection to other sites. When we set a new Logistics centre in City 2 to serve the logistics demand between cities, it will be much more effective than in other cities. Therefore, we can program the logistics planning with the second purpose as shown in Figure. 3.

When we compare Figure 2 with Figure 3, we can find that, the influence area of the new logistics centre with the second purpose is smaller than the influence area with the first purpose. With different purpose, the logistics planning will be changed and the standard of the logistics canter will be changed too.

If we choose the last purpose, building new logistics canters to satisfy the most needed city, we can find that City 1 is the most appropriate choice in the three cities in this paper. In this part, we suppose that we just can build one logistics canter, so we have to find out the strongest logistics demand connection for the city self in the whole logistics demand. And we can program the logistics planning with the purpose and the suggestion, mentioned above, as Figure. 4.

The logistics flow for the City 1 itself increase with the building of new logistics canter. And we also can figure out that the standard with this purpose is lower than that mentioned above. The scope of influence area is much smaller and it is limited in the City 1.

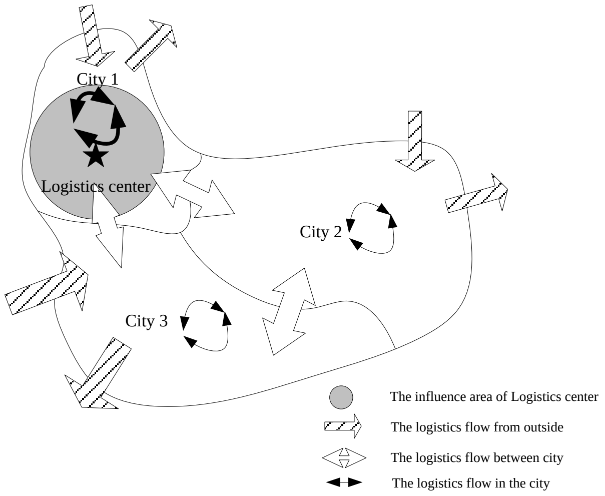

City 1
Logistics center
City 2
City 3
● The influence area of Logistics center
▨ The logistics flow from outside
The logistics flow between city
The logistics flow in the city

FIGURE 4 The Regional Logistics Planning to satisfy the most needed city

5 Conclusions

This study analysis the different parts of regional logistics demand at first. There are three parts we are considered in this paper, they are logistics demand in the city, logistics demand between cities and the logistics demand from or to outside the area. We study their relationships by the grey theory, and make a numerical example to show the way how to analysis the logistics demand structure in the regional logistics planning. Then we program the logistics planning with different regional logistics planning purposes. Based on the numerical example, we got different plans and different influence scopes.

Acknowledgments

This research is supported by “the Fundamental Research Funds for the Central Universities” (10501B10096017) in China.

References

[1] Chen Si, Yan Ying, Song Hongjiang 2010 Optimal Logistics Hubs Location on the multimodal transportation network *IEEE International Conference ICLEM*

[2] Chen Si, Tang Yinying 2011 Research on the relationship between the economic development and industry city *IEEE International Conference ICCASM*

[3] Hu Zhaoyin. Gray 2003 Theories and Its Application *Journal of Wuhan University of Technology* **27** 405-11

[4] Chen Mianyun, Li Zhijun 2001 A control approach to “poor” information systems *Advances in Systems Science and Application* **2** 152-8

[5] Wang Jianqiang 1997 A New Computation Method of Grey Interconnect Degree and Its Application *System engineering theory and Practice* 119-23

[6] Tan Guanjun 2000 The Structure Method and Application of Background Value in Grey System GM (1, 1) Model (I) *System engineering theory and Practice* 98-103

[7] Sifeng Liu, Jeffrey Yi-Lin Forrest 2010 *Grey Systems: Theory and Applications* Springer-Verlag Berlin Heidelberg

[8] Saaty T L 2005 *Analytic Hierarchy Process* Encyclopedia of Biostatistics

[9] Saaty T L, Peniwati K 2008 *Group Decision Making: Drawing out and Reconciling Differences* Pittsburgh, Pennsylvania: RWS Publications

[10] Smola A J, Scholkopf B 2004 A tutorial on support vector regression *Statistics and Computing* **14** 199-222

[11] Si Chen, Mi Gan, Yinying Tang 2013 Analysis of predicting the diversity regional logistics demand based on SVR: the case of Sichuan in China *Applied Mathematics & Information Sciences Sci.* **7**(2) 645-51

[12] Ackermann J, Muller E 2007 Modelling, planning and designing of logistics structures of regional competence-cell-based networks with structure types *Robotics and computer-integrated manufacturing* **23** 601-7

Authors

	Chen Si, born in December, 1982, Chengdu **Current position, grades:** assistant professor in Southwest Jiaotong University **Scientific interest:** logistics management, supply chain management, transportation network optimization and logistics system optimization. **Publications:** 15 **Experience:** Si Chen got PHD degree in Transportation planning and management from Southwest Jiaotong University in 2013. And she studied in University of Arizona from 2010 to 2011. She now is an assistant professor in Southwest Jiaotong University. Her research interests are
	GanMi, born in July, 1984, Yue Yang **Current position, grades:** assistant professor in Southwest Jiaotong University **Scientific interest:** supply chain management, logistics network, mathematical model, transportation planning. **Experience:** Mi Gan received the PHD degree in Logistics engineering from Southwest Jiaotong University in 2012. And she is an assistant professor in Southwest Jiaotong University. Her research interests are
	Shuai Dingqi **Current position, grades:** ideological and political counselor in Southwest Jiaotong University **Scientific interest:** transportation planning, mathematical model. **Experience:** Dingqi Shuai received the undergraduate degree in Southwest Jiaotong University in 2013.

Integrating TTF and TAM perspectives to explain mobile knowledge work adoption

Yongqin Jin[1], Dongsheng Liu[2, 3]*

[1] *Zhejiang Digital Financial Management Center, HangZhou, China*

[2] *College of Computer Science & Information Engineering Zhejiang Gongshang University, Hangzhou, China*

[3] *Centre for Studies of Modern Business Zhejiang Gongshang University, Hangzhou, China*

Received 1 March 2014, www.tsi.lv

Abstract

It is an advanced research subject to information technology as well as a great influence to the development of mobile work that how mobile work service is adopted and how it provides effectiveness. This article refers to 1) Analyse principles of TAM & TTF models, clarify the precondition and strength & weakness of the model, and propose a new mobile work service model combined with TAM & TTF; 2) Practical study to the new model. The conclusion for this model is as follow: a) two basic characteristics of mobile work and support from up-level managers in a firm are the preconditions whether service will be adopted. b) task-technology matching is the significant factor on service acceptation. c) It could improve employees' efficiency within the practical use of task-technology matching mobile work service.

Keywords: mobile work, TTF, TAM, Embedded model

1 Introduction

Mobile knowledge workers or the mobile official workers, generally refer to people who spend more than 20% of the time away from office [1], such as sale staff, negotiators, after-service staff and the like. With the sharp development of Internet, Mobile Device (Cell phone, PDA, etc.) and the 3G communication network, the scale of mobile knowledge workers has been expanded continuously. Yankee Group's research results show that mobile knowledge workers amounted to 55 million in USA in 2005, close to 40% of the total workers, and it even rose to 65% the next year [2, 3]. Based on the mobile office service condition supported by each corporation to their staff and the pressure of the balance between work and life, experts from IDC predicate that in 2015, USA will become the world second mobile applications popular country with narrowly 5-percentage-point weakness to Japan, the leader of mobile knowledge workers in the world with 80% mobile knowledge works. Currently mobile knowledge works in Japan occupy 53 percentage points. Apparently, most workers are eager to apply mobile technology in working so that they can do things much more efficiently out of office [3].China owns the largest mobile-phone users and Internet users [4, 5], due to 3G's spring up, Internet and mobile service tend to blend itself with the other. China's mobile knowledge work-groups will certainly expand rapidly with mobile commerce for its largest subscribers. However, according to past experience, it's acceptability problems that always becomes barrier to new system and new technology's adoption, therefore, we should pay more attention to its acceptability. Acceptability makes great sense to both service providers and users. To providers, they can find out factors to user's acceptability, users' needs and expectation further so that they could offer services that users are willing to use; to users, they can realize their needs with awareness of providers' motivation so that they could choose the best service for themselves.

Compared to mobile commerce in China and abroad, Dehua He and Yaobin Lu concluded that whereas special conditions of China, acceptance and involved factors that effect users' adoption must have something different to others. To this point, after probing into the structure, complexity, frequency, urgency and mobility and so forth of mobile commerce, Chu and Huang set up Task-technology fit mobile commercial applications adopt model to instruct industry's development, it is a pity that they didn't explain relevant technology in detail, so the model always acts as a referenced framework for the latter period [6]. Chaohua Deng, et al came up with a model integrated with IDT/TTF, and studies enterprise mobile message service as an example, achieving good results to explain. Nevertheless, current mobile service has been far beyond message service, so the research is not universal. In addition, the personalization, location awareness, task complexity, time constraint, ubiquitous and so on of mobile knowledge worker make users' actions quite different from ordinal users' in mobile business cooperation [7, 8], therefore, acceptance of mobile knowledge worker problems cannot be simply

****Corresponding author*** e-mail: lds1118@zjgsu.edu.cn

equated with mobile commerce research. So far, little research has been done but for some papers presented by Zheng and Yuan. Zheng and Yuan proposed a concept framework of mobile knowledge worker, through the definition of the technical characteristics, they explained the nature of mobile, offering good basis for the research and development of mobile devices, but they did not go on further study on factors that lead users adopt mobile knowledge worker, lacking of convincingness in practice [9].

This paper hopes to explore the main theory framework of users' acceptance and method to study based on summary of domestic and internal related problems about the acceptability and adoption of mobile knowledge worker, to pave the way for deep researches. There're two main points in this paper:(1) based on actual System use and Task-technology fit, construct an integrated relation model which is able to explain and predicate mobile knowledge worker's mobile service adoption; (2) empirical research on the model, confirm validity of the model.

2 Literature review

On the base of TRA, Davis extends to a relationship among attitude, behaviour and intention, proposing the Technology Acceptance Model (TAM) to successfully explain the decisive factors about the personal computer which is widely accepted [10]. Because TAM has a solid theoretical foundation and can be effectively used to understand and explain the behaviour of IS, once made, having aroused extensive attention in theoretical circles. In the past 20 years, TAM model is widely used in more than 50 kinds of information technology adoption and use of research.

An Empirical Study of the above not only prove the versatility of the TAM model in the field of information technology, but also validate the model perceived usefulness and perceived ease of use of the powerful explanatory power. However, we also have noticed that as the TAM model empirical results suggest that the behavioural intention is proportional to actual usage and the unpredictable nature of actual usage behaviour, the follow-up studies based on TAM simply take the behavioural intention as an end-user adoption of standards, only a few studies discuss the correlation of the behavioural intention and the actual usage. Standish Group in 1998 on the implementation of IS's findings on the secondary permitted the final thesis of these scholars that the behavioural intention could not be representative of the actual usage. Only 26% of the MIS projects and less than 23.6% of the projects of large companies could be completed according to requirements, on time and on budget. Close to 1/3 (28%) of the projects were cancelled in the end [11]. The system acceptance and use of the end users is mandatory, for example, in the mobile office system applications, the end-users' usage behaviour is not only mandatory, but the tasks of each user are also tightly integrated with the coupling [12, 13]. For these reasons, Rawstorne, etc. and Karahanna, etc. proposed to replace the accepted model of behavioural intentions by Symbolic Adoption [14, 15]. Symbolic acceptance refers to the individual's acceptance of new technologies in the ideological, while the actual acceptance refers to the actual use of technology. Based on this understanding, Fiona FuiHoon Nah, etc. figured that, for the application of complex systems, the concept of symbolic acceptance is more suited in acceptance model [16]. In this study, we also use the concept of symbolic acceptance. However, in the interpretation of its practical application, due to the TAM model's lack of concern about the organization's mission, it results in confusion between the perceived usefulness of information technology and the perceived usefulness of this information technology for specific tasks, there exist obvious structural defects in its theory.

To make up for deficiencies in information technology cognitive for of TAM, Goodhue and Thompson focused on information systems and the relationship among individual behaviours of users, and proposed the Task/Technology Fit Model [17]. In their theory, the function of information technology will be adopted only when it supports work well, on the contrary, the technology will not be accepted [17, 18]. Since the TTF has been raised, it has been quoted as many as more than 1000 times during the short period of 10 years of development history. The representative studies include: a model of Task/Technology Fit was used to assessing organizational decision-making system by Zigurs and Buckland, which further fully proved the availability of TTF at the organizational level [19]; Dishaw and Strong used TTF to explore the impact on the willingness of organizations using software maintenance tools, it not only conducted a study on the organizational level successfully, but also successfully described the TTF can be used to study the adoption of information systems [20]. There are still many research articles about TTF contributed to the development of TTF and enhanced the operationalization of TTF. Such as Benslimane, Plaisent and Bernard used TTF in the adoption studies of E-Commerce technology and so on [21].

In spite of the comprehensive application of the Task/Technology Fit Model within information technology research, Goodhue and Thomp son mentioned the limitation exists when we focus on matching model, which lack of the notice that information technology application should be prior to the effect generated by it. In other word, the premise of the foundation of TTF is adoption of this technique. On the other side, although two factors, task demands of enterprise and information technology, have been involved in this model, the absentation of a communication between technology fit and information technology adoption which leads the failure of internal mechanism reflection of individual behaviour exposed to technology fit. This is the bottleneck of the further development of TTF [22].

In the research, Dishaw and Strong found that, the model of Task/Technology Fit is counterbalanced by TAM model on the care to IT application, the key point of TAM. Therefore, in 1999 they supposed to combine TAM and TTF together to solve the challenge of technique adoption. The result demonstrated that TTF worked better than TAM when using in explain "job-related" task, in their research. It was found, that job and technique affect TTF, TTF make an impact on "usability of recognition" and "practical application". However, no significant influence was shown in the preconceived effect of "usefulness of recognition". Another conclusion was drawn from this research, which is the combination model of TTF and TAM can better explained than TTF or TAM respectively did.

Throughout the history of the development of TAM and TTF, scholars have applied TTF into the adoption studies of more than 50 emerging technology, and have got fruitful achievement. However, these studies either remain in the theory development phase, or lack of representation in the empirical research process, and the study for the mobile office is inadequate. Therefore, both in theory and in practice, it is very meaningful to apply TAM and TTF theory into more representative and universal adoption research of mobile office.

3 Hypothesis and model

3.1 CHARACTERISTICS OF MOBILE KNOWLEDGE WORKER

(1) *Task-carried context*

The user's need is a sequence over time and changes of events. When an event occurred, the needs will also change, and the change of events depends on a certain context. When some changes in certain context cause the failure in occurrence of the event, the subsequent series of events may be affected. Changes in the new context may also lead to a new task events, such as the event of heavy rain, traffic disruption, leading to the suppression of existing business itinerary, and create a new itinerary task. Thus, completion of the tasks requires some contexts; however, the contexts are also the conditions or triggers for completing a certain task. Extant research has found that situation relevance influence perceived usefulness significantly [23]. Based on the above analysis, the following hypotheses are proposed:

H1a: Task-carried context has a direct effect on perceived usefulness.

H1b: Task-carried context has a direct effect on perceived ease of use.

(2) *Resource Availability*

The effective personalized service to meet the user's needs which occurs in moving is the cornerstone of mobile knowledge worker. When the user is moving and constant interaction occurs with a fixed remote service, network performance will not very stable sometimes. The worst case is the network link interruption, so that the running services will be also interrupted; In addition, users may sometimes need the appliance of services previously in their own devices, but the current system cannot configure this service. In the face of above problems, the design of mobile knowledge worker that whether the system allows the user to select the provision of remote services, allows the user's local machine to run off-line services, or supporting the users device to run a new service in a more powerful neighbour, or supporting the re-orientation of running service has a direct impact on users' perceived ease of use and perceived usefulness. Therefore, the following hypotheses are proposed:

H2a: The availability of resources has a direct effect on perceived usefulness.

H2b: The availability of resources has a direct effect on perceived ease of use.

3.2 PERCEIVED EASE OF USE

Perceived ease of use can be regarded as the extent of user understanding of technology. In general, the technology which is easier to understand and use has greater chance of being accepted. The technology that the more easy to be use is the more to be applied under the same conditions.

Szajna puts experience into the model and proves that Perceived ease of use influents behavioural intention no matter before or after people taking up the technology [24]. David Gefen made a gender-specific study about the influence of Perceived ease of use. The empirical results show that the impact to Behavioural Intention to Use is still there [25]. Furthermore, Venkatesh in 2003, and Davis, Morris co-operation of the article, is also demonstrated in detail the existence of such relationship [26]. Therefore, the study in the mobile office, the same applies to the following theorem: Users who have higher perceptual level of perceived ease of use about mobile office technology, namely, regard it requires less effort and resources, is easier to accept the technology. On the contrast, users who regard the technology require more effort will reduce the positive effect to Behavioural Intention to Use. The above theory put forward this hypothesis:

H3a: Perceived ease of use of mobile office service has positively correlated with the symbolic acceptance.

Perceived ease of use can also indirectly, by influencing the role of perceived usefulness on the symbolic acceptance. Szajna who made a more detailed exploration about the difference of perceived ease of use in two stages -before and after, found that, before use, because users have no enough experience to the new technology, perceived ease of use fails to impact perceived usefulness, after use, the experience of ease of use increased, in addition to direct effects on the symbolic acceptance. Perceived ease of use will indirectly, by influencing the role of perceived usefulness on the symbolic acceptance. The impact of perceived ease of use on perceived usefulness, although unlike their direct effect on the symbolic acceptance, but it reflects the association between the two main factors, the situation also reflects the people's cognitive habits. It is also adopted by many scholars. The technology with stronger perceived ease of use will be regarded as more useful, however, the technology which is not ease to use will be limited to improve the performance no matter how many strong functions it carried and how much useful it will be to work. Based on the above analysis, put forward this hypothesis:

H3b: The perceived ease of use of mobile office service has positively correlated with the perceived usefulness.

3.3 PERCEIVED USEFULNESS

Perceived usefulness is on behalf of the information systems users' awareness of the use of that technology can bring to improve the job performance. As the main factor in the prototype of Technology Acceptance Model, perceived usefulness is considered strong predictor that impacts the symbolic acceptance. Davis proposed in the original model that perceived usefulness and perceived ease of use, two factors would have a direct impact on symbolic acceptance, also stressed the perceived usefulness is relatively stronger in terms of influencing factors, and it has a closer relationship with symbolic acceptance. Although, Taylor etc. focused on greater level comparison the difference between Technology Acceptance Model and TPB model, the result further verified the perceived usefulness' significant role in symbolic acceptance [27]. Szajna (1996), Patrick Chau (2001), Morris (2003) is all validated the relationship in the study. In a word, in the research field of Technology Acceptance Model, the perceived usefulness is widely recognized as the most basic factors. The users believe that the more useful a technology, the more inclined to use the technology. On the contrary, if users think the technology cannot significantly improve performance, usefulness is low, and then its use tendency will be significantly lower. On the above theories, this hypothesis is:

H4: Perceived usefulness of mobile knowledge worker operations and the symbolic acceptance are positive correlated.

3.4 SUBJECTIVE NORM

Subjective norm refers to the individual whether to take a specific act perceived social pressure. Azjen think, on the mandatory conditions, the behaviour affected by the pressure impact of the social environment will be greater than the individual's own attitude. Theory of Reasoned Action (TRA) pointed out that Subjective Norm is one of the determinants of Behavioural Intention. In 2000, by the study of Venkatesh and Davis, information system usage scenarios proposed to be divided into two types, voluntary use and mandatory use. Research shows that when the user is in a voluntary state, the Subjective Norm have little effect on the user's acceptance, but when the user is in a mandatory state, the Subjective Norm have a significant effect on user's acceptance. In the application of mobile knowledge worker services, end user's usage behaviour is not only mandatory, but also the task of each user is also tightly integrated with the coupling. No matter how the end-user attitudes and acceptance of the system will be, they generally do not have the right to choose the system. Therefore, in this case, the subjective norm has a significant impact on the symbolic acceptance. The above theories, this hypothesis:

H5a: Subjective norm and symbolic acceptance are positive correlated.

In the relevant practice, more is the company's high-level decision-making, the involvement of middle-level leadership, which separate the user from the involvement in decision-making stimulate the staff's usefulness of perception due to the higher mandatory. In addition, as a result of high-level participation, in contrast, funding for the system implementation is in a certain degree of protection, thus it will facilitate the realization of a variety of technical training to enhance the user perceived ease of use. Accordingly, proposed this hypothesis:

H5b: Subjective norms and perceived usefulness are positive correlated.

H5c: Subjective norms and perceived ease of use are positive correlated.

3.5 SYMBOLIC ACCEPTANCE

Symbolic acceptance refers to the individual's acceptance of new technologies in the thought, but the actual acceptance refers to the actual use of technology. In the aforementioned article, we have already discussed in the mandatory conditions, symbolically accept the feasibility of alternatives to the use of intent, but also discussed the TAM does not apply to the reasons for the organizational level as well as TTF as a basic theory research is required to use this technology has been a prerequisite. Therefore, users for the use of the technology is uncertain and with the organizations to adopt the decision-making and change, this time it is only "forced to adopt", which TTF of the state already meet the prerequisite. Based on the above analysis, we put forward this hypothesis:

H6: Symbolic acceptance is a prerequisite for task-technology fit.

3.6 TASK-TECHNOLOGY FIT

People are always using an information technology in order to meet certain mission requirements. Weil and Olson found that research assignments matching the better performance is better. Goodhue thought only when the use of information technology to support the task, the employees could not show a good performance [28]. Accordingly, we assume that: H7a: Task-technology fit forward for users of the perceived usefulness of mobile knowledge worker.

Because different development goals and tasks of the organization there is diversity, and therefore the demand for services on the mobile knowledge worker is not the same. The mobile knowledge worker services, only for a particular task to design in order to meet their needs, did not meet the needs of the organization's mission will reduce the organization to the service or understanding of the usefulness of technology will reduce the ease of use of the technology organization Awareness, when the information technology is looked as the research object, only when the information technology function can well support the organization's mission requirements, it will be adopted, otherwise, not be adopted [28]. Accordingly, we assume that:

H7b: Task-technology fit with the use of a positive correlation.

4 Scale and data collection

4.1 SCALE

The proposed hypothetical model in mobile knowledge work was tested empirically using a survey study. Questionnaires consist of two parts. The first part is demographical information include gender, age, education, etc. The second part is the variables in the model and eight key constructs was measured with multiple items. These items were either adopted or adapted from the extant literature. The scales of perceived usefulness, perceived ease of use and symbolic intention were adopted from TAM. The characteristics of mobile knowledge worker including task-carried context and availability were adopted from some mobile commerce and situation perception literatures [29-31]. The scales of task-technology fit were adapted from the research of Goodhue D.L. Each item was measured with the five-point Likert scale, ranging from ‘strongly disagree’ to ‘strongly agree’.

4.2 SAMPLING

Empirical data for this study was collected via a survey of a retail group in Hangzhou, a southeastern city of China. We adopted stratified sampling in a certain proportion in order to get representative sample. Firstly, departments are divided into three categories on the principle of demand for mobile office service: high, medium and low. Then, respondents needed decrease proportionally. The objects of sampling are 15 branches from five cities. In order to improve the precision of data, the author increases the size of the sample and participation of managers promotes the number of finished questionnaires. In total, 300 questionnaires were returned, of which 63 questionnaires were discarded as they were only partly completed. Therefore, 237 questionnaires were retained for analysis, giving an effective response rate of 79 percent. Demographically 57% of the respondents were male and 43% were female.

5 Data analysis

5.1 SACLE RELIABILITY AND VALIDITY

As showed in Table 1, the Cronbach’s alpha for all constructs exceeded 0.6 and most is 0.8 around, indicating that the item reliability were judged to be adequate.

TABLE 1 Reliability test

Variables	Items	Mean	Standard deviation	Scale	Alpha
Task-driven situation	3	2.9660	0.367	1-5	0.6952
Availability of resources	3	3.5625	0.0132	1-5	0.8067
Perceived ease of use	4	3.3439	0.0184	1-5	0.8601
Perceived usefulness	4	3.3449	0.0279	1-5	0.8159
Subjective norm	3	3.6326	0.0213	1-5	0.8437
Symbolic adoption	3	3.5971	0.0179	1-5	0.8742
Task-technology fit	3	3.1449	0.0722	1-5	0.7839
Task-driven situation	3	3.5935	0.0075	1-5	0.7915

Table 2 shows KMO and Bartlett’s Test. KMO is the coefficient of sampling adequacy. The bigger KMO is, the more common factors in variables are and the more appropriate factor analysis is. According to the viewpoints of Kaiser, it is inappropriate to conduct factor analysis if KMO < 0.5. KMO here is 0.815, shows significant. Data is suitable to factor analysis.

TABLE 2 KMO and Bartlett's Test

Kaiser. Meyer. Olkin Measure of Sampling Adequacy		.815
Bartlett’s Test of Sphericity	Approx. Chi. Square	1380.347
	df	476
	sig.	.000

Validity includes convergent and discriminant validity. Convergent validity measures whether items can effectively reflect their corresponding factor, whereas discriminant validity measures whether two factors are statistically different. In these models, CR and Cronbach’s alpha are bigger than 0.8, which shows good reliability [32]. The CRs and AVEs of all the constructs exceed the recommended threshold of 0.8 and 0.5 respectively, thereby indicating good internal consistency.

TABLE 3 Result of exploratory fact analysis

	Fact1	Fact2	Fact3	Fact4	Fact5	Fact6	Fact7	Fact8
C1	0.878	0.070	0.102	0.078	0.055	0.102	0.139	0.035
C2	0.878	-0.027	-0.011	0.175	0.149	0.073	0.130	0.081
C3	0.716	0.151	0.048	0.033	0.054	0.067	-0.008	0.212
R1	0.068	0.860	0.062	0.091	0.073	-0.046	-0.014	-0.034
R2	0.016	0.849	-0.086	0.067	0.048	0.074	0.137	0.011
R3	0.035	0.789	0.045	0.160	-0.032	-0.113	0.188	0.058
PU1	0.057	0.074	0.832	-0.029	0.098	0.112	0.021	0.191
PU2	0.118	-0.027	0.792	0.057	0.046	0.112	0.125	0.090
PU3	0.070	0.135	0.707	0.068	0.361	0.131	0.192	0.064
PU4	0.153	0.103	0.509	-0.051	0.203	-0.096	0.243	-0.007
PE1	0.004	0.024	0.127	0.818	0.154	-0.015	0.059	0.029
PE2	0.058	0.160	-0.080	0.787	0.052	0.124	0.097	0.033
PE3	0.129	0.062	0.026	0.766	0.100	-0.076	-0.022	0.011
PE4	0.122	0.196	-0.032	0.756	-0.074	-0.032	-0.041	0.039
SA1	0.029	0.079	0.132	0.054	0.856	0.049	0.185	0.127
SA2	0.110	0.060	0.201	0.045	0.823	0.207	-0.008	0.067
SA3	0.140	0.025	0.165	0.152	0.815	0.169	-0.060	0.061
TTF1	0.204	-0.018	0.094	0.092	0.197	0.729	0.153	0.045
TTF2	0.210	-0.063	0.017	-0.028	-0.113	0.720	-0.019	0.106
TTF3	0.022	-0.094	-0.021	0.014	0.221	0.716	0.195	0.032
SN1	0.158	0.067	0.132	0.089	0.130	0.126	0.800	0.178
SN2	0.242	0.124	0.330	0.044	0.011	0.160	0.660	0.134
SN3	0.353	0.049	0.156	-0.056	-0.015	0.113	0.631	0.309
A1	0.151	-0.005	0.199	0.066	0.042	0.080	0.315	0.763
A2	0.283	0.023	0.289	0.031	0.041	0.230	-0.045	0.690
A3	0.181	0.027	-0.045	0.031	0.151	0.002	0.203	0.665

To examine the discriminant validity, we compared the square root of AVE and factor correlation coefficients. As listed in Table 3, for each factor, the square root of AVE is significantly larger than its correlation coefficients with other factors. This indicates that each construct is more closely related to its own measures than to those of other constructs. Therefore, discriminant validity is supported.

5.2 STRUCTURED EQUATION MODELLING

This paper tests the validity of the model using structured equation modelling by combining measurement equation and structured equation.

We assessed the structural model to determine its explanatory power and the significance of the hypothesized paths. Figure 2 shows the results of the PLS analysis for the research model. Explained variance of perceived usefulness, Task-technology fit, Actual System use is 43%, 51% and 32% respectively.

The level of significance is 0.000 and most of path coefficients are significant in the level of P<0.001 significance level, and all the hypotheses were supported. X2 /d.f. value is 2.87. According to suggestion of Joreskog and Sorbom, it is acceptable between 2 and 5. Other results are shown as Table4.

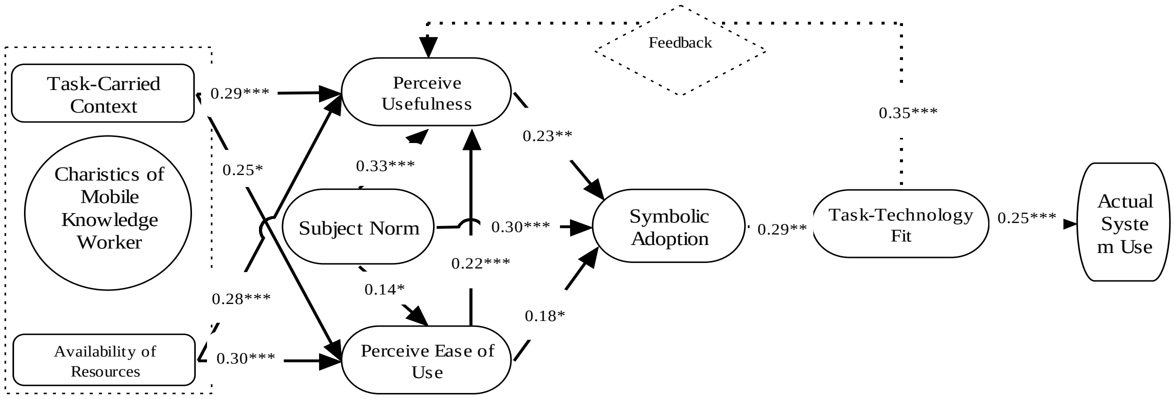


FIGURE 1 Hypothetical model

TABLE 4 Model evaluation overall fit measurement

Measure	Value
X2/df.(2-5)	2.89
Root mean square residual(RMR)(<0.05)	0.03
Goodness of fit index(GFI)(>0.9)	0.91
Normed fit index(NFI)(>0.9)	0.95
Non-normed fit index(>0.9)	0.93
Comparative fit index(>0.9)	0.94
Root mean square error of approximation(RMSEA)(<0.05-0.08)	0.04

5.3 RESULT DISCUSSION

On the above data analysis, we can make the following discussion:

(1) The results show that, the features of mobile office have a significant positive effect on perceived ease of use and perceived usefulness. Therefore, a successful mobile office service system should not only provide the functions required, it should also be possible to provide maximum convenience for the customers, Such as the user interface vivid, friendly and intuitive, with easy to operate, easy maintenance and so on. In terms of the staff, useful perception is also not limited to the software interface, different software designs that affect the matching degree of the mission requirements and Information-technology will also influence it.

(2) In this study, users' perceived usefulness have a significant positive impact on perceived ease of use, while the impact on symbolic acceptance is not obvious. According to Davis and others' study, users' perceived ease of use will intensify its perceived usefulness of technology, the influence of perceived usefulness on people's attitude is relatively significant at the stage of the technology was just launched. With the technology develop after a period; its significance will be obviously decreased. However, this research shows that, usefulness in the initial stage of technology does not affect the technology adoption significantly; this may be the impact of subjective norm. Because Venkatesh and Davis think, on the mandatory condition, the users will break the perception of ease of use and usefulness and be forced to choose a new technology application. The conclusion on Subjective Norm in this paper seems to support this view from the side: the positive influence of Subjective Norm on perceived ease of use and symbolic acceptance is enormous, but little on perceived usefulness.

(3) In this study, the matching relationship between symbolic acceptance and task-technology is proposed the first time, but it has also been supported by good data. Investigating its cause deeply, mainly because symbolic acceptance has been the substantial but not necessary conditions of tech-application when mandatory, and has meet the pre-condition of TTF research. The task-technology match makes a further promotion of users' ease of use perception, and is further upgraded to increase self-efficacy, thus has solved the issue how the technology affects the effectiveness.

6 Conclusion

Mobile knowledge worker is not a new concept, but because of the limitations of technology and a variety of factors, its practical application is still in its infancy. With the rapid development of mobile commerce, industry and researchers are interested in mobile office for its innovative technical characteristics, so to understand what factors affect the needs of the Enterprises of the adoption of mobile office technology and the nature of the adoption of mobile office brought with it the need. Based on TAM theory and TTF theory, this paper presents an embedded TAM/TTF model, empirical studies have shown that this model has a better interpretation of results.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No. 71071140 and 71301070005), National Natural Science Foundation of Zhejiang Province (Grand No.Y1090617), Key Innovation Team of Zhejiang Province (Grand No. 2010R50041) and Modern business centre of Zhejiang GongShang University.

References

[1] Gartner 2002 *Trends and Developments in Wireless Data Applications* (TCMC-WW-FR-0116) Aug 13

[2] Yankee Group 2005 Mobile workers number almost 50 million *Business Communications Review* **35**(8) 8-12

[3] The IDC study Worldwide Mobile Worker 2007-2011 Forecast and Analysis (IDC #209813) Business Wire, Jan 15, 2008

[4] *The People's Republic of China Ministry of Information Industry* Telecommunications industry in August 2008 Monthly Bulletin of Statistics [EB/OL] [2008-09-13] http://www.mii.gov.cn/col/coll66/index.html

[5] China Internet Network Information Center, 24th China Internet Development Statistics Survey July, 2009

[6] Chu Yan, Huang Lihua 2005 Mobile Business Applications Adoption Model Based on the Concepts of Task-Technology Fit *Proceedings of ICSSSM '05. IEEE Publishers* **2** 13-15

[7] Yao-bin LV, Chao-hua DENG, Zhi-yu CHEN 2008 *Mobile commerce application mode and the adoption of research* Science Press **9**

[8] Chen Tianjiao, Xu Zhengchuan 2005 Mobile Commerce Adoption Research: A Literature Review and a Proposed Framework *Proc. of CoDE, July 29-30, Shanghai, China*

[9] Zheng W, Yuan Y 2007 Identifying the differences between stationary office support and mobile work support: a conceptual framework *Int. J. Mobile Communications* **5**(1) 107–22

[10] Davis F D, Bagozzi R P, Warshaw P R 1989 User acceptance of computer technology: A comparison of two theoretical models *Management Science* **35**(8) 982-1003

[11] Legris Paul, Ingham John, Collerette Pierre 2003 Why do people use information technology? A critical review of the technology acceptance model *Information & Management* **40**(3) 191-204

[12] Brown S A, Massey A P, et al. 2002 Do I really have to? User acceptance of mandated technology *European Journal of Information Systems* **11**(4) 283-95

[13] Pozzebon M 2002 Combining a saturation approach with a behavioral-based model to investigate ERP usage *Americas Conference on Information Systems* Houston, TX

[14] Rawstorne P, Jayasuriya R, Caputi P 1998 An integrative model of information systems use in mandatory environments *International Conference on Information Systems* Helsinki, Finland

[15] Karahanna E 1999 Symbolic adoption of information technology *International Decision Science Institute* Athens, Greece

[16] Fiona Fui-Hoon Nah, Xin Tan 2004 Soon Hing Teh An Empirical Investigation on End - Users' Acceptance of Enterprise Systems *Information Resources Management Journal* **17**(3) 32-53

[17] Goodhue D L, Thompson R L 1995 Task-technology Fit and Individual Performance *MIS Quarterly* **19**(2) 213 - 36

[18] Goodhue D L 1995 Understanding User Evaluations of Information Systems *Management Science* **41**(12) 1827-44

[19] Zigurs I, Buckland B K A 1998 Theory of Task/Technology Fit and Group Support Systems Effectiveness *MIS Quarterly* **22**(3) 313-34

[20] Dishaw M T, Strong D M 1998 Assessing software maintenance tool utilization using task-technology fit and fitness-for-use models *Journal of Software Maintenance: Research and Practice* **10**(3) 151-79

[21] Benslimane Y, Plaisent M, Bernard P 2003 Applying the Task-Technology Fit Model to WWW-based Procurement: Conceptualization and Measurement *IEEE Proceedings of the 36th Hawaii Internat ional Conference on System Sciences* (I) 6-9

[22] Hao Tian, Dongsheng Liu, Jianming Lin, Yongqin Jin 2010 What drives mobile office service? A revised technology acceptance model based on the characteristics of Wireless mobile office technology *International Conference of Information Science and Management Engineering*

[23] Dishaw M T, Strong D M 1999 Extending the Technology Acceptance Model with Task-technology Fit Constructs *Information & Management* **36**(3) 9-21

[24] Szajna, Bernadette Davis et al. 1996 Empirical evaluation of the revised technology acceptance model *Management Science* **42**(1) 85-92

[25] Gefen D, Straub D 1997 Gender differences in the perception and use of E-mail: An extension to the technology acceptance model *MIS Quarterly* **21**(4) 389-400

[26] Patrick Y K, Chau Paul, Jen-Hwa Hu 2001 Information Technology Acceptance by Individual Professionals: A Model Comparison Approach *Decision Sciences* **32**(4) 699-719

[27] Kamal A Munir 2003 Competitive dynamics in face of technological discontinuity: a framework for action *Journal of High Technology Management Research* **14**(1) 93-109

[28] Elena Karahanna, Detmar W Straub, Norman L 1999 Chervany. Information Technology Adoption Across Time: A Cross-Sectional Comparison of Pre-Adoption and Post-Adoption Beliefs *MIS Quarterly* **23**(2) 183-213

[29] Chiu D K W, Cheung S C, Kafeza E, Leung H F 2003 A Three-Tier View Methodology for adapting M-services *IEEE Transactions on Systems Man and Cybernetics* Part A **33**(6) 725-41

[30] Leung F S K, Cheung C M K Consumer Attitude toward Mobile Advertising *Proceedings of the Tenth Americas Conference on Information Systems, New York, August 2004*

[31] Anderson J, Gerbing D W. 1988 Structure Equation Modelling in Practise: A Review and Recommended Two-step Approach *Psychological Bulletin* **103**(3) 411-23

Authors

Yongqin Jin, born on June 6, 1968

Current position, grades: a senior engineer in Department of Finance of Zhejiang Province
Scientific interest: data mining, electronic commerce and wireless network.
Experience: Yongqin Jin received his master degree in 2002, Zhejiang University, China. He is currently a senior engineer in Department of Finance of Zhejiang Province. His research interests include data mining, electronic commerce and wireless network.

Dongsheng Liu

Current position, grades: an associate Professor in the school of information engineering at ZJGSU
Scientific interest: data mining, electronic commerce and wireless network
Experience: Dongsheng Liu received his PHD degree in the school of information engineering, in 2008, Zhejiang Gongshang University (ZJGSU), China. He is currently an associate Professor in the school of information engineering at ZJGSU. His research interests include data mining, electronic commerce and wireless network

Application of magnified BP algorithm in forecasting the physical changes of ancient wooden buildings

Jie Liu[1]*, Qiusi Dai[2], Huifeng Yan[3]

[1] *School Of Civil Engineering & Architecture, Chongqing Jiaotong Univ., Chongqing, China*

[2] *College of Architecture and Urban Planning, Chongqing University, Chongqing, China*

[3] *Chongqing University of Posts and Telecom, Chongqing, China*

Received 1 March 2014, www.tsi.lv

Abstract

CA using neural network model of ancient buildings to predict changes in the physical properties of Applied X-ray detector collection of ancient buildings grey wood elements, so that the ancient wooden building components of each pixel grayscale and Neural Network CA model correspond to each cell, using the CA model "grey" concept learning through the improved BP Algorithm to calculate the grey value of each cell changes, so as to arrive ancient architectural wood elements over time the case of damage by example through the projections obtained wood over time damage to the picture.

Keywords: ancient building, BP neural network, cellular Automata (CA)

1 Introduction

In order to protect the ancient architecture, the destruction of the natural environmental factors is predicted on ancient building components. Since changes in the various factors that affect the objects often have irregular and unpredictable, difficult to describe traditional mathematical methods ancient building components change with time in the process. With the artificial neural network theory and technological developments, the theory has been widely applied, we use cellular automata (Cellular Automata) CA) and magnified BP algorithm [2] combines elements of ancient buildings damaged by natural environmental factors to predict. Application of X-ray scanning of ancient building components for 2D data, through the CA calculated for each neuron grey value changes that alter the ancient building components corresponding to a single bitmap pixel grey values to predict the damage of ancient building components extent.

2 CA model

In CA model, the state of the cell is only 0 or 1 and a discrete value, it can not reflect the state of continuous change process objects in the CA model binding, use ”grey” concept $G_d^t\{x,y\}$ to represent and reflect element mesh {x, y} state of continuous change process when the grey value $G_d^t\{x,y\}$ gradient from 0 to 1, it indicates that the cell directly in good condition from the beginning is completely transformed into damage state.

CA based binding model formula is:

$$P_d^t\{x,y\} = f\left(S^t\{x,y\},N\right)\times CONS_d^t\{x,y\}, \qquad (1)$$

$P_d^t\{x,y\}$ is the state of development, S^t is the state, N is the close range, $CONS_d^t\{x,y\}$ is the total binding conditions, {x, y} is the cell specific location.

Total binding coefficient formula is:

$$CONS_d^t\{x,y\} = \prod_{i=1}^{n} cons_{id}^t\{x,y\}. \qquad (2)$$

Gradation value increases $\Delta G_d^t\{x,y\}$ with transition probability and the coefficient value is proportional to the total binding:

$$\Delta G_d^t\{x,y\} = P_d^t\{x,y\}\times CONS_d^t\{x,y\}. \qquad (3)$$

At a time, *t + 1* the grey value of $G_d^{t+1}\{x,y\}$ may be used to make predictions iteration:

$$G_d^{t+1}\{x,y\} = G_d^t\{x,y\} + \Delta G_d^t\{x,y\},\ G_d^t\{x,y\}\in(0,1). \qquad (4)$$

Based on different grey, a cell {x, y} at a certain moment *t+1*, There are 3 possible states: in part or develop or remain undisturbed. Be expressed as:

***Corresponding author** e-mail: 41934023@qq.com

$$S_d^{t+1}\{x,y\} = \begin{cases} part-develop, & G_d^t\{x,y\}=1 \\ develop, & 0<G_d^t\{x,y\}<1 \\ S^t\{x,y\} & G_d^t\{x,y\}=0 \end{cases}, \qquad (5)$$

Through the introduction of local, regional and global constraint conditions and the “grey” concept, binding objects CA model for predicting the evolution of the ability to significantly enhance.

3 Ancient building components analysis

3.1 PHYSICAL PROPERTIES OF THE EVOLUTION OF ANCIENT COMPONENT ANALYSIS

In order to predict the erosion of ancient building components over time by natural factors, we use the model of the ancient building element numerical processing. Numerical ancient component model is a 2D grayscale graphics that make access to ancient architecture model for each pixel grayscale and Neural Network CA model each cell corresponds use pixels as cellular, CA model can easily be combined and image processing systems.

The ancient building element numerical information is captured. The neural network model is utilized to calculate the CA of the grey value of each cell. The temperature, humidity and other natural factors predictive value of historical data and neural networks to predict the combination of CA model, through changes in individual neurons grey, visually display the ancient architectural elements about the trend.

The physical properties of the evolution of ancient component analysis system structure shown in Figure 1.

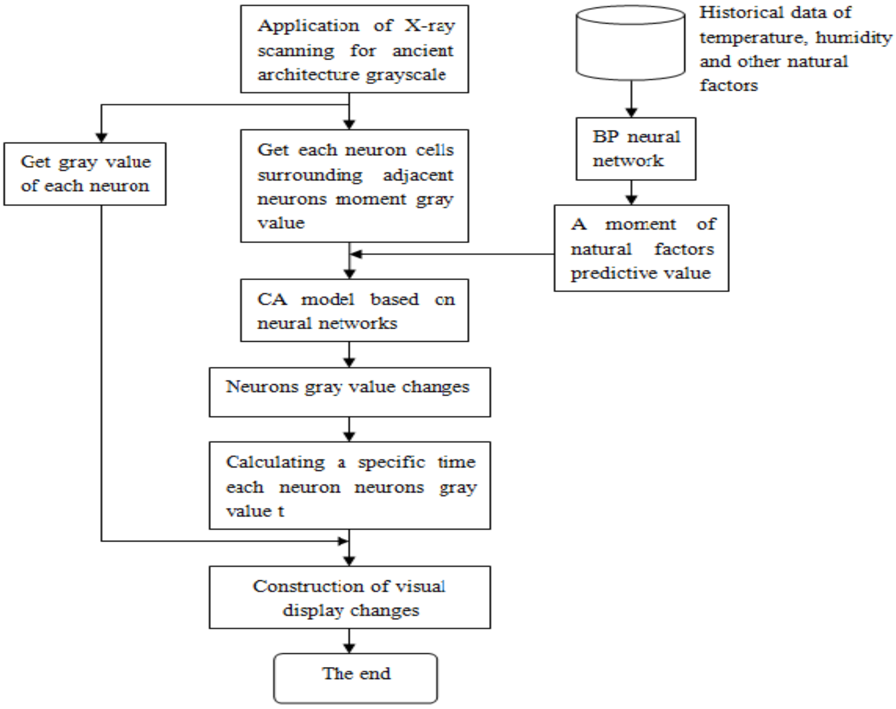

Application of X-ray scanning for ancient architecture grayscale
Historical data of temperature, humidity and other natural factors
Get gray value of each neuron
Get each neuron cells surrounding adjacent neurons moment gray value
BP neural network
A moment of natural factors predictive value
CA model based on neural networks
Neurons gray value changes
Calculating a specific time each neuron neurons gray value t
Construction of visual display changes
The end

FIGURE 1 The physical properties of ancient building components analysis system architecture evolution

3.2 GRAY VALUE OF A SINGLE CELL

Ancient building element numerical model of treatment, the each pixel as a cell, a grey image (grey-scale) to said neural network cellular CA model of a single grey value.

Input from the image, each pixel with a particular grey value will correspond. The element of the grey value is set to D = f (x, y), after changing the gradation grey value, the grey value changes can be expressed as or. Among them, the set D and are in the image grey values within the specified range.

Function for the grey-scale transformation function, which describes the input grey value and the output gradation value conversion relationship between the neural network model is calculated for each cell CA grey value changes of the cell and through to change the grey levels of the corresponding cell and the grey value of the pixel.

3.3 CA ALGORITHM BASED ON MAGNIFIED BP NEURAL NETWORK MODEL[2]

Research on magnified BP [2] neural network is used. According to projections need to set 11 in the input layer neurons, each of the input neurons were 11 decisions ancient architectural elements correspond to the grey value changes.

The selection of numbers of neurons in the hidden layer affects the outcome of accuracy, training time and fault tolerance. Generally, the more hidden layer neurons, the more accurate the results are, but too many neurons in the hidden layer will increase the training time while hidden layer neurons will cause an increase in network fault tolerance decreases. studies show that for the n-layer neural network, the hidden layer neurons number is at least , where n is the number of neurons in the input layer after the above analysis, neural network hidden layer neurons number, in order to predict the best results, the hidden layer neurons is set to 6 in the output layer, by a neuron to the output grey numerical value by changing the output of each cell corresponding pixel grey value established neural network model shown in Figure 2.

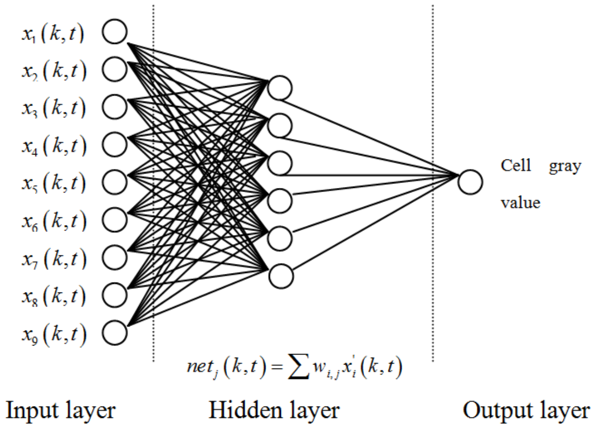


FIGURE 2 BP neural network model

Neural network hidden layer using Sigmoid transfer function, which makes the learning process is not high speed and sensitivity training and easy access to the saturation order to improve the speed and agility training and Sigmoid function effectively avoid the saturation region, the general requirements of the input data values between 0 and 1. This phenomenon in order to reduce the possibility of platforms, accelerate the learning speed, the input samples are normalized, handling as follows:

$$P_{\max}=\max\{P\},P'=P/P_{\max}, \tag{5}$$

wherein, P is input, $P_{\max}$ is processed through the normalized experimental data.

3.4 APPLIED ALGORITHM DERIVATION

Wood in the natural environment by a variety of natural factors, the influence of different factors, wood elements will produce different changes, shown in Figure 3.

Changes affecting the natural wood ancient architecture factors as temperature, humidity, microbes, the above three factors as the input layer three input data, however, the model of ancient buildings at the development of a cellular change is not accidental, for each element cell consists of a state transition to another state is a continuous process. Grey value of each cell to be affected by changes in natural wood ancient buildings factors as temperature, humidity, microbes, the above three factors as the input layer, three input data, however, ancient architecture model development at a cellular change is not accidental, each cell consists of a state to another state is a continuous process. Each cell of the grey values are it is subject to the grey values of the adjacent cell, therefore, needs to be around 6 neighbouring cell at time t_{n-1} of the 6 grey value as the input data input layer. The input is:

$$X(k,t)=\left[x_1(k,t),x_2(k,t),x_3(k,t),...,x_8(k,t),x_9(k,t)\right]^T ,(6)$$

where, $x_i(k,t)$ represents t time's i variable of the k cell in the simulation time.

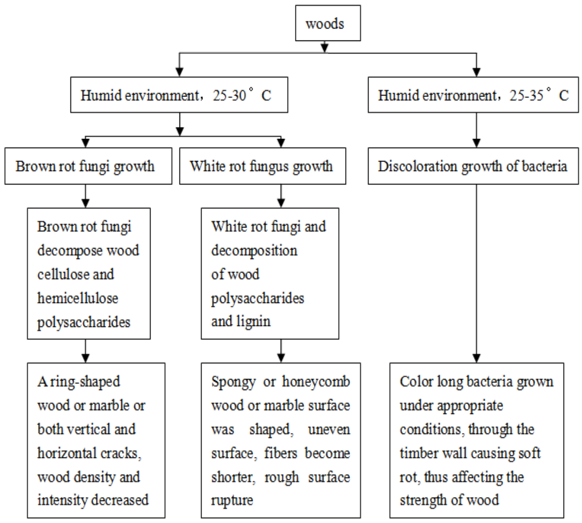

woods
Humid environment, 25-30° C
Humid environment, 25-35° C
Brown rot fungi growth
White rot fungus growth
Discoloration growth of bacteria
Brown rot fungi decompose wood cellulose and hemicellulose polysaccharides
White rot fungi and decomposition of wood polysaccharides and lignin
A ring-shaped wood or marble or both vertical and horizontal cracks, wood density and intensity decreased
Spongy or honeycomb wood or marble surface was shaped, uneven surface, fibers become shorter, rough surface rupture
Color long bacteria grown under appropriate conditions, through the timber wall causing soft rot, thus affecting the strength of wood

FIGURE 3 Wood is affected by changes in environmental factors

Input data normalization, the data passed to neural network input layer and then the data is output to the input layer hidden layer, each hidden layer neurons will accept various input layer neuron input. Neural network structure by the CA It can be seen, there are eight hidden layer neurons, neuron j-th received signal is expressed as

$$net_j(k,t)=\sum_i w_{i,j}x_i(k,t), \tag{7}$$

wherein, $net_j(k,t)$ represents the j neuron of the received signal, $w_{i,j}$ is the weight value.

Hidden layer obtained signal, the transmission signal to the output layer, to obtain the final output. Hidden layer response function is:

$$f\left[net_j(k,t)\right]=1\Big/_{1+e^{-net_j(k,t)}} . \tag{8}$$

The output of the output layer (i.e., the grey value of each cell) is $\sum w_{j,l}\, 1\Big/_{1+e^{-net_j(k,t)}}$, where $w_{j,l}$ is the weight value.

At this point, you can through the cell to change the grey value changes ancient architectural model of each pixel grayscale display changes.

In order to make the CA model simulation results closer to the actual situation, the introduction of random variables RA.

$$RA = 1 + \left(-\ln\gamma\right)^{\alpha}, \tag{9}$$

wherein γ represents the [0, 1], a random number, α is the size of the control parameters of the random variable.

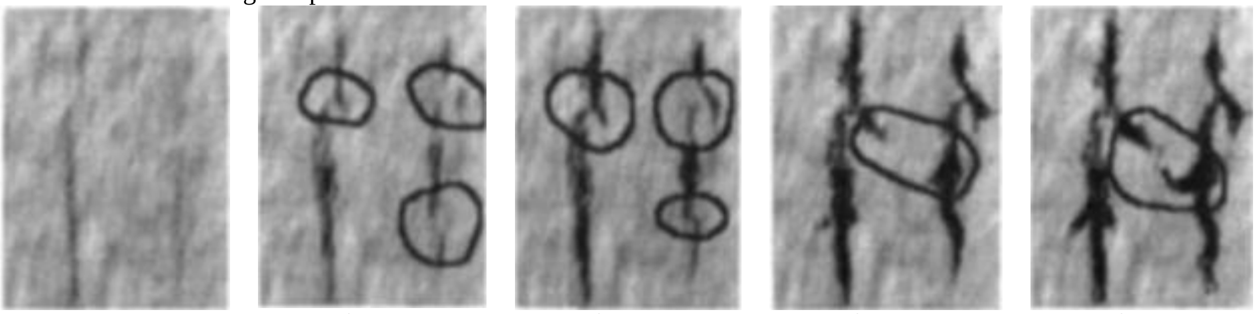
Adding random variable, the grey value equation becomes:

$$P\left(k,t,l\right) = RA \times \sum_{j} w_{j,1} \, {}^{1}\!\Big/_{1+e^{-net_j(k,t)}} \,. \tag{10}$$

4 Application Examples

There is using CA model neural network structure of the ancient wooden building components to an area within 20 years to predict the evolution of the physical properties analysis.

Figure4(a) is an ancient architectural wood elements grey area image, Figure4(b) is the prediction image using CA model of neural network after 5 Years. It is seen that Figure4(a) in the shallow cracks in Figure4(b) to further expand, and the original grey value of the cracks at the light grey transformed into a dark grey colour, while in Figure4(b) new cracks is generated at the circle. Figure4(c) to predict the results of 10 years, and Figure4(b) contrast can be seen further expansion cracks, Old cracks at the original grey levels improved, in the circle the cracks are combined. Figure4(d) and Figure4(e) are image of results after 15 years and 20 years. Two figures contrast can be seen, while expanding cracks extending in new small cracks, cracks have to mesh trend. Circle from two graphs at clearly see that after five years of change, in drawing circle the rift is combined together.



a) Original image b) After 5 years c) After 10 years d) After 15 years e) After 20 years

FIGURE 4 Predict the evolution of the wood elements

Based on the above, for the collection of ancient buildings wood elements grayscale, CA model of neural network is used to predict the wood elements. through neural network training the "grey" concept of CA model is used to predict each a cellular changes in the grey values and to calculate the wood elements of ancient buildings and damaged over time the situation. Therefore, the neural network model of architectural wood elements CA extent of the damage prediction analysis over time had a significant effect.

5 Conclusion

In the study, measures are taken to make numerical treatment of the ancient building models and neural networks for each pixel in the CA model each cell one correspondence, and the use of the "grey" concept of CA model, based on neural network training to learn to predict the grey value of each cell changes, and then come to the ancient architectural models each pixel grey change, obtaining a prediction ancient buildings being eroded due to natural environmental factors effect. Studies have shown that this method of ancient buildings due to natural environment in a variety of factors and the degree of erosion prediction is valid.

Acknowledgments

The authors wish to thank Li Qiang. This work was supported in part by a grant from Zhang Wei.

References

[1] Wolfram S 1986 *Theory and Applications of Cellular Au-tomata* Singapore *World Scientific*

[2] Wang-Genda, Liu-Heping, Wang-yunjian 2008 A Magnified BP Algorithm with Fast Convergence *Computer Simulation* **25** 162-72 June 2008

[3] Gong Ping, Shipei Ji, Wei Wei 2012 Based on BP Artificial Neural Network regional ecological warning *Agricultural Research* **30** 211-6 May 2012

[4] Yuan Shou Qi, Shen Yan Ning, Zhang Jinfeng 2009 Based on improved BP neural network composite impeller centrifugal pump performance prediction *Agricultural Machinery* **25** 211-6 May 2009

[5] He Qingfei, Chen Guiming, Chen Xiaohu 2011 Life Prediction of Hydraulic Pump Based on the Improved Grey Forecasting Model *Lubrication Engineering* **36**(7) 27-31 March 2011

Authors

Liu Jie, born on November, 1973, Sichuan, China

Current position, grades: a teacher in School Of Civil Engineering & Architecture, Chongqing Jiaotong Univ.
University studies: Bachelor and Master degree in Architecture from the ChongQing University in China.
Scientific interest: Digitized intelligent building, data mining, intelligent control.
Publications: She has more than 8 publications to his credit in international journals and conferences.

Dai Qiusi, born on May, 1973, Sichuan, China

Current position, grades: a teacher in College of Architecture and Urban Planning, Chongqing University, Chongqing University.
University studies: Bachelor and Master degree in Architecture from the ChongQing University in China.
Scientific interest: Architectural Theory and History, digitized intelligent building, data mining, intelligent control.
Publications: more than 16 publications to his credit in international journals and conferences.

Yan Huifeng, born on June, 1976, Henan, China

Current position, grades: Bachelor and Master degree in computer science from the ChongQing University in China.
University studies: a teacher in College of Mobile Telecommunications ChongQing University of Posts and Telecom
Scientific interest: machine learning, data mining and Pattern Recognition.
Publications: more than 10 publications to his credit in international journals and conferences.

The Semi-quantitative evaluation method and application of the risks of geological disaster of the Shaan–Jing pipeline

Cunjie Guo[1*], Wei Liang[2], Laibin Zhang[2]

[1] *College of Mechanical and Transportation Engineering, China University of Petroleum-Beijing, Beijing, China*

[2] *College of Mechanical and Transportation Engineering, China University of Petroleum-Beijing, Beijing, China*

Received 1 March 2014, www.tsi.lv

Abstract

Based on the index scoring method, the semi-quantitative method for assessment of pipeline geological disaster risks calculates the relative risk of a disaster by investing and assessing the objective existence state of index in accordance with pre-determined scores and weights, and meets the requirements of risk prioritizing and ranking at the geological disaster investigation stage so as to guide the development of risk control planning. A geological disaster risk semi-quantitative assessment system and risk grading standards both of which are applicable to oil and gas pipelines have been established. What has been developed also includes the pipeline geological disaster risk management system software, which integrates the risk semi-quantitative assessment technique based on the index-scoring-method, and other techniques such as information management and risk management, and thus provides a platform of information, technology and management for the management of pipeline geological disaster risks. This method has been used for a unified risk assessment of more than 3300 disaster points along the oil and gas pipeline, and satisfactory evaluation results are obtained, thus providing an important basis for the development of planning of pipeline geological disaster risk remediation.

Keywords: oil and gas pipeline, geological disasters, index scoring method, semi-quantitative assessment, risk grading

1 Introduction

Geological disasters are one of the major risks for oil and gas pipelines and even become the NO.1 risk threatening the safety of oil and gas pipelines in mountain areas with complex geological and geomorphic conditions. Due to randomness and unpredictability of geological disasters, pipeline geological disaster risk management has gradually become a major means for prevention of pipeline geological disasters. The risk assessment technique is one of the main supporting techniques for the management of pipeline geological disaster risks, and its purpose is to conduct risk calculation and evaluation of identified risk points, rank risks according to the size of them and provide a basis for risk control planning. Currently, pipeline geological disaster risk assessment methods can be divided into three categories: qualitative assessment, such as the risk matrix method; semi-quantitative assessment, such as the risk index method; quantitative assessment, such as the probability assessment method [1]. For the pipelines laid in mountain areas in Midwest China, each one is faced with different kinds of geological disasters along the line, and the number of disaster points for each pipeline is often as many as several hundreds or even several thousands [2-5]. The qualitative method cannot meet the requirements while the quantitative method is inoperable to rank risks for such a huge number of disaster points, so the semi-quantitative method is the ideal choice.

The semi-quantitative assessment method developed for the GRM system by Canadian BGC Company in 2002 can be used for semi-quantitative risk assessment and risk ranking of landslides, dilapidation and brook & road flood destruction. The West-East Gas Pipeline Environmental Geological Disaster Risk Assessment System [6] developed by West-East Gas Pipeline Company together with Southwest Petroleum University in 2006 can conduct semi-quantitative risk assessment of 9 kinds of disasters faced by the west-east gas pipeline, such as water destruction, collapsible loess, collapse of mined-out area and debris flow.

2 Risk assessment model

The Department of Humanitarian Affairs of the United Nations published a definition of the natural disaster risk in 1992, and used Formula (1) to represent a risk [6]:

$$\text{Risk} = \text{Hazard} \times \text{Vulnerability}. \tag{1}$$

The formula is the rationale for assessment of natural disaster risks. Hong Kong Civil Engineering and Development Department uses Formula (2) to evaluate the annual risk of landslides. This formula also applies to risk assessment of other disasters such as pipeline landslides (with the pipeline as the hazard-bearing body), dilapidation, debris flow, collapse of mined-out area and water destruction.

$$R_{\text{prop}} = P_{\text{L}} \cdot P_{\text{T,L}} \cdot P_{\text{S,T}} \cdot V_{\text{prop,S}} \cdot E_{\text{s}}\,, \tag{2}$$

**Corresponding author* e-mail: guocunjie@sina.com where: R_{prop} is the annual loss of property value, with the unit being a currency or in another form; P_{L} is the frequency of landslides, dimensionless, ranging from 0 to 1; $P_{\text{T, L}}$ is the probability of landslides reaching the hazard-bearing body, dimensionless, ranging from 0 to 1; $P_{\text{S,T}}$ is the space-time probability of the hazard-bearing body, dimensionless, ranging from 0 to 1; $V_{\text{prop, S}}$ is the vulnerability of the hazard-bearing body to landslides, dimensionless, ranging from 0 to 1; E_{S} is the hazard-bearing body (such as value or existing net property value), with the unit being a currency or in another form.

Once laid, the location of the oil and gas pipeline is generally fixed and can be clearly identified, so $P_{\text{T, L}} = 1$. The probability of disaster occurrence is relevant to the geological disaster's natural state and environment, and will also be relevant to the effectiveness of engineering measures if these measures are taken against the geological disaster; the probability of the disaster reaching the pipeline is related to the spatial locations of the two; the probability of destruction of the pipeline under the influence of disaster is relevant to parameters such as the disaster scale, the length of affected part of the pipeline and the pipeline's strength, and the vulnerability of the pipeline will also be related to the effectiveness of protective measures if these measures are taken against the disaster. Therefore, Formula (2) can be adjusted to Formula (3):

$$R = H\left(1-H'\right)SV\left(1-V'\right)E\,, \tag{3}$$

where: R is the pipeline's risk of a geological disaster with the unit being a currency or in another form; H is the probability of occurrence of a geological disaster under natural conditions, i.e. disaster liability, dimensionless, ranging from 0 to 1; H' is the probability of preventive measures playing a full role (completely preventing occurrence of the disaster), dimensionless, ranging from 0 to 1; S is the probability of a geological disaster affecting the pipeline after its occurrence, dimensionless, ranging from 0 to 1; V is the probability of pipeline failure (strength failure or instability) without any protections against a geological disaster that has occurred, i.e. pipeline vulnerability, dimensionless, ranging from 0 to 1; V' is the probability of pipeline protective measures playing a full role (completely preventing pipeline damages), dimensionless, ranging from 0 to 1; E is the consequence of pipeline failure, i.e. economic losses due to transmission media leaks, service outages and adverse effects of leaked media on the environment after pipeline failure, with the unit being a currency or in another form.

If E, the consequence, is not to be considered. Formula (3) can be adjusted to Formula (4):

$$P_{\text{R}} = H\left(1-H'\right)SV\left(1-V'\right), \tag{4}$$

where: P_{R} is the risk probability of the pipeline suffering from a geological disaster, i.e. the probability of pipeline failure under the current disaster environment and pipeline state, dimensionless, ranging from 0 to 1.

This model is a quantitative evaluation model, which can be used to calculate the real risk of a pipeline geological disaster, and can be used for semi-quantitative risk assessment. The purpose of semi-quantitative assessment of pipeline geological disaster risks is to prioritize and rank risks. Therefore, semi-quantitative risk assessment only needs to assess the relative risk of a disaster instead of calculating the actual risk of the disaster, and can use a relative value (index) to represent each parameter in Formula (3) to respectively evaluate risk probability and the consequence.

3 Risk probability evaluation

A. Index scoring method

The index scoring method is selected to evaluate the risk probability exponent of a pipeline geological disaster. The index scoring method is a major influence factor for statistical analysis of risk probability. All influential factors are used as evaluation indexes to establish an evaluation index system. Taking into account the impact of each index on risk probability, each index is given an appropriate weight. Then the possible state of each index is analysed to judge the level of risk probability in different states and give a score to each index (Table 1). Therefore, the evaluation system is made up of evaluation index system, weight, score and algorithm. In practical application, after the state of each index is investigated, the risk probability exponent can be calculated based on the weights, the index state score and the algorithm.

TABLE 1 evaluation parameters of different methods

Index u	Weight w	Index State	Score
		State 1	u_{11}
Index 1	w_1	State 2	u_{12}
		…	…
		State 1	u_{21}
Index 2	w_2	State 2	u_{22}
		…	…
…	…	…	…
		State 1	un_1
Index n	w_n	State 2	un_2
		…	…

B. Calculation method

According to Formula (4), risk probability assessment can be divided into five assessment parts. An index system is established for each assessment part and the sum of weights in system is one. The maximum score of index state scoring is 10. The value of risk probability is a decimal no greater than 1. According to the index scoring method, Formula (4) can be expressed as Formula (5):

$$P_R = \frac{\sum_{i=1}^{n_1} u_{1i} \times w_{1i}}{10} \left(1 - \frac{\sum_{i=1}^{n_2} u_{2i} \times w_{2i}}{10} \right) \frac{\sum_{i=1}^{n_3} u_{3i} \times w_{3i}}{10} \frac{\sum_{i=1}^{n_4} u_{4i} \times w_{4i}}{10} \left(1 - \frac{\sum_{i=1}^{n_5} u_{5i} \times w_{5i}}{10} \right), \quad (5)$$

where: n_1 is the number of evaluation indexes for the occurrence probability of a geological disaster under natural conditions; u_{1i} is the score of the state of the i-th (No. i) evaluation index for the occurrence probability of a geological disaster under natural conditions; w_{1i} is the weight of the i-th (No. i) evaluation index for the occurrence probability of a geological disaster under natural conditions.

Therefore, the establishment of a risk probability calculation system needs to determine the risk probability evaluation index system, index weights and the index state score.

C. Establishment of a risk probability assessment index system

To ensure the operability and applicability of the evaluation system and objectivity of the evaluation result, the determination of evaluation indexes and index states shall follow the following principles: selecting main factors having significant impacts on the evaluation result as indexes to form a slimmer but better index system; index states should be intuitive, easy to get, and not experience-dependent to avoid the interference of human factors; indexes and index states should be widely applicable, and can be applied to most disasters of the same category. Based on this, an index system (Table 2) has been established for 6 categories (11 kinds) of disasters, namely landslides (clay landslides, debris landslides, loess landslides, rock landslides), dilapidation, debris flow, collapse of mined-out area, water destruction (slope water destruction, platform & farmland water destruction, brook & road water destruction) and loess collapses. Each index generally has 3 to 4 index states.

TABLE 2 evaluation parameters of different methods

Disaster Category	Number of Indexes
Clay Landslides	21
Debris Landslides	25
Loess Landslides	19
Rock Landslides	21
Dilapidation	27
Debris Flow	23
Collapse of Mined-out Area	25
Slope Water Destruction	13
Platform & Farmland Water	11
Destruction Brook & Road Water	14
Destruction Loess Collapses	14

A. Determination of index weights and index state scores

Index weights are essential for the accuracy of the result. Calculate the weight of each evaluation index using the Analytical Hierarchy Process (AHP) method and according to the relative importance of each index provided by several experts of pipeline geological disasters. Score each index state according to the level of risk in each index state and the score should be between 10 and 0.

Because the evaluation index system varies as the disaster varies, the evaluation results of different disasters cannot be directly compared. To rank risk probabilities of different kinds of geological disasters, a unified standard has been adopted for the scoring of all states of all pipeline vulnerability evaluation indexes, making relative risk probabilities of different kinds of disasters basically comparable.

4 Failure consequence evaluation

The pipeline failure consequence evaluation model established by Muhlbauer W Kent [7] is used for risk consequence assessment. The pipeline failure consequence (expressed as an exponent) is calculated using Formula (6):

$$E = \mathrm{PH} \cdot \mathrm{SP} \cdot \mathrm{DI} \cdot RC, \quad (6)$$

where: PH is the product hazard coefficient, to be selected according to the product type, ranging from 5 to 10; SP is the leakage coefficient, to be selected according to the amount of leakage, ranging from 1 to 5; DI is the diffusion coefficient, to be selected according to the nature of the surrounding environment, the higher the value, the more conducive to diffusion, ranging from 1.5 to 5; RC is the receptor whose value is selected according to factors such as nearby residents, ecological environment and the resulted public opinion, ranging from 0.5 to 4.

The values of all parameters are determined by the corresponding indexes. The greater the value of E, the more serious the losses are.

5 Risk probability evaluation

A risk can be determined based on the risk probability and the consequence. As the calculated results are the risk probability exponent and the consequential loss exponent, a risk is expressed in the risk matrix form and the risk level is determined according to the risk probability exponent and the consequential loss exponent. Generally speaking, the ultimate goal of pipeline geological disaster prevention is to prevent pipeline failure accidents as far as possible, extra emphasis should be placed on classification of risk probabilities when ranking risks. To meet different production needs, the risk probability exponent, the risk consequence exponent and risks are respectively classified into 5 levels: (1) high level of risk, unacceptable, for which risk mitigation measures need to be implemented within the specified time; (2) relatively high level of risk, undesirable, for which preventive measures should be taken to reduce it, but the prevention cost needs to be assessed and limited, and selective inspection, professional monitoring or risk mitigation measures can be adopted; (3) moderate level of risk, conditionally acceptable, for which there should not be a substantial increase in cost of risk control, and selective inspection or easy monitoring can be adopted; (4) relatively low level of risk, acceptable, for which tour-inspection measures can be adopted; (5) low level of risk, negligible, for which no measures need to be taken and no file records need to be kept.

Risk grading standards should be determined based on the risk tolerance of pipeline operators. The risk tolerance, which is changing all the time, is different for different pipeline operators. Therefore, risk grading standards adopted by different pipeline operators are different and can be adjusted at any time. As risk grading is intended to provide a basis for risk control planning, so grading standards need to be developed based on previous prevention and control planning as well to ensure a certain degree of continuity. Risk grading standards are developed according to PetroChina's prevention and control planning for major pipelines in 2008. As the management cost of water destruction and loess collapses is relatively low, the risk probability grading standards are adjusted (in Figure 1, risk grading exponent values outside the brackets apply to landslides, dilapidation, debris flow, collapse of mined-out area, etc. while risk grading exponent values inside brackets apply to slope water destruction, platform & farmland water destruction, brook & road water destruction, and loess collapses) in consideration of risk - cost effectiveness.

TABLE 3 evaluation parameters of different methods

=0.4 (0.2)	High	High	High	High	High	High
0.2~0.4 (0.1~0.2)	Relatively High	Relatively High	Relatively High	Relatively High	Relatively High	High
0.1~0.2 (0.05~0.1)	Moderate	Moderate	Moderate	Moderate	Relatively High	Relatively High
0.05~0.1 (0.01~0.05)	Relatively Low	Relatively Low	Relatively Low	Moderate	Moderate	Moderate
<0.05 (0.01)	Low	Low	Low	Relatively Low	Relatively Low	Relatively Low
		Low <10	Relatively Low 10-90	Moderate 90-300	Relatively High 300-860	High =860

Risk Grading Exponent →

Consequential Loss Exponent →

6 Validation and application

A team made up of several experts of pipeline geological disasters is organized to conduct on-site confirmatory applications on the Shaanxi-Beijing gas pipeline in order to verify the operability and applicability of the pipeline geological disaster semi-quantitative risk assessment method which is based on the index scoring method and to verify the accuracy of the evaluation result. According to the result of applications, this method is easy to operate, and using this method, the field investigation generally takes no more than 20 minutes for a single disaster point; evaluation indexes are easily accessible and the evaluation result is less affected by human factors; (3) the reliability of the result is ensured as there is a high degree of consistency between the grading result worked out by the expert team through qualitative evaluation and the calculation result obtained using this method; (4) this method has considered similarities of disaster points, and therefore applies to most of disaster points.

As there are so many indexes and disaster points in an index system, the computer is needed for data management and computing. The pipeline geological disaster risk management system (software) is developed with the needs of pipeline geological disaster risk management work taken into account. This system, which integrates the index-scoring-method based pipeline geological disaster risk semi-quantitative assessment technique, can conduct semi-quantitative evaluation computing, and have functions such as information management (including information, photos, documents, etc. other than evaluation indexes), quantitative evaluation of landslides and dilapidation, risk control planning and management, and workflow management, thus providing an information platform, a technology platform and a management platform for risk management of pipeline geological disasters. The system is based on browser/server mode (B/S architecture), and authorized users only need to log on the Internet to use it. Currently widely used in the Shaan–Jing gas pipeline, this system evaluates risks of more than 3,000 disaster points along the pipeline in a unified way, and the evaluation result provides an important basis for the development of pipeline geological disaster risk control planning.

7 Conclusion

(1) Pipeline geological disaster risk management is a mainstream method of pipeline geological disaster protection. A risk assessment method needs to be established to make risk control planning for thousands of geological disaster points along the pipeline. The relative risks of disaster points should be ranked and graded.

(2) The pipeline geological disaster risk assessment model, which is established on the basis of the landslide disaster risk assessment model, can perform quantitative risk assessment and can also be used for semi-quantitative risk assessment.

(3) The index-scoring-method based pipeline geological disaster risk semi-quantitative assessment method can be used to calculate the relative risk of a disaster. Not only is the method very simple, but the assessment result is reliable, so it can meet the needs of pipeline geological disaster risk management.

Acknowledgments

Authors wishing to acknowledge assistance or encouragement from colleagues, special work by technical staff or financial support from organizations should do so in an unnumbered Acknowledgments section immediately following the last numbered section of the paper.

References

[1] Porter M, Reale R 2006 Geohazard Risk Management for the Nor Andino Gas Pipeline Calgary, *IPC* 2006-10208 5-4 February 2006

[2] Deng Q L 2007 A Report on Geological Disaster Investigation and Remediation Planning for the Zhongxian-Wuhan Gas Pipeline Wuhan: China University of Geosciences (Wuhan) **12** 12-3 August 2007

[3] Sichuan Institute of Geological Engineering Survey *A Report on Geological Disaster Investigation and Remediation Planning for the Lanzhou - Chengdu - Chongqing Pipeline* Chengdu: Sichuan Institute of Geological Engineering 18-9 April 2007

[4] Institute of Geological Disaster Prevention and Control under Gansu Academy of Science *A Report on Geological Disaster Investigation and Remediation Planning for the Sebei - Xining - Lanzhou Gas Pipeline* Lanzhou: Institute of Geological Disaster Prevention and Control under Gansu Academy of Science 21-2 August 2007

[5] PetroChina West-East Gas Pipeline Company *West-East Gas Pipeline Environmental Geological Disaster Risk Assessment* Shanghai: PetroChina West-East Gas Pipeline Company 522-4 November 2007

[6] Lu Q Z 2007 Geological Disaster Risk Assessment (Evaluation) Research Summary *Science of Disaster* **18** 59-63

[7] Muhlbauer W K 2005 *Pipeline Risk Management Manual (2nd Edition)* Translated by Yang Jiayu, Zhang Deyan, Li Chinhua, et. Beijing: China Petrochemical Press 98-122

Authors	
	Cunjie Guo was born in Shandong province, China **Current position, grades:** a Ph.D candidate in China University of Petroleum-Beijing, China **University studies:** bachelor and master degrees from China University of Petroleum-Beijing, China in 2002 and 2005, respectively. **Scientific interest:** pipeline geological disaster risk management and complex system reliability assessment
	Wei Liang, born on June 5, 1978, Shaanxi **Current position, grades:** Prof., Vice Dean **University studies:** China University of Petroleum, Beijing **Scientific interest:** Safety Monitoring **Publications:** 30 **Experience:** He has been engaged in the research of oil & gas pipeline leak detection and fault diagnosis of the storage and transportation equipment since 2002. He is now severing as an Prof. and doctoral supervisor in China University of Petroleum. And He is also severing as the vice dean of the Mechanical & Transportation College, and the dean of the Safety Engineering Department.
	Laibin Zhang, born on September 17, 1961, Anhui **Current position, grades:** Prof., President **University studies:** China University of Petroleum, Beijing **Scientific interest:** Mechanical Fault Diagnosis **Publications:** 120 **Experience:** Professor Lai-bin Zhang has been mainly engaged in the teaching and research work of the fault diagnosis technology of mechanical equipment and the technology of safety science and engineering. He has published more than one hundred papers in journals. Professor Lai-bin Zhang is the young academic pacesetter in Beijing and one of outstanding young teachers of CNPC (China National Petroleum Corporation). Besides, he is the deputy director of safety engineering education guiding committee, which is affiliated to The Ministry of Education. Also, he is the member of Beijing academic degrees committee and the executive director of Beijing HE (higher education) Academy. Furthermore, he is the member of SPE (Society of Petroleum Engineers) and the safety science and engineering evaluation group, which is affiliated to the State Council.

Estimation of forest volume based on LM-BP neural network model

Dasheng Wu[1, 2*]

[1] *Key Laboratory of Zhejiang Province about Forestry Intelligent Monitoring and Information Technology Research, Zhejiang A & F University, Lin'an 311300, Zhejiang, China*

[2] *School of Information Engineering, Zhejiang A & F University, Lin'an 311300, Zhejiang, China*

Received 1 March 2014, www.tsi.lv

Abstract

Since cost factors are of primary importance, continuously searching for more efficient and reliable estimation models that could integrate or, in some cases, substitute the traditional and expensive measuring techniques for forest investigation is necessary. The evaluation indexes set, which included 10 factors: elevation, slope, aspect, surface curvature, solar radiation index, topographic humidity index, tree ages, the depth of soil layer, the depth of soil A layer, and coarseness, was established. Then, using the integration data of the administrative map, Digital Elevation Model (DEM), and forest resource planning investigation data of the key forestry city of Longquan, Zhejiang Province, PRC, the membership of each factor was empirically fitted by polynomials, and the forest volume was estimated via an improved back propagation (BP) neural network(NN) model with Levenberg-Marquardt(LM) optimization algorithm(LM-BP). The results show that the individual average relative errors (IARE) were from 23.29% to 47.87% with an average value of 33.06%; The groups relative errors (GRE) were from 0.38% to 9.31% with an average value of 3.65%, this meant that groups estimation precision was more than 90% which is the highest standard of overall sampling accuracy about volume of forest resource inventory in china.

Keywords: LM-BP; Forest Volume; Estimation

1 Introduction

Forest inventories provide objective and scientifically reliable information on key forest ecosystem processes, and constitute an effective tool for forest management and forest resource monitoring. Forest inventory data define the extent, size distribution, and species composition of forested and non-forested lands and through periodical updating, they track the changes that occur in natural resources over time [1].

In China, the traditional large-scale survey of forest resources include forest inventory and forest resource planning investigation, where, forest inventory repeated once every 5 years, and the forest resource planning investigation conducted once every 10 years interval. Whether from an ecological viewpoint or from the perspective of the use of forest products, the traditional long cycle of forest resources survey has been unable to meet the actual demand.

Since cost factors are of primary importance, forest managers are continuously searching for more efficient and reliable estimation models that could integrate or, in some cases, substitute the traditional and expensive measuring techniques[1].

Many simulation models have been built to provide managers with predictions of forest growth and yield response to treatments. But, all developed models presented a very low R^2 implying that the variables explained a small portion of the growth processes. Traditional statistical methods are not always suited to solve unstructured problems occurring in natural resource assessment [2] mainly because statistical methods are based on some assumptions on the data distribution. Moreover, they have shown to have several limitations when variables that are involved interact in a complex manner and have difficulties in handling poor and noisy data. Such conditions are very frequent in forest data where classes may display a range of distributions, relationships between variables may be non-linear, and outliers and noise may exist in the data [1].

Thanks to their flexibility and adaptability, artificial neural networks (ANNs) constitute an alternative and valid approach for modelling non-linear and complex long-lived dynamic biological ecosystems such as forests. ANN models have become very popular because they can learn complex patterns and trends in the data, they are slightly affected by data quality problems and bias, and they are robust to data structures with highly interrelated relationships. Artificial neural networks were applied to develop models to predict aboveground forest carbon storage according to sample data and Land sat Thematic Mapper TM data. The results showed the BPNN algorithm could accurately generate the spatial distributions of forest carbon density and changes, where

Corresponding author e-mail: dashengwu@sina.cn the mean estimate of carbon density for the whole study area was 0. 98mg (10. 89 mg/hm 2) which was smaller than the average from the sample plots with a relative error of only 13%[3]. Two BPNNs were constructed, and their performance in estimating the height of pure uneven-aged stands of common beech (Fagus sylvatica L.) in north-western Spain was compared with that of the models most commonly used to estimate tree height (nonlinear calibrated local and generalized mixed-effects models and generalized fixed-effects models). Comparison results showed that BPNNs require less sampling effort because no height measurements are required for their implementation [4].

Although ANNs have been showing potential for solving some difficult problems in forest resources management, research on ANNs applications has been very limited compared to other artificial intelligence techniques.

To facilitate the monitoring of forest resources by the forest managers, a simulation model based on the improved back propagation neural network (BPNN) with Levenberg-Marquardt(LM) algorithm(BP-LM) has been developed. In which, first, a comprehensive evaluated factor set were established to cost-effectively estimate forest volume, including 10 factors: elevation, slope, aspect, surface curvature, solar radiation index, topographic humidity index, tree ages, the depth of soil layer, the depth of soil A layer, and coarseness. Then, using the integration data of the administrative map, DEM, and forest resource planning investigation data of the key forestry city of Longquan, Zhejiang Province, PRC, the membership of each factor was empirically fitted by polynomials, and the forest volume was estimated via an improved BPNN model with LM optimization algorithm.

2 Study area

The key forestry city of Longquan, 3059 km^2 in extent, is a largely mountainous area located in the southwestern part of Zhejiang province in China, where the longitude is from 118°42′to 119°25′E, and latitude is between 27°42′to 28°20′N.

There are abundant forest resources with 3,985,000 mu of areas, forest volume reached 14.56 million cubic meters, and the forest coverage rate up to 84.2%.

3 Materials and methods

3.1 MATERIALS AND METHODS

Formation and development of forest resources is actually a forest cultivation process from tree seeds, seedlings, planting forest trees to mature, throughout the process of cultivating and nurturing, forest development must be carried out under certain site conditions which commonly evaluated by site factors including environmental factors, forestry vegetation factors and human activity factors[5].

Typically, in the natural state, the development of forest resources most affected by the environmental factors which include 3 classes:

- climate, mainly includes solar radiation and precipitation.
- topography, directly related to water potential and soil conditions, including elevation, aspect, slope, slope position, slope type, and small terrain, etc.
- soil, including soil type, soil depth, soil texture, soil structure, soil nutrients, soil humus, soil PH, soil erosion degrees, all levels of gravel reserves in the soil, soil salinity, soil-forming rock, and parent material type, etc.

In fact, site factors are not always suited to solve the monitoring of forest resources because the variety of site factors greatly increased the costs of data acquisition and greatly increased the complexity of research, which led to many experts and scholars try to select part of these factors involved in their experiment and have got some good results [6, 7]. Index system which monitor the forest resources changes should follow the basic principles of scientific, systematic, practicality and economy.

To estimate forest volume, a comprehensive evaluated index set including 10 factors: elevation, slope, aspect, surface curvature, solar radiation index, topographic humidity index, tree ages, the depth of soil layer, the depth of soil A layer, and coarseness, was established.

3.2 DATA SOURCES

There were data sources as following:

(1) Administrative map about Longquan city.

(2) The 2007 forest resource planning investigation data consisting of 83078 subplots, which composed a group of sub compartments with the size of 39377. In order to eliminate erroneous and incorrect data, a preliminary phase of the study comprised of severe data quality control. During preliminary processing, those samples, which were non-volume or located in non-forested land, were removed. Thus, the 2007 forest resource planning investigation data were remained with 28707 sub compartments and 40249 subplots.

(3) DEM with 30 meter resolution, which was generated by the first version data of ASTER GDEM (V1) in 2009, with a data type of IMG and a projection of UTM/WGS84.

3.3 IMPROVED BP NEURAL NETWORK MODEL BASED ON LM ALGORITHM

ANNs have received a great deal of attention over the last 3 decades as a valid alternative to traditional statistical methods to predict the behaviours of non-linear systems. The importance of neural networks is in their ability to learn very complex and correlated patterns.

As previously underlined, multilayer feed-forward neural networks trained by back, propagation algorithm has been the most prominent and well-researched class of ANNs in classification and pattern recognition. A back propagation system usually comprises three types of successive layers: input layer, hidden layer and output layer. During training, the input signal propagates through the network in a forward direction, from left to right on a layer-by-layer basis, generating a set of values on the output units and fixing all networks synaptic weights. Then, difference between the actual and desired output values is measured, and the network model connection strengths are changed so that the outputs produced by the network become closer to the desired outputs. A backward pass achieves this during which connection changes are propagated back through the network starting with the connections to the output layer and ending with those to the input layer [1].

However, the traditional BPNN has some shortcomings, such as slow convergence speed and easy to fall into local minimum, etc.

Fortunately, LM algorithm, which is actually a combination of gradient descent algorithm and newton algorithm, compare to the traditional BPNN, significantly reduce the number of iterations, accelerate the convergence speed, and get a higher accuracy. Especially, whose convergence speed is the fastest of all traditional and other improved BPNNs for medium-sized networks. In recent years, the improved BPNN by LM algorithm has been widely used in the fields of evaluation and forecasting and had some good effects [8-10].

In order to obtain a better result for the experiment, the improved BP neural network model based on LM algorithm was chosen to estimate the volume of forest resources.

3.4 Data preprocessing

3.4 .1 Data integration

The average volume per unit (m3/mu) of forest resources was the only estimated factor, whose data was stored in the database of forest resource planning investigation. The depth of soil layer, the depth of soil A layer, tree ages, coarseness were stored in the same database also.

However, the data about elevation, slope, aspect, surface curvature, solar radiation index, topographic humidity index were derived from DEM.

To take full advantage of the database management systems(DBMS) in the data storage and analysis, all data should be integrated into the same database of forest resource planning investigation.

3.4.2 Membership about evaluation indexes

Generally, membership is solved as follows:

- to group each evaluation index data according to the experience;
- to statistics its distribution area or the average volume per unit of forest resources by the each grouped evaluation index, and to obtain their polynomial fitting curves and fitting formulas;
- to get the fitted values for each evaluation index according the fitting formulas, and to get their membership with normalization through equation as shown in formula 1.

$$z_i = \left| y_i / \max\left(y_i\right) \right| , \qquad (1)$$

where, y_i was the fitted value of each index of every monitoring unit, max(yi) was the maximum of all y_i, z_i was the membership of each index.

Exceptionally, in this article, the indexes of the depth of soil layer, the depth of soil A layer, the coarseness and the aspect, whose membership had its own special rules.

Specifically, the membership of each evaluation index was solved:

(1) Elevation: whose values were between 156 and 1806. Considering the distribution of species are usually within a certain elevation range, which is often hundreds of meters across, in order to speed up training, they first must be classified.

Step 1, to divide elevation values into 50 classes equidistantly.

Step 2, to statistics their distribution area of forest resources by the 50 classes.

Step 3, to obtain their polynomial fitting curves and the fitting formulas, in which the elevation was independent variable and the distribution area of forest resources was dependent variable.

Step 4, to get the membership of elevation according to formula 1.

(2) Slope: whose values were from 1 to 49. Considering the interval of slope values are relatively small, they were firstly rounded to the nearest integer and only classified into 47 classes. Correspondingly, the solution steps of slope membership should be a referring to the elevation.

(3) Aspect: firstly, according to their degree range, to divide aspect into 9 classes: flat, north, northeast, east, southeast, south, southwest, west, northwest, north; secondly, to statistics their distribution area of forest resources grouped by the 9 classes; finally, to get the membership of aspect according to formula 1. The classification and membership about aspect showed as table 1.

(4) Surface curvature: whose values were from -1.55555999279 to 1.46667003632 with an average value of 0.00703782594377091. Since the interval of curvature was too small, they were rounded to the nearest integer after amplification to 100 times of their original value and classified into 218 classes. Correspondingly, the solution steps of curvature membership should be a referring to the elevation.

(5) Solar radiation index: their values were firstly divided by 10000, then rounded to the nearest integer and classified into 95 classes. Other solution steps of

membership of solar radiation index should be a referring to the elevation.

(6) Topographic humidity index: their values were from 9.67430973053 to 34.2994995117. Considering the interval of the values were relatively small, they were rounded to the nearest integer after amplification to 10 times of their original value and classified into 228 classes. Correspondingly, the solution steps of membership of topographic humidity index should be a referring to the elevation.

(7) Tree ages: tree ages did not directly affect forest distribution area, but would affect forest volume, and thus the membership of the index should be based on the relationship between tree ages and the average volume per unit rather than the distribution area of forest resources. In addition, the solution steps of membership of tree ages should be a referring to the elevation.

(8) The depth of soil layer: a positive correlation between soil thickness and plant height has been presented [11]. Similarly, in this research, the experimental data also reflected a generally positive linear correlation between the depth of soil layer and the volume of forest resources. So, the membership of this index was calculated by formula 1 directly.

(9) The depth of soil A layer: according to the data from forest resource planning investigation, the depth of soil A layer qualitatively recorded as thick, medium, thin or null. Accordance with experts' experience, the membership of the depth of soil A layer were quantified as: thick to 1; medium to 0.7; thin to 0.4 and null to 0.

(10) Coarseness: Similarly, with tree ages, the creating membership of this index should be based on the relationship between the coarseness and the average volume per unit of forest resources. In addition, other solution steps of coarseness membership should be a referring to the elevation.

The polynomial fitting curves about those evaluation indexes, which included elevation, slope, surface curvature, solar radiation index, topographic humidity index and Coarseness, were showed as figure 1-figure 7. Correspondingly, their polynomial fitting formulas were showed as table 2.

TABLE 1 Classification and membership of aspect

Aspect classification	Degrees range	Actual distribution area of forest	Membership
Flat	<=0	-	-
North	(>0 and <=22.5) or (>337.5 and <=360)	4501	0.006683138
Northeast	>22.5 and <=67.5	139973	0.20783357
East	>67.5 and <=112.5	360173	0.534789142
Southeast	>112.5 and <=157.5	534362	0.793427035
South	>157.5 and <=202.5	652845	0.969351998
Southwest	>202.5 and <=247.5	673486	1
West	>247.5 and <=292.5	473574	0.703168292
Northwest	>292.5 and <=337.5	172661	0.2563691

TABLE 2 The polynomial fitting formulas for the 7 evaluation indexes

Index name	Polynomial fitting formula
Elevation	$y=-6.78e^{-08}x^{4}+0.00042981\ x^{3}-0.941288539x^{2}+738.4373822x-79871.40674$
Slope	$y=-6.78e^{-15}x^{15}+2.63e^{-12}x^{14}-4.63e^{-10}x^{13}+4.85e^{-08}x^{12}-3.38e^{-06}x^{11}+0.000165019x^{10}-.005798663x^{9}+0.148429674x^{8}-2.770327014x^{7}+37.376799x^{6}-357.7547367x^{5}+2353.107885x^{4}-0098.0831x^{3}+26373.10656x^{2}-34728.49104x+16580.03954$
Surface Curvature	$y=3.72e^{-36}x^{20}-1.91e^{-34}x^{19}-4.81e^{-31}x^{18}+1.69e^{-29}x^{17}+2.59e^{-26}x^{16}-6.58e^{-25}x^{15}-7.70e^{-22}x^{14}+1.52e^{-0}x^{13}+1.40e^{-17}x^{12}-2.31e^{-16}x^{11}-1.63e^{-13}x^{10}+2.40e^{-12}x^{9}+1.25e^{-09}x^{8}-1.68e^{-08}x^{7}-6.32e^{-06}x^{6}+7.37e^{-05}x^{5}+0.02123x^{4}-0.1805x^{3}-45.6071x^{2}+189.1521x+50263.14$
Solar Radiation Index	$y=-6.16e^{-20}x^{15}+1.07e^{-16}x^{14}-8.42e^{-14}x^{13}+3.91e^{-11}x^{12}-1.18e^{-08}x^{11}+2.30e^{-06}x^{10}-0.000260073x^{9}+0.002569461x^{8}+5.022223928x^{7}-1044.693319x^{6}+122333.9409x^{5}-9544261.604x^{4}+508159595.3x^{3}-17869026938x^{2}+3.75995e^{+11}x-3.59689e^{+12}$
Topographic Humidity Index	$y=-5.072435697e^{-26}x^{16}+1.816994302e^{-22}x^{15}-3.022650299e^{-19}x^{14}+3.098679650e^{-16}x^{13}-2.190274111e^{-13}x^{12}+1.131479979e^{-10}x^{11}-4.417347193e^{-8}x^{10}+1.328940533e\text{-}05e^{-05}x^{9}-0.003112417x^{8}+0.569125647x^{7}-80.949868992\ x^{6}+8858.396346998x^{5}-730836.145378965x^{4}+43927720.9766960x^{3}-1813419841.016x^{2}+45921882977.516x-537240051906.498$
Tree Ages	$y=2.34e^{-05}x^{3}-0.003269209x^{2}+0.220835719x+0.84902787$
Coarseness	$y=-0.007507773x^{3}+0.18633539x^{2}-0.027133501x+0.195631142$

3.4.3 Volume normalization

In order to unify the dimension for all variables, the volume of forest resources should also be normalized according to formula 1 before they were input into BPNN. In which, the volume referred to the average volume per unit(m^3/mu).

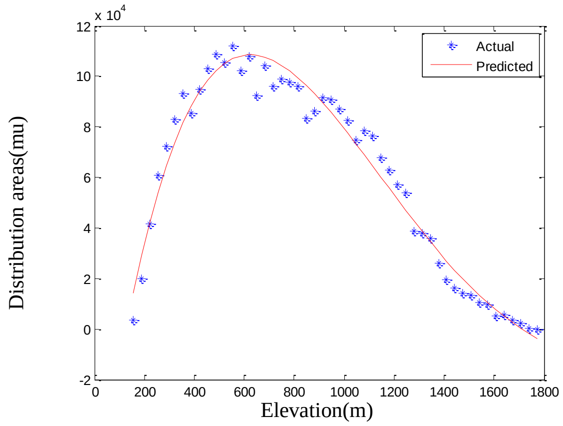


FIGURE 1 Polynomial fitting curves of elevation

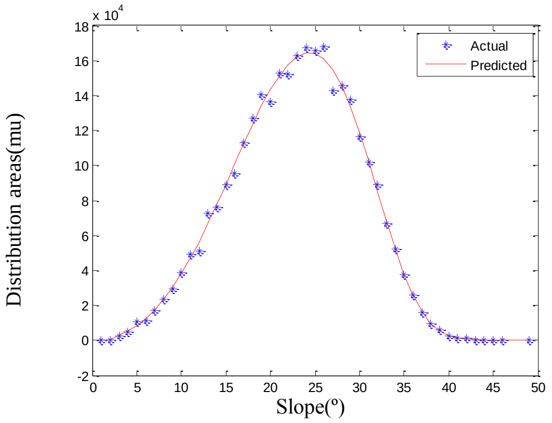


FIGURE 2 Polynomial fitting curves of slope

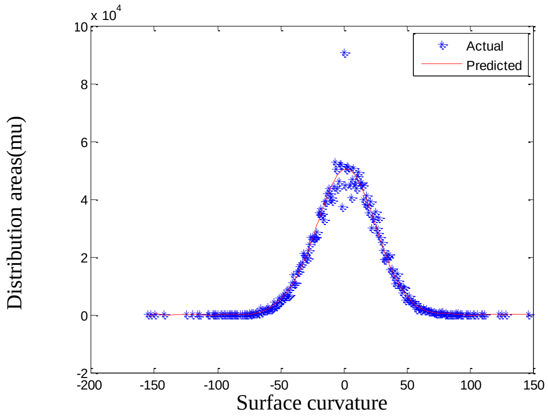


FIGURE 3 Polynomial fitting curves of surface curvature

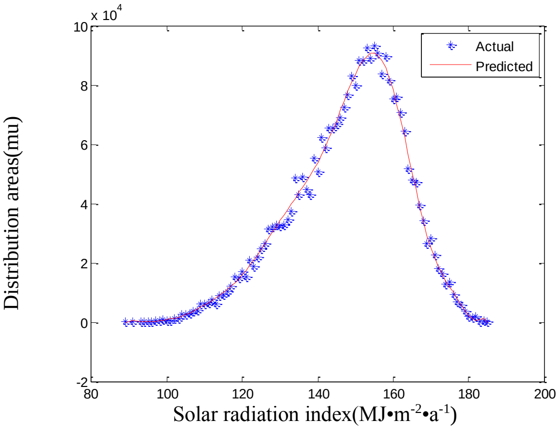


FIGURE 4 Polynomial fitting curves of solar radiation index

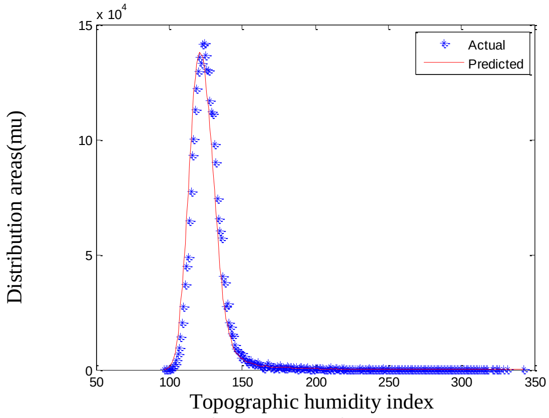


FIGURE 5 Polynomial fitting curves of topographic humidity index

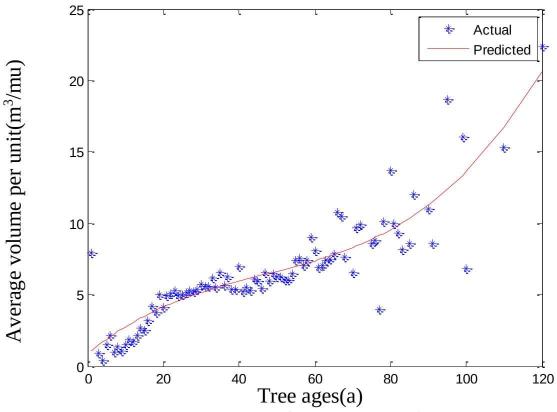


FIGURE 6 Polynomial fitting curves of tree ages

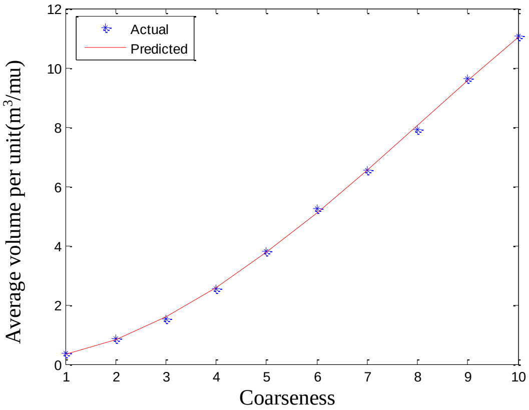


FIGURE 7 Polynomial fitting curves of coarseness

3.5 ESTIMATION FOR FOREST VOLUME BASED ON BP-LM NEURAL NETWORK MODEL

3.5.1 Determining of modelling sample set and simulating sample set

The pre-processed data were divided by administrative unit into 22 groups including 3 streets, 8 towns, 8 townships, a scenic spot and two forest farms. After removing a forest farm(city forest farm) for its too small sample size with only 13, the other 21 groups of samples were independently divided into 2 sets: a modelling sample set and a simulating sample set.

3.5.2 Setting model parameters

The improved BPNN based on LM algorithm comprised three successive layers: input layer, hidden layer and output layer. In which, the nodes of hidden layer were calculated by formula 2.

$$Hidden\_Num = 2 * Input\_Num + Output\_Num\,, \qquad (2)$$

where, Hidden_Num was the number of nodes about the hidden layer, Input_Num was the number of nodes of the input layer, and Output_Num was the number of nodes of the output layer.

Specifically, those model parameters were set as follows:

Epochs = 1000; % the maximum of epochs was 1000.

Input_Num=10; %the nodes of input layer was 10.

Output_Num=1; %the nodes of output layer was 1.

Hidden_Num=2*Input_Num+ Output_Num; %the nodes of the hidden layer

TransferFcn= {'tansig' 'purelin'}; %tansig was the transfer function transferring values from the input layer to the hidden layer, and purelin was the transfer function transferring values from the hidden layer to the output layer.

TrainFcn = 'trainlm'; % training function was trainlm corresponding to LM algorithm.

LearnFcn = 'learngdm'; %learning function was learngdm.

PerformFcn = 'mse'; % performing function was mse(mean square error).

3.5.3 Creating net

Net =newff(P,T, Hidden_Num), where P was the input vector and T was the output vector.

3.5.4 Training net

[Net TR] = train(Net,P,T);%training net

3.5.5 Simulation

y =sim(Net,P_test), where, Net was the trained net, P_test was the input vector of simulating samples and y was the estimation result.

4 Results and discussion

Estimation results of forest volume based on improved BPNN with LM algorithm were showed as table 3.

TABLE 3 Estimation results of forest volume

Administrative unit name	**Total samples**	**Modelling samples**	**Simulating samples**	**Observed value(m^3/mu)**	**Calculated value(m^3/mu)**	***IARE* (%)**	***GRE* (%)**
Longyuan street	1498	1200	298	0.307747	0.317165	25.90	3.06
Jianchi street	421	221	200	0.272585	0.274099	47.87	0.56
Xijie street	634	434	200	0.329917	0.354546	43.98	7.47
Zhulong town	3155	2500	655	0.324390	0.308923	28.77	4.77
Badu town	2457	2000	457	0.368117	0.353936	29.18	3.85
Pingnan town	3264	2500	764	0.144387	0.148159	29.46	2.61
Jinxi town	2111	900	211	0.297376	0.290909	26.75	2.17
Xiaomei town	1358	1000	358	0.280108	0.276572	25.50	1.26
Chatian town	1144	900	244	0.319118	0.323485	28.95	1.37
Shangyang town	1906	1500	406	0.251733	0.240635	24.95	4.41
Anren town	2870	2000	870	0.268594	0.253586	31.74	5.59
Daotai township	5507	4500	1007	0.111731	0.112151	27.27	0.38
Chengbei township	4054	3000	1054	0.161777	0.153023	35.23	5.41
Zhuyang township	1553	1200	353	0.300755	0.313140	42.75	4.12
Tashi township	1165	900	265	0.379292	0.343975	35.00	9.31
Baoxi township	1696	1400	296	0.227732	0.225360	35.11	1.04
Yanzhang township	826	600	226	0.230112	0.238182	46.82	3.51
Lanju township	1240	1000	240	0.229669	0.219144	25.12	4.58
Longnan township	2723	2400	323	0.139618	0.146168	23.29	4.69
Fengyang mountain	550	400	150	0.247506	0.251961	35.43	1.8
Forest farm of shankeng	104	70	34	0.325962	0.310383	45.29	4.78
Average value				0.262773	0.259786	33.06	3.65

Note: *IARE* was the individual average relative error which calculated by formula 3, and *GRE* was the group relative error which calculated by formula 4.

$$IARE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{t_i - y_i}{t_i}\right|, \qquad (3)$$

$$GRE = \frac{1}{n}\left|\frac{\sum_{i=1}^{n}(t_i - y_i)}{\sum_{i=1}^{n}(t_i)}\right|, \tag{4}$$

where, n was the number of simulating samples, t_i was the observed value of the i-th sample, y_i was the calculated value of the i-th sample.

As shown in table 3, *IARE* were from 23.29 to 47.87 with an average value of 33.06, and *GRE* were from 0.38 to 9.31 with an average value of 3.65. There was a considerable reason why those 4 groups *GRE* were more than 5%, that is, most of their tree ages were between 18 years and 30 years that almost all observed points were located above the fitting curve(figure 6), respectively, there were 75% in Tashi township(*GRE* 9.31), 80% in Xijie street(*GRE* 7.47), 54% in Anren town t(*GRE* 5.59), and 55% in Chengbei township(*GRE* 5.41).

5 Conclusions

In this study, the forest volume of the key forestry city, Longquan in Zhejiang province of China, was estimated dynamically. First, the evaluation index set was established, which included 10 factors: elevation, slope, aspect, surface curvature, solar radiation index, topographic humidity index, tree ages, the depth of soil layer, the depth of soil A layer, and coarseness. Then, the membership of each evaluation set was empirically fitted by polynomials, and the forest volume was estimated via an improved BPNN model with LM optimization algorithm. The results showed that the average individual relative errors(IARE) were from 23.29% to 47.87% with an average value of 33.06%; the groups relative errors (GRE) were from 0.38% to 9.31% with an average value of 3.65%, this meant that groups estimation precision was more than 90% which is the highest standard of overall sampling accuracy about volume of forest resource inventory in china.

Acknowledgments

This work was financially supported by Zhejiang provincial major science and technology project (2011C12047), Zhejiang provincial natural science foundation (Y5110145), pre-research development fund project of Zhejiang A & F university (2008FK60).

References

[1] Scrinzi G, Marzullo L, Galvagni D 2007 Development of a neural network model to update forest distribution data for managed alpine stands *Ecological Modelling* **206** 333-46

[2] Gimblett R H, Ball G L 1995 Neural network architectures for monitoring and simulating changes in forest resources management *AI Appl.* **9**(2) 103-23

[3] Wang Shaohua, Zhang Maozhen, Zhao Pingan, Chen Jinxing 2011 Modelling the spatial distribution of forest carbon stocks with artificial neural network based on TM images and forest inventory data *Acta Ecologica Sinica* **31**(4) 998-1008

[4] Castaño-Santamaría J, Crecente-Campo F, Fernández-Martínez J L, Barrio-Anta M, Obeso J R 2013 Tree height prediction approaches for uneven-aged beech forests in north-western Spain *Forest Ecology and Management* **307** 63-73

[5] Shen Guofang 2001 *Silviculture* China Forestry Press: Bei Jing

[6] Xie Huiqin 2004 Establishment on growth model of Chinese fir and application of multilinear regression *Journal of Fujian Forestry Science and Technology* **1** 34-7

[7] Xu Weimin, Chen Youfei1, Lin Guangfa, Chen Minghua 2011 Dynamic visualization of Chinese fir volume driven by site condition and growth model *Journal of Fujian College of Forestry* **31**(2) 151-5

[8] Miao Xinying, Chu Jinkui, Du Xiaowen 2011 Application of LM-BP neural network in predicting dam deformation *Computer Engineering and Applications* **47**(1) 220-2

[9] Jian Xiaochun, Wang Liwei, Min Feng 2012 BP Neural Network Based on LM Algorithm for the Forecasting of Vehicle Emission *Journal of Chongqing University of Technology (Natural Science)* **26**(7) 11-6

[10] Wang Zeping 2013 Application of LM - BP Neural Network in Lake Trophic Evaluation *Environmental Science Survey* **32**(3) 98-101

[11] Li Chengcheng, Cheng Xing, Yang Shichao 2012 Study on The Soil Thickness Factor of Plant Growth in Karst Mountains - Take The Guizhou Xiangbao Mountain as An Example *Journal of Guizhou Normal College* **28**(9) 38-41

Authors

Dasheng Wu, born on November, 1972, Qingyuan County, Zhejiang Province, China

Current position: Lin'an City, Zhejiang Province, China
University studies: Zhejiang A & F University
Scientific interest: remote sensing and information system application
Experience:

- **Education:** September 2008 to present, Zhejiang University, majored in remote sensing and information system application; September 2000 to December 2003, Beijin Forestry University, majored in information system, got master's degree; September 1990 to July 1994, Zhejiang Forestry College, majored in forestry, got bachelor's degree.
- **Work experience:** August 1994 to present, Zhejiang A&F University.

Required screw length measurement in distal tibia based on three-dimensional simulated screw insertion

Kun Zhang, Yanxi Chen*, Minfei Qiang

Department of Orthopaedic Trauma, East Hospital, Tongji University School of Medicine, 150 Jimo Rd, Shanghai 200120, China

Received 1 March, www.tsi.lv

Abstract

The objective of the study was to provide morphological data of the distal tibia to offer guidance on the required screw length. Computed tomography scans of the ankle in 225 patients were reviewed. Then parameters in the three-dimensional reconstruction images were measured by three independent, qualified observers on 2 separate occasions. The anteroposterior length increases from medial to lateral margin at the level of the base of the tibiofibular syndesmosis. On both proximal and distal planes of tibiofibular syndesmosis, the medial-lateral width increases from posterior to anterior margin. Significant differences were observed in all parameters between male and female and in the minimum width at the level of the roof of the syndesmosis between left and right limbs (P<0.05). All of the parameters exhibited moderate to excellent intra-class correlation coefficient. The anteroposterior screws would probably penetrate the far cortex and injure the structures surrounding the distal tibia if longer than 35.35 mm and 32.53 mm in male and female. The screws should not longer than the maximum diagonals which are 51.29 mm and 46.58 mm on distal plane and 43.64 mm and 38.24 mm on proximal plane in male and female respectively, or inadvertent distal tibiofibular syndesmosis penetration may occur.

Keywords: Tibia, Tibiofibular syndesmosis, Tomography, X-ray computed, Imaging, three-dimensional

1 Introduction

Distal tibia fracture is one of the most common lower limb fractures as it is contained in both intra-articular and extra-articular fractures such as metaphyseal and pilon fractures [1-3]. The directions of the screws in distal tibia are flexible and should be determined by different types of plates and the pattern of fracture lines [4-8]. Sometimes length determined by depth gauge is imprecise in distal tibia, especially when the far cortex is fragmentized in tibiofibular syndesmosis. The screws inserted may be too long or too short. Then the screws should be replaced and more fluoroscopy is inevitable. The more frequent the fluoroscopy was used, the more radiation exposure surgeons and patients will receive [9]. Therefore, it is important for surgeons to understand the morphology of the distal tibia in all directions.

As more and more orthopaedic surgeons tried to treat distal tibia fracture with minimally invasive plate osteosynthesis (MIPO) [4-8], complications associated with this technique have been described and included infection, damage to neurovascular structures and bone and tendon impingement [10, 11]. Previous studies demonstrated that key anatomical structures including neurovascular bundles and tendons are in very close proximity to the distal tibia [11, 12]. Furthermore, only a small gap exists in the distal tibiofibular syndesmosis [5, 13, 14]. Frequent use of depth gauge, percutaneous insertion and replacement of screws, and poor implantation may lead to injury to nerves, vessels, tendons and articular surfaces. Exposed screw tail may stimulate and injure adjacent anatomic structures. Therefore, there is little room for error treating distal tibia fractures and extreme caution should be exercised. Knowledge of the distal tibia morphology in all directions may help to decrease the possibility of key anatomical structure damage and inadvertent tibiofibular syndesmosis penetration. However, scarce data are available concerning the morphology of the distal tibia.

The objective of the study was to provide morphological data of the distal tibia to offer guidance on the required screw length in all directions.

2 Materials and Methods

2.1 SUBJECTS

Institutional ethical approval for this study was obtained from the Ethics Committee of our hospital, and conforms to the provisions of the Declaration of Helsinki. (East Hospital Ethics Committee, Ethics number 2012-020). The patients were collected from the foot and ankle clinic of our hospital and informed consent was obtained. Patients were excluded if they had a history of distal tibia fracture, pilon fracture, internal or posterior malleolus fracture confirmed by radiological examination or surgery. Patients with congenital or acquired malformation, osteoarthritis, rheumatoid arthritis or a history of bone tumour of ankle were also excluded.

**Corresponding author* e-mail: cyxtongji@126.com

Patients less than 20 years old or older than 65 years were excluded to avoid skeletal immaturity or degeneration. Finally, 225 patients were enrolled in this study, with 122 men (with 122 ankle joints) and 103 women (with 103 ankle joints). There were 104 left and 121 right ankle joints. The average age was 39.5 years (range, 20 to 65 years).

2.2 IMAGE ACQUISITION AND POST PROCESSING

The thin-slice CT images (DICOM 3.0 format) of all the patients scanned by 16-row spiral CT (Light Speed, GE, USA) were collected. Main CT scanning parameters were as follows: thickness, 0.625 mm; voltage, 120 kV; current, 200 mA; image matrix, 512×512.

The thin-slice CT axial images of all research subjects were firstly uploaded to picture archiving and communication system (PACS) of hospital, then these CT data were imported into the digital orthopaedic clinical research platform (SuperImage orthopaedics edition 1.0, Cybermed Ltd, Shanghai, China) via removable storage devices. On this platform, three-dimensional (3-D) images of ankle joints were generated by performing surface shaded display with a bone algorithm at 0.625 mm slice thickness. All component bones of ankle joints were distinguished by computer in 3-D images. Then parameters in the 3-D reconstruction images could be measured and calculated.

2.3 MEASUREMENTS AND CALCULATIONS

Design of the parameters in distal tibia closely combined clinical experiences. The trajectories which were most commonly used in distal tibia were visualized by inserting simulated screws (Figure 1, 2). In order to standardize the measurement process so that other researchers could replicate the study, we designed the steps below.

Firstly, on the distal tibia articular surface, the turning point of posterior malleolus and medial malleolus (point A), the turning point of medial malleolus and anterior ankle (point B) and the peak of the lateral margin of the tibial plafond (point C) were selected to define the cross-section (plane ABC) which was corresponded to the most distal slice at the level of the plafond as well as the base of the distal tibiofibular syndesmosis (Figure 3).

Secondly, on the cross-section ABC, points were randomly selected in the antero-medial, antero-lateral, postero-medial and postero-lateral arcs (point E, F, G and H, respectively). The midpoint of the medial margin (point D) and the point in the most front of the anterior margin (point I) were also selected. The distances between point G and E, between point A and B and between point H and I were measured and analysed to obtain the anteroposterior lengths. The distances between point D, E, G and point C, F, H were measured respectively (line DC, DF, DH, EC, EF, EH, GC, GF and GH) and analysed to obtain the medial-lateral widths and diagonal distances (Figure 3).

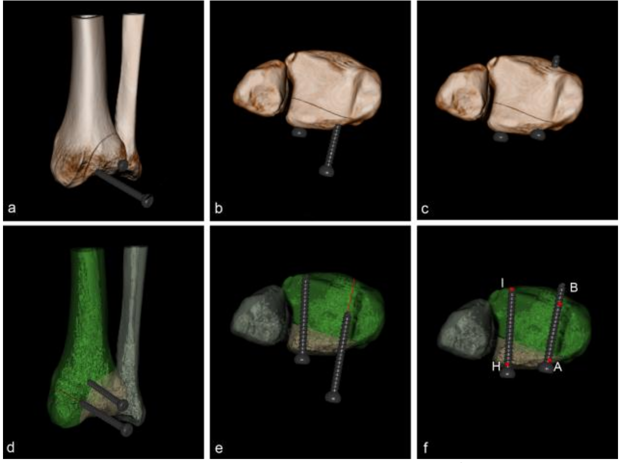


FIGURE 1 An example illustrates the technique of simulating screw insertion and clinical application of the anteroposterior length. (a) A posterior malleolar fracture was simulated in 3-D volume rendering mode. After reduction of the fracture, two screws were simulated to fix the fragment. A screw was inserted and the other to be inserted. (b) A screw was being inserted into the trajectory. (c) After insertion, the figure revealed that the left screw was appropriate and the other was too long. (d-f) The same simulation in surface shaded display mode, which corresponds to Figure 2a-c. (e) The perspective figures revealed the inner part of the tibia. The red arrow was the trajectory. (f) The distance between point A and B and between H and I presented the proper length of the screw. The right screw was too long and penetrated the far cortex.

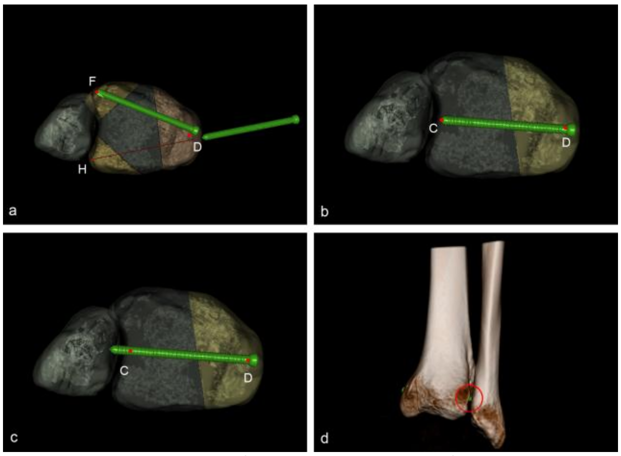


FIGURE 2 An example of clinical application of the medial-lateral width of the distal tibia. The distance between point D and H, between D and F (a) and between D and C (b) presented the proper length of the screw. (c) The figure revealed that the screw was too long and the far cortex was penetrated. (d) The tibiofibular syndesmosis penetration of the screw (marked with red circle) could be observed clearly in volume rendering mode.

Next, plane J was determined through the most proximal point of the anterior tubercle of the distal tibia as the roof of the distal tibiofibular syndesmosis, which is parallel to the plane ABC (Figure 4).

Finally, on this profile, points were randomly selected in the antero-medial, antero-lateral, postero-medial and postero-lateral arcs (point M, N, O and P). The midpoints of medial and lateral margins (point K and L) were also selected. The distances between point K, M, O and point L, N, P were measured respectively (line KL, KN, KP, ML, MN, MP, OL, ON and OP) and analysed to obtain the medial-lateral widths and diagonal distances (Figure 4).

The measurements were performed on 2 separate occasions by 3 independent, qualified observers. All observers were blinded to the other's analysis. The average was taken as the final data.

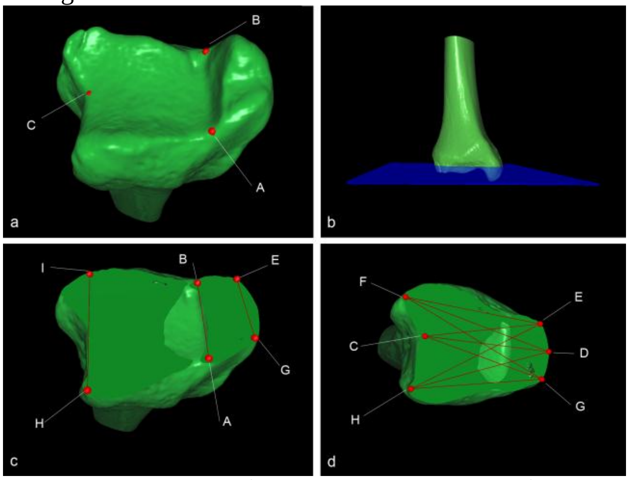


FIGURE 3 Measurement of the distal tibia. (a, b) Base of the distal tibiofibular syndesmosis (plane ABC) was defined by selecting three anatomical landmarks (point A, B and C). (c) Three lines (line GE, AB and HI) were measured and analysed to obtain the anteroposterior lengths. (d) Measuring 9 lines (line DC, DF, DH, EC, EF, EH, GC, GF and GH) to obtain the medial-lateral widths and the maximum distance.

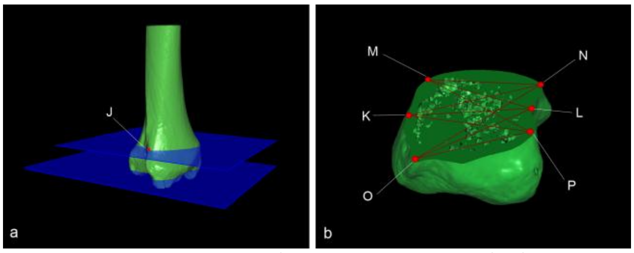


FIGURE 4 The dimension of the tibia on the roof of the distal tibiofibular syndesmosis. (a) Cross-section J was determined through the most proximal point (point J) of the anterior tubercle of the distal tibia. (b) Obtain the medial-lateral width and the maximum distance by measuring 9 lines (line KL, KN, KP, ML, MN, MP, OL, ON and OP).

2.4 STATISTICAL ANALYSIS

SPSS 18.0 (SPSS Inc, Chicago, IL, USA) was used for statistical analysis. The parameters between male and female, and between left and right limbs were compared using the two-samples t test. The lengths and widths were analyzed using the one-way ANOVA. The intra-class correlation coefficient (ICC) was used to assess intraobserver and interobserver reliability. $P<0.05$ was considered to be statistically significant.

3 Results

The mean dimensions of the distal tibia were 41.37 ± 3.22 mm (line DC, range 35.32-49.47 mm, 95% confidence interval [CI]: 40.67-42.11) medial-lateral and 23.83 ± 2.12 mm (line AB, range 20.10-30.47 mm, 95% confidence interval [CI]: 23.45-24.23) anteroposterior at the level of the base of the distal tibiofibular syndesmosis and 35.49 ± 3.29 mm (line KL, range 27.72-45.20 mm, 95% confidence interval [CI]: 34.75-36.24) medial-lateral at the level of the roof of the syndesmosis.

At the level of the base of the syndesmosis (plane ABC), the anteroposterior length increases from medial to lateral margin. The minimum anteroposterior length of distal tibia was 19.14 ± 1.79 mm (line GE, range 15.28-23.60 mm, 95% confidence interval [CI]: 18.82-19.50) and the maximum was 34.09 ± 2.47 mm (line HI, range 27.57-38.88 mm, 95% confidence interval [CI]: 33.63-34.57). The medial-lateral width increases from posterior to anterior margin. The minimum and maximum widths were 38.29 ± 3.39 mm (line GH, range 31.27-46.25 mm, 95% confidence interval [CI]: 37.59-39.04) and 45.36 ± 3.66 mm (line EF, range 38.28-54.43 mm, 95% confidence interval [CI]: 44.56-46.18), respectively.

At the level of the roof of the syndesmosis (plane J), the medial-lateral width increases from posterior to anterior margin. The minimum width was 30.28 ± 3.67 mm (line OP, range 22.98-38.78 mm, 95% confidence interval [CI]: 29.50-31.12), and the maximum was 35.94 ± 3.09 mm (line MN, range 30.20-41.85 mm, 95% confidence interval [CI]: 35.27-36.66).

The longest distance on both planes were diagonals from postero-medial to antero-lateral arc which were 49.01 ± 3.57 mm (line GF, range 41.43-57.13 mm, 95% confidence interval [CI]: 48.19-49.82) on plane ABC and 41.04 ± 4.25 mm (line ON, range 30.18-49.08 mm, 95% confidence interval [CI]: 40.06-42.07) on plane J.

Significant differences were observed in all parameters between male and female (Figure 5), and in the minimum width on the cross-section J between left and right limbs (P<0.05) (Figure 6).

Table 1 shows the reliability of all parameters and the intra-observer reliability between three observers.

4 Discussion

Because of the difficulty in defining coronal, sagittal and axial planes in a 3-D image, a novel method was presented in this study to explain the relativity among different structures in a stereo space. A straight line and a plane can be defined by two and three points respectively in a stereo space. Some specific mark points are recognizable in distal tibia. So this method was also used in this study. Surface shaded display was applied. It was the first 3-D rendering technique applied to medical imaging and was mainly applied in orthopaedics because of its superiority for bony surface reconstructions [15-17]. Besides, the distinct surfaces can facilitate clinical measurements [18]. In addition, it is rather difficult to collect a large sample of cadaver specimens. The morphological measurement on 3-D CT post processing images can be based on a large sample size and thus provide a more reliable and accurate data set.

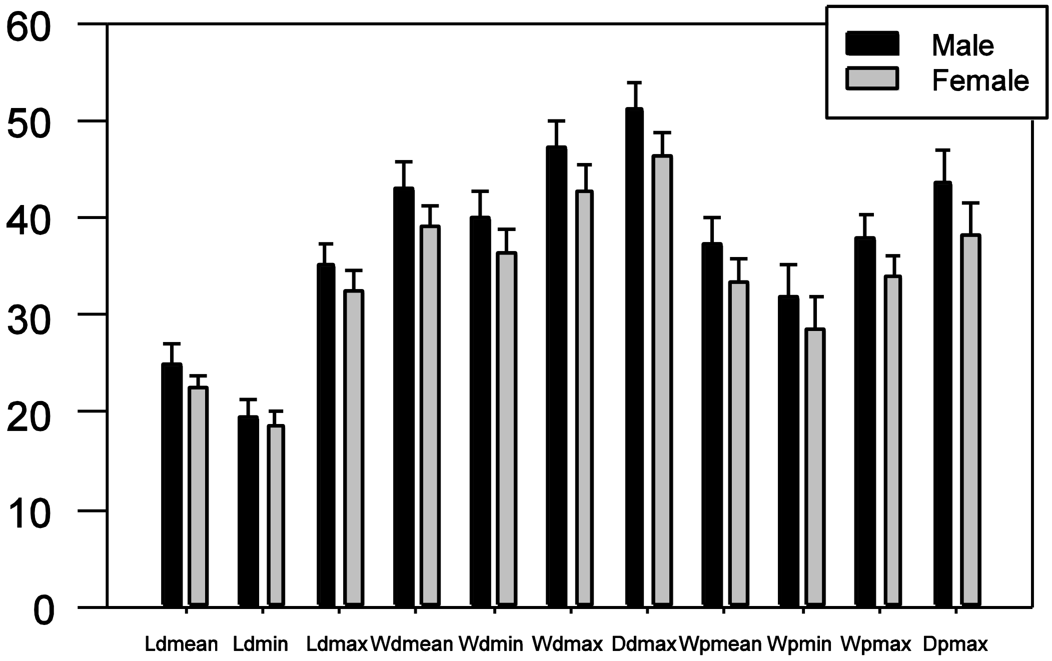


FIGURE 5 Significant differences were observed in all parameters between different genders (P<0.05)

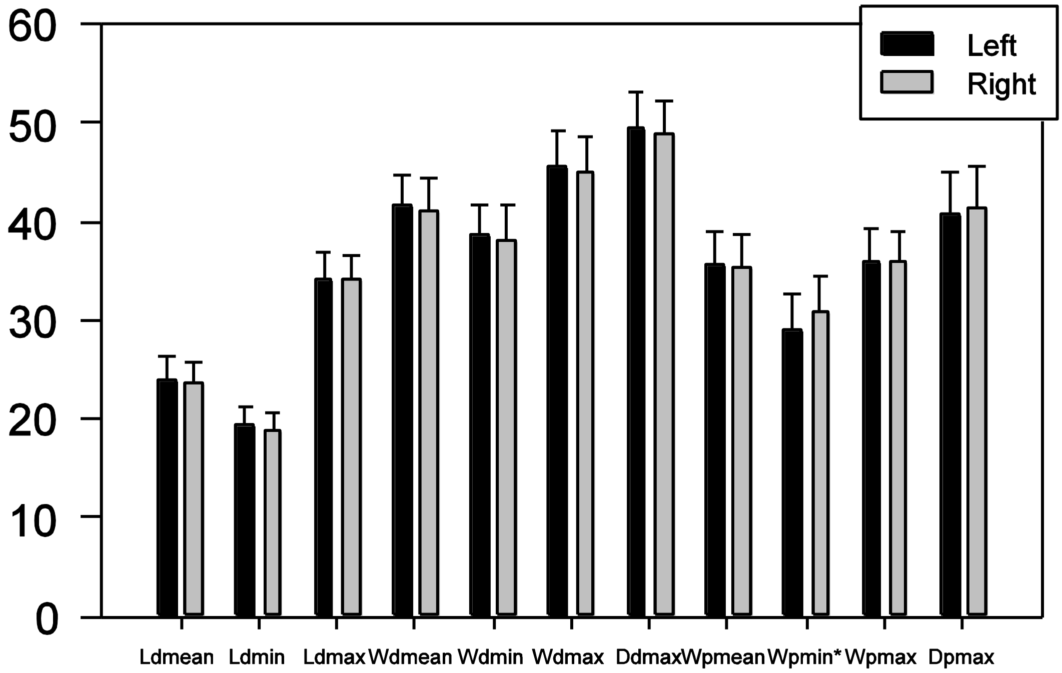


FIGURE 6 *Significant differences were observed in the minimum width on the level of the roof of the distal tibiofibular syndesmosis between left and right tibia (*P*<0.05)

The intra-observer and inter-observer reliability in this study are relatively high. It demonstrates that the values produced by three observers using 3-D CT post processing techniques are accurate and reproducible.

The dimensions of distal tibia offer guidance on the required screw length. As we all know, the length of implanted screws in the plane of distal tibia should be sufficient but no more than the maximum. Or exposed screw tail may stimulate and injure adjacent anatomic structures. If screws inserted were too long or too short, replacement and more fluoroscopy are inevitable. Frequent use of depth gauge, percutaneous insertion and replacement of screws, and poor implantation may lead to injury to nerves, vessels, tendons and articular surfaces. Many important anatomical structures are very close to distal tibia cortex. The anterior tibial compartment

TABLE 1 Intra-observer and inter-observer reliability of all parameters

	Observer 1	Observer 2	Observer 3	ICC1	ICC2	ICC3	ICC12	ICC13	ICC23
Ldmean (mm)	23.77 ± 2.19	23.96 ± 2.12	23.76 ± 2.11	0.98	0.98	0.97	0.97	0.98	0.98
Ldmin (mm)	19.08 ± 1.80	19.20 ± 1.85	19.15 ± 1.78	0.90	0.93	0.91	0.95	0.97	0.96
Ldmax (mm)	34.05 ± 2.61	34.26 ± 2.49	33.96 ± 2.56	0.91	0.92	0.92	0.91	0.91	0.90
Wdmean (mm)	41.72 ± 3.29	41.22 ± 3.59	41.26 ± 2.98	0.96	0.72	0.95	0.83	0.81	0.84
Wdmin (mm)	38.67 ± 3.38	38.18 ± 3.43	38.29 ± 3.30	0.92	0.95	0.95	0.90	0.82	0.93
Wdmax (mm)	45.73 ± 3.58	45.25 ± 3.73	44.90 ± 3.67	0.92	0.95	0.87	0.91	0.89	0.90
Ddmax (mm)	49.56 ± 3.61	48.98 ± 3.70	49.05 ± 3.34	0.90	0.93	0.95	0.93	0.87	0.91
Wpmean (mm)	35.87 ± 3.43	35.55 ± 3.42	35.07 ± 3.21	0.95	0.97	0.94	0.96	0.93	0.95
Wpmin (mm)	30.46 ± 4.13	30.04 ± 3.75	30.34 ± 3.44	0.93	0.95	0.96	0.92	0.89	0.94
Wpmax (mm)	35.71 ± 2.95	36.19 ± 3.36	35.94 ± 3.27	0.77	0.94	0.94	0.88	0.87	0.95
Dpmax (mm)	41.29 ± 4.63	40.99 ± 4.38	40.85 ± 4.14	0.72	0.97	0.97	0.89	0.87	0.96

Ldmean: mean anteroposterior length on the distal plane (base of the distal tibiofibular syndesmosis), Ldmin: minimum length on the distal plane, Ldmax: maximum length on the distal plane, Wdmean: mean medial-lateral width on the distal plane, Wdmin: minimum width on the distal plane, Wdmax: maximum width on the distal plane, Ddmax: maximum diagonal on the distal plane, Wpmean: mean width on the proximal plane (roof of the distal tibiofibular syndesmosis), Wpmin: minimum width on the proximal plane, Wpmax: maximum width on the proximal plane, Dpmax: maximum diagonal on the proximal plane

ICC1, ICC2, ICC3: intraclass correlation coefficient for the first, second and third observer, ICC12, ICC13, ICC23: interclass correlation coefficient between the first and second, between the first and third and between the second and third observer, respectively

Reliability is excellent if ICC is greater than or equal to 0.75, moderate if between 0.4and 0.74, poor if less than or equal to 0.4

contains tibialis anterior tendon, deep peroneal nerve, anterior tibial neurovascular bundle, the posterior includes tibialis posterior tendon, posterior tibial neurovascular bundle, and the lateral has tibiofibular syndesmosis [19]. Ali et al. [12] reported that the anterior neurovascular bundle was 3 mm from the anterior tibial cortex and the posterior cortex was 1 mm from the tibialis posterior tendon and 3 mm to the posterior neurovascular bundle. Iatrogenic injury may occur during surgical manoeuvres for insertion of the screws including incision, blunt dissection down to the plate, drilling screw holes and finally the insertion itself [11]. Deangelis et al. [20] found the superficial peroneal nerve to be at significant risk during percutaneous screw placement in distal tibia. Bono et al. [10] found a relatively high incidence of about 63% of at least one structure damage after anteroposterior locking bolts inserted into the distal metaphyseal tibia. According to our results, from medial to lateral margin, the anteroposterior length increase from 19.65 mm and 18.52 mm to 35.35 mm and 32.53 mm in male and female at the level of the base of the tibiofibular syndesmosis respectively. The screws would probably penetrate the far cortex and injure the structures surrounding the distal tibia if longer than 35.35 mm and 32.53 mm in male and female (Figures 1 and 2). The results of this study present a possibility for surgeons to estimate the length of screws before operation.

Special attention needs to be paid to avoid any disruption of the ankle syndesmosis, which is a joint between the distal tibia and fibula [21, 22]. The distal fibula can be rotated externally and translated postero-medially by an external rotation force [23]. And with ankle dorsiflexion, the distal fibula moves proximally, posteriorly and rotates externally [22]. There is a small area where the tibia and fibula are in direct contact at the base of the syndesmosis [14, 24]. Moreover, the width of the distal tibiofibular syndesmosis 9-12 mm proximal to the tibial plafond was 2-4 mm [13]. Therefore, only a small gap exists in the distal tibiofibular syndesmosis [5, 13, 14]. On both proximal and distal planes of tibiofibular syndesmosis, the medial-lateral width increases from posterior to anterior margin. On distal plane, from posterior to anterior margin, the mean medial-lateral widths increase from 40.03 mm and 36.33 mm to 47.29 mm and 44.22 mm in male and female respectively. Similarly, on distal plane, the widths increase from 31.80 mm and 28.64 mm to 37.82 mm and 33.92 mm. The screws should not longer than the maximum diagonals which are 51.29 mm and 46.58 mm on distal plane and 43.64 mm and 38.24 mm on proximal plane in male and female respectively, or inadvertent distal tibiofibular syndesmosis penetration may occur (Figure 2).

However, the extent of distal tibiofibular syndesmosis is not clearly defined [14]. Kelikian et al. [21] postulated that the distal tibiofibular syndesmosis begins at the level of origin of the tibiofibular ligaments from the tibia and ends where these ligaments insert into the fibular malleolus. The anterior tibiofibular ligament originates close above the anterior tubercle of the distal tibia and the posterior ligament extends from the posterior tibial malleolus [14, 24]. The anterior tubercle is larger than the posterior tubercle [14]. In addition, it is sometimes difficult to distinguish between the proximal margin of the posterior tibiofibular ligament and the interosseous tibiofibular [24]. Therefore, the proximal margin of the anterior tubercle of the tibia was defined as the roof of the tibiofibular syndesmosis.

We acknowledge there are some limitations of the study. The cross-sections in this study were defined by several points. Therefore, the measured values may be inconsistent with those measured by two-dimensional MRI scans and CT scans. Moreover, there is no clear statement about the exact extent of the tibiofibular syndesmosis [14]. In this study, the plane through the most proximal point of the anterior tubercle was defined as the most proximal extent of the joint, which worthy of continuous further study and improvement. Further clinical results are needed to test the findings of the study.

5 Conclusion

The study provides 3-D CT post processing techniques-based detailed values of distal tibia. The dimensions of distal tibia offer guidance on the required screw length. The anteroposterior length increases from medial to lateral margin at the level of the base of the tibiofibular syndesmosis. On both proximal and distal planes of tibiofibular syndesmosis, the medial-lateral width increases from posterior to anterior margin. The anteroposterior screws would probably penetrate the far cortex and injure the structures surrounding the distal tibia if longer than 35.35 mm and 32.53 mm in male and female. The screws should not longer than the maximum diagonals which are 51.29 mm and 46.58 mm on distal plane and 43.64 mm and 38.24 mm on proximal plane in male and female respectively, or inadvertent distal tibiofibular syndesmosis penetration may occur.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (81000818, 81271989 and 81370053)

References

[1] Mandracchia V J, Evans R D, Nelson S C, Smith K M 1999 Pilon fractures of the distal tibia *Clin Podiatr Med Surg* **16**(4) 743-67

[2] Ristiniemi J 2007 External fixation of tibial pilon fractures and fracture healing *Acta Orthop Suppl* **78**(326) 3, 5-34

[3] Singer B R, McLauchlan G J, Robinson C M, Christie J 1998 Epidemiology of fractures in 15,000 adults: the influence of age and gender *J Bone Joint Surg Br* **80**(2) 243-8

[4] Hasenboehler E, Rikli D, Babst R 2007 Locking compression plate with minimally invasive plate osteosynthesis in diaphyseal and distal tibial fracture: a retrospective study of 32 patients *Injury* **38**(3) 365-70

[5] Khoury A, Liebergall M, London E, Mosheiff R 2002 Percutaneous plating of distal tibial fractures *Foot Ankle Int* **23**(9) 818-24

[6] Redfern D J, Syed S U, Davies S J 2004 Fractures of the distal tibia: minimally invasive plate osteosynthesis *Injury* **35**(6) 615-20

[7] Ronga M, Longo U G, Maffulli N 2010 Minimally invasive locked plating of distal tibia fractures is safe and effective *Clin Orthop Relat Res* **468**(4) 975-82

[8] Sitnik A A, Beletsky A V 2013 Minimally Invasive Percutaneous Plate Fixation of Tibia Fractures: Results in 80 Patients *Clin Orthop Relat Res* **471**(9) 2783-9

[9] Mehlman C T, DiPasquale T G 1997 Radiation exposure to the orthopaedic surgical team during fluoroscopy: "how far away is far enough?" *J Orthop Trauma* **11**(6) 392-8

[10] Bono C M, Sirkin M, Sabatino C T, Reilly M C, Tarkin I, Behrens F F 2003 Neurovascular and tendinous damage with placement of anteroposterior distal locking bolts in the tibia *J Orthop Trauma* **17**(10) 677-82

[11] Pichler W, Grechenig W, Tesch N P, Weinberg A M, Heidari N, Clement H 2009 The risk of iatrogenic injury to the deep peroneal nerve in minimally invasive osteosynthesis of the tibia with the less invasive stabilisation system: a cadaver study *J Bone Joint Surg Br* **91**(3) 385-7

[12] Ali A A, Gregory J J, Ockenden M, Hill S O, Makwana N K 2012 Anatomic description of the distal tibia: implications for internal fixation *J Foot Ankle Surg* **51**(3) 296-8

[13] Elgafy H, Semaan H B, Blessinger B, Wassef A, Ebraheim N A 2010 Computed tomography of normal distal tibiofibular syndesmosis *Skeletal Radiol* **39**(6) 559-64

[14] Hermans J J, Beumer A, de Jong T A, Kleinrensink G J 2010 Anatomy of the distal tibiofibular syndesmosis in adults: a pictorial essay with a multimodality approach *J Anat* **217**(6) 633-45

[15] Chen Y X, Lu X L, Bi G, Yu X, Hao Y L, Zhang K, Zou L L, Mei J, Yu G R 2011 Three-dimensional morphological characteristics measurement of ankle joint based on computed tomography image post-processing *Chin Med J Engl* **124**(23) 3912-8

[16] Zhang K, Chen Y X, Qiang M F, Li H B, Dai H 2014 The morphology of medial malleolus and its clinical relevance *Pak J Med Sci* **30**(2) 348-51

[17] Chen Y X, Zhang K, Hao Y N, Hu Y C 2012 Research status and application prospects of digital technology in orthopaedics *Orthop Surg* **4**(3) 131-8

[18] Kuszyk B S, Heath D G, Bliss D F, Fishman E K 1996 3-D CT: advantages of volume rendering over surface rendering *Skeletal Radiol Skeletal* **25**(3) 207-14

[19] Davidovitch R I, Egol K A 2010 *Rockwood & Green's fractures in adults* (7th edition) ed Bucholz RW, Heckman JD, Court-Brown CM and Tornetta P Lippincott Williams & Wilkins,Philadelphia PA, United States pp 1993-5

[20] Deangelis J P, Deangelis N A, Anderson R 2004 Anatomy of the superficial peroneal nerve in relation to fixation of tibia fractures with the less invasive stabilization system *J Orthop Trauma* **18**(8) 536-9

[21] Kelikian H and Kelikian A S 1985 *Disorders of the ankle* Saunders: Philadelphia

[22] Pena F A, Coetzee J C 2006 Ankle syndesmosis injuries *Foot Ankle Clin* **11**(1) 35-50, viii

[23] Beumer A, Valstar E R, Garling E H, Niesing R, Ranstam J, Löfvenberg R, Swierstra B A 2003 Kinematics of the distal tibiofibular syndesmosis: radiostereometry in 11 normal ankles *Acta Orthop* **74**(3) 337-43

[24] Bartonicek J 2003 Anatomy of the tibiofibular syndesmosis and its clinical relevance *Surg Radiol Anat* **25**(5-6) 379-86

Authors	
	Kun Zhang, born on November 3, 1988, Jiangsu, China **Current position, grades:** Resident Doctor **University studies:** Tongji Unversity, Shanghai, China **Scientific interest:** Traumatic orthopaedics, foot and ankle surgery, computer-assisted orthopaedic surgery **Publications:** 2 papers were indexed by SCI **Experience:** Graduated from Tongji Unversity School of Medicine, has 3-year clinical experience in the field of traumatic orthopaedics, foot and ankle surgery and computer-assisted orthopaedic surgery.
	Yanxi Chen, born on November 20, 1975, Chongqing, China **Current position, grades:** Associate Chief Doctor **University studies:** Tongji Unversity, Shanghai, China **Scientific interest:** Traumatic orthopaedics, foot and ankle surgery, computer-assisted orthopaedic surgery **Publications:** 7 papers were indexed by SCI **Experience:** Associate professor, MD, ph.D, has a high reputation in the field of traumatic orthopaedics, foot and ankle surgery and computer-assisted orthopaedic surgery in China; Member of the Société Internationale de Chirurgie Orthopédique et de Traumatologie (SICOT).
	Minfei Qiang, born on September 15, 1989, Shanghai, China **Current position, grades:** Resident Doctor **University studies:** Tongji Unversity, Shanghai, China **Scientific interest:** Traumatic orthopaedics, foot and ankle surgery, computer-assisted orthopaedic surgery **Publications:** 2 papers were indexed by SCI **Experience:** Graduated from Tongji Unversity School of Medicine, has 3-year clinical experience in the field of traumatic orthopaedics, foot and ankle surgery and computer-assisted orthopaedic surgery.

Disruption management for resource-constrained project scheduling based on differential evolution algorithm

Weiming Chen*, Xiaoyang Ni, Hailin Guo

Faculty of Engineering, China University of Geosciences, Wuhan, 430074, China

Received 1 March 2014, www.tsi.lv

Abstract

In this paper, we study the problem of how to react when an ongoing project is disrupted. The focus is on the resource-constrained project scheduling problem with finish–start precedence constraints and the recovery strategies based on disruption management for the different types of disruptions are proposed. The goal is to get back on track as soon as possible at minimum cost, where cost is now a function of the deviation from the original schedule. The problem is solved with a differential evolution (DE) algorithm that can be solved more perfectly on the objective function. The new model is significantly different from the original one due to the fact that a different set of feasibility conditions and performance requirements must be considered during the recovery process. Project scheduling problem library (PSPLIB) has been taken into account so as to test the effect of novel hybrid method. Simulation results and comparisons determine the effects of different factors related to the recovery process and show that the differential evolution algorithm is competitive and stable in performance.

Keywords: disruption management, scheduling, resource-constrained, differential evolution

1 Introduction

Project scheduling has attracted an ever-growing attention in recent years both from science and practice. Nowadays, enterprises have to focus more on improving product development efficiency due to the global economic crisis and the increasingly intense market competition. To gain more market share, it is critical for enterprises to reduce product development time and cost with limited resources and shorten the time-to-market. One of decision problems that in practice often involve uncertain information is resource-constrained project scheduling problem (RCPSP).

Projects are often performed under high levels of uncertainty related to such factors as resource availability, unproven technology, team competence, and the commitment of upper management. Sometimes, even the project goal is not well defined when the work begins. For most projects, though, a schedule specifying the implementation details must be developed before uncertainties are resolved. Without any historical data or past experience, expert opinion and rough estimates might be the only way to quantify activity costs and durations in the initial planning stages. What results is an initial schedule designed to optimize some objective within the limits of uncertainty. Eden et al [1] pointed out that it is very hard to estimate the cost of delay and disruption for most real world projects. The primary purpose of this paper is to apply the growing field of disruption management (DM) [2, 3] for examining and resolving this type of problem. Disruption management is an emerging field in which operations research techniques are applied to help resolve uncertainties as they unfold. The problem that must be solved may be significantly different from the initial planning problem because it contains new decision variables, new constraints, and a new objective. Beside air traffic and airline-related scheduling, the domains of machine scheduling and production planning have been at the centre of research in disruption management [4]. Bean et al. [5] were among the first to consider deviation costs in their approach to match up scheduling, which is based on the idea of identifying an updated schedule that converges with the original one at some early point in the future. While Clausen et al. [1] discuss disruption management in the execution of shipbuilding processes; Xia et al. [6] investigate DM in the context of a two-stage production and inventory system, evaluating solutions for fixed and flexible setup epochs as well as different forms of penalty functions. Yang et al. [7] consider cost and demand disruptions occurring on a single-product manufacturing plant and propose a pseudo-polynomial dynamic programming procedure for the general cost case and present advanced solution procedures for specific forms of cost functions. Additional information and comprehensive overviews of DM in the context of production planning can be found in [8] and [3]. Apart from the areas of application mentioned above, disruption management plays a crucial role in the context of many other real-world processes. Research, for example, has been conducted in the domains of telecommunication [2], project management [9, 3, 10], supply chain coordination [11, 12] and logistics management [3]. Al-Fawzan et al

**Corresponding author* e-mail: chenpool@126.com

[13] introduce the concept of schedule robustness and we develop a bi-objective resource-constrained project scheduling model and several variants of the algorithm are tested and compared on a large set of benchmark problems. Hur et al [14] define such a setting as the real-time work schedule adjustment decision, proposes mathematical formulations of the real-time adjustment and develops efficient heuristic approaches for this decision. Herroelen et al [15] review the fundamental approaches for scheduling under uncertainty: reactive scheduling, stochastic project scheduling, fuzzy project scheduling, robust (proactive) scheduling and sensitivity analysis and discuss the potentials of these approaches for scheduling under uncertainty projects with deterministic network evolution structure. Howick [16] et al use System dynamics (SD) to analyse disruption and delay, making an informed decision about the appropriateness of SD as a modelling approach to support any specific claim for compensation. Vonder et al [17] discuss computational results obtained with priority-rule based schedule generation schemes, a sampling approach and a weighted-earliness tardiness heuristic on a set of randomly generated project instances. Demeulemeester et al [18, 19, 20] review the fundamental approaches for scheduling under uncertainty and discuss the potentials of various approaches for scheduling under uncertainty projects with deterministic network evolution structure.

Despite of noticeable progress in the researches on DM (disruption management), DM applications to date to solving practical difficulties are still in great constraints, especially in the field of project scheduling, where fruits of multi-mode DM study are yet to be added to present few. With hope of advancing further, this section introduces a framework for solving the problem of multi-mode project scheduling with uncertainties, including task-disruption and resource-disruption. In section 2, we give an outline of DM studies on project scheduling with uncertainties, followed by section 3, adjustment strategies of DM and its mathematical model. In the section 4, differential evolution algorithm is introduced to tackle project scheduling under uncertainties. Differential evolution algorithm is employed to calculate examples and analyse the outcomes in section 5. The last section is conclusion of the whole paper.

2 Principle of Project Scheduling in DM

Project scheduling in DM is to bring the project deviating from the original schedule back on track with recovery strategies when the project is still in progress, as well as to minimize the effect of disruption. As is shown in Figure 1, the flow of project scheduling mainly includes disruption types, recovery strategies, recovery objectives and recovery constraints. This section will explain and analyse the first two, based on which, the latter two aspects will be explained in the next section.

2.1 DISRUPTION TYPES ANALYSIS

This section will make further analysis and explanation on the specific types of disruptions.

1) Task delay disruption

Task delay disruption is mainly caused by project structure disruption. During the implementation of project, some disruptions due to requirements change or other environmental factors may contribute to new task needed to be added or to task stop, which triggers shift in task number. Meanwhile, some change in task priority will result in difference of project structure in project scheduling, which may bring about disruption in the project structure itself. In this way, project structure disruption would make project scheduling vulnerable and infeasible, and eventually tasks in original schedule could not proceed as planned so as to produce delay disruption.

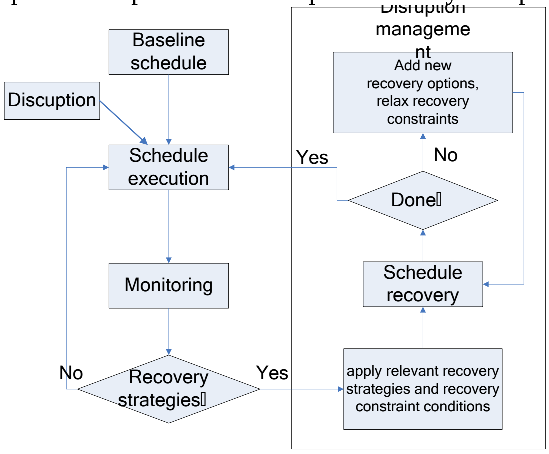


FIGURE 1 Flow of project scheduling in DM

2) Task completion time disruption and resource consumption disruption

Task disruption main consist of completion date disruption and resource consumption disruption. Task completion date disruption refers to the failure of task completion on time as scheduled; in most cases, delay is expected. Resource consumption disruption is the overuse of resource compared with that allocated in original schedule. The impacts of task disruption are as follows: for one thing, delay of task completion will postpone the opening of next task, thereby impairing the implementation of original schedule and raising the cost of the project; for the other, the postponed task will grab the resource initially allocated to other task, which may put off the execution of parallel task.

3) Resource available disruption

Resource disruption refers to that of current available resource, usually denoting the resource available cannot meet the needs of project implementation, namely resource deficiency. This is the commonest type of disruption in project scheduling. The main reasons for this disruption are equipment malfunction, stuff insufficiency, over-consumption of resource by other tasks or projects. The impacts of resource available disruption are the following: on the one hand, the task about to begin needs to be postponed, in order to wait for the completion of other tasks when available resource is restored; on the other hand, backup resource needs to be manoeuvred to meet the requirement of task execution.

4) Delivery time disruption

Delivery time disruption will not impact the implementation of project; however, its influence on project itself is evident. The delay of delivery will cause the offence of project against the contract, thus leading to the failure of the project to some extent. Mainly affected by previous three disruptions, delivery time disruption is valuable when acting as a cost punishment measure to restrict the options of disruption recovery strategies, as those strategies are usually used to curb the disruption in project execution.

2.2 DISRUPTION RECOVERY STRATEGIES

As the key to disruption management, disruption recovery strategies are flexible decision related to disruptive events. These strategies share the same model parameters with disruptive events, while the essence is distinct. Disruption is the fact that project scheduling deviate from the original under the impact of external uncertainties, whereas disruption recovery strategies serve as a decision variable to amend the disrupted process in project scheduling.

1) Task Execution Mode Substitution Strategy

With respect to multi-mode project scheduling, the task execution modes are diverse implementation mode of the project, each of which responds to different completion date and resource consumption. When the task execution is delayed, alternative execution mode could be employed to restore the disrupted scheduling, so as to ensure the task completion on schedule and keep other tasks free from disruption. On the other hand, when the task is delayed for the deficiency of resource, the execution modes of some tasks could be altered to restore the execution of subsequent task. The change in execution mode of tasks may accelerate the task-implementing speed, bringing project scheduling back on track faster, at the expense of project cost, however. To keep the balance between project scheduling and cost control, we define a variable C_{im} as mode-switching cost, the extra cost produced when the execution mode of task *i* is converted from m in original schedule to *m*’ in disruption management. C_{im} reflects the decision cost on the conversion of execution mode of the task.

2) Resource Substitution Strategy

Resource is the foundation of project implementation. The impact of resource insufficiency on project is decisive. Considering this, when disrupted, additional resource could be provided to ensure the project free from impedance of disruption. However, increase of available resource may bring about rise in cost, which means that additional resource provision is restricted within the project budget. Based on the balance between project scheduling and cost control, we define *g*(*r*) as resource punishment coefficient, to express the cost of resource *r* in units.

3 Disruption management model for project scheduling

The current universal form of disruption management model is

$$\min f(x), \tag{1}$$

$$subject\,to\, x \in X . \tag{2}$$

Expression (1) is the objective function, and *f*(*x*) is the disruption degree function; expression (2) is the constraint condition. The objective of disruption management is to minimize the degree of deviation of new scheme from the original after disruption.

According to uncertain disruption types, we propose a disruption management model based on disruption recovery strategies in this section. This model consists of disruption recovery objective function and disruption recovery constraint condition. The disruption recovery objective function is to minimize the project disruption, the component variable includes not only recovery cost caused by the adopted disruption recovery strategy, but also punishment for execution delay and delivery delay due to disruption; the disruption recovery constraint condition defines the constraint conditions for each factor in amended project scheduling.

3.1 OBJECTIVE FUNCTION

We define recovery objective function as the following:

$$\min Q(x) = D(x) + C(x) + G(x) + P(x), \tag{3}$$

where *D*(*x*) is the deviation between project scheduling and implementation after delay of task start, and we define w_i as a weight of delay in start-up punishment, which means the cost caused by the deviation of start delay for per time unit from the original schedule; let the start time of task *i* in the original schedule be s_i, the real start time s'_i.

According to the definition above, *D*(*x*) can be expressed as below,

$$D(x) = \sum_i w_i \left(\left[s_i^{'} - s_i \right]^+ + \left[s_i - s_i^{'} \right]^+ \right), \tag{4}$$

where [z]=max{0, *z*}.

C(*x*) is the extra cost produced when substitution strategy of task execution mode is employed; for multi-mode task, we define the following resource substitution decision variable

$$x_{imt} = \begin{cases} 1 & \textit{Task i is implemented in}\ \text{mode m at the time t} \\ 0 & \textit{Otherwise} \end{cases} . \quad (5)$$

According the definition above, we have

$$C(x) = \sum_{i,m} C_{im} \sum_{t} x_{imt} \ . \quad (6)$$

$G(x)$ is the extra cost of additional resource increase when resource substitution strategy is employed, where we define resource substitution decision variable y_r

$$y_r = \begin{cases} 1 & \text{Resource substitution strategy r is employes} \\ 0 & \textit{Otherwise} \end{cases} . \quad (7)$$

Therefore, G(x) can be expressed as below

$$G(x) = \sum_{r} g(r) y_r \ \cdot \quad (8)$$

$P(x)$ is the contract punishment for project delivery delay; let variable t_θ be the delivery time on schedule, variable η be the contract punishment cost of delivery delay for per time unit, we have

$$P(x) = \eta_i ([t_\theta - \sum_{i,m,t} tx_{imt}]^+ + [\sum_{i,m,t} tx_{imt} - t_\theta]^+) , \quad (9)$$

The objective function is to minimize the sum of $D(x)$, $C(x)$, $G(x)$ and $P(x)$, so as to minimize the effect of disruption on project.

3.2 CONSTRAINT CONDITION

The ultimate objective of disruption management is to ensure the smooth execution of the project, in the way that the disruption recovery strategy can tackle the disruption somehow; thereby, the disruption recovery strategy is expected to meet the constraint condition of project scheduling. We express the constraint condition as below

$$\sum_{m \in M_i} \sum_{t \in T_i} x_{imt} = 1, i \in A \cup A_N , \quad (10)$$

$$\sum_{m \in M_i} \sum_{t \in T_i} tx_{imt} - \sum_{m \in M_i} \sum_{t \in T_i} (t - p_{jm}) x_{imt} \le 0, (i, j) \in P_N \cup P_A \cup (P \setminus P_R) , \quad (11)$$

$$\sum_{t \in \pi_k} \sum_{i \in A \cup A_N} \sum_{m \in M_i} \sum_{q=t}^{t+p_{im}-1} r_{imk} x_{img} \le R_{\pi k} - \rho_{\pi k} + y_{\pi_k} R(\pi_k), \pi_k \in \prod k, k \in K , \quad (12)$$

$$x_{im_0 f_i} = 1, \forall i \in A_F , \quad (13)$$

$$\sum_{m \in M_i} \sum_{t_1 \le t \le t_2} x_{imt} = 1 \cdot \quad (14)$$

Expression (10) ensures that each task has a unique completion time; Expression (11) and Expression (12) are task precedence relationship and resource constraint condition respectively; Expression (13) means the task outside of the time-restoring window is implemented as originally scheduled; Expression (14) is task-restoring constraint.

4 Differential evolution for disruption management in project scheduling

The advancements in meta-heuristics in recent years, related mainly to the development of more efficient computational algorithms have enabled the solution of complex problems by means of numerical optimization algorithms [21]. One of these modern meta-heuristics is the Differential Evolution (DE), an evolutionary computation method. The DE developed by Storn and Price [22] is one of the most superior algorithms. The DE have become widely used in engineering optimization [23–28] due to its simple structure, ease of use, convergence speed, versatility, and robustness. The main difference between genetic algorithms and DE is that, in genetic algorithms, mutation is the result of small perturbations to the genes of an individual (potential solution) while in DE, mutation is the result of arithmetic combinations of individuals.

Stom and Price [22] first introduced the DE algorithm a few years ago. DE is similar to genetic algorithms in that a population of individuals is used to search for an optimal solution. DE combines simple arithmetic operators with the classical operators of crossover, mutation and selection to evolve form a randomly generated starting population to a final solution.

DE offers the advantage of incorporating a relatively simple and efficient form of self-adapting mutation. The fundamental idea behind DE is a scheme whereby it generates the trial parameter vectors. The population of a DE is subject to operators of mutation, crossover and selection. In each time step, DE mutates vectors by adding weighted random vector differentials to them. If the cost of the trial vector is better than that of the target, the target vector is replaced by trial vector in the next generation.

Stom and Price [22] proposed 10 different strategies for DE based on the individual being perturbed, the number of individuals used in the mutation process and the type of crossover used. The strategy implemented here was DE/rand/1/bin, meaning that the target vector is randomly selected, and only one difference vector is used. The bin acronym indicates that the recombination is controlled by a binomial decision rule.

The optimization procedure of DE/rand/1/bin is given by the following steps:

Step 1: Choice of the control parameters, including population size (M), boundary constraints of optimization variables, mutation factor (f_m), crossover rate (c_r), and the stopping criterion (t_{max}).

Step 2: Initialization of population with *M* individuals. Set generation *t*=0. Initialize a population of *i*=1,2,…,*M* individuals (real-valued *N*-dimensional solution vectors) with random values generated according to a uniform probability distribution in the *N*-dimensional problem space as following equation.

$$x_{ij}^{(0)} = x_{\min,j}^{(0)} + rand \cdot (x_{\max,j}^{(0)} - x_{\min,j}^{(0)}), \qquad (15)$$

where *i*=1,2,…,*M* is the individual's index of population; $x_i^{(t)} = [x_{i1}^{(t)}, x_{i2}^{(t)}, \ldots, x_{in}^{(t)}]^{\mathrm{T}}$ stands for the *i*th individual of a population of real-valued *N*-dimensional solution vectors; *t* is the generation (time); *rand* is random value generated according to a uniform probability distribution in [0,1]; x_{max},j and x_{min},j stand for the upper bound and lower bound of the *i*th individual of *j*th real-valued vector.

Step 3: For each individual, evaluate its fitness value.

Step 4: Mutation operation (or differential operation). Mutation is an operation that adds a vector differential to a population vector of individuals according to equation:

$$z_i^{(t+1)} = x_{i,r_1}^{(t)} + f_m^{(t)} \cdot [x_{i,r_2}^{(t)} - x_{i,r_3}^{(t)}], \qquad (16)$$

where $z_i^{(t)} = [z_{i1}^{(t)}, z_{i2}^{(t)}, \ldots, z_{in}^{(t)}]^{\mathrm{T}}$ for the *i*th individual of a mutant vector, *r*1, *r*2 and *r*3 are mutually different integers and are also different from the running index *i* randomly selected with uniform distribution from the set {1,2,…,*i*-1,*i*+1,…,*N*}. The mutation factor $f_m^{(t)} > 0$ is a real parameter, which controls the amplification of the difference between two individuals with indexes r_2 and r_3.

The mutation operation using the difference between two selected randomly individuals may cause the mutant individual to escape from the search domain. If an optimized variable for the mutant individual is outside of the domain search, then this variable is replaced by its lower bound or its upper bound so that each individual should be restricted with the search domain.

Step 5: Evaluate Operation. Evaluate is employed to generate a trial vector by replacing certain parameters of the target vector by the corresponding parameters of a randomly generated donor vector. For each vector $z_i^{(t+1)}$, an index $rnbr(i) \in \{1,2,\ldots,n\}$ is randomly chosen using uniform distribution, and a trial vector, $u_i(t+1) = [u_{i_1}(t+1), u_{i_2}(t+1), \ldots, u_{i_n}(t+1)]^T$, is generated via:

$$u_{i_j}(t+1) = \begin{cases} z_{i_j}(t+1) & \text{if } randb(j) \le CR \quad \text{or } j = rnbr(i) \\ x_{i_j}(t) & \text{if } randb(j) > CR \quad \text{or } j \ne rnbr(i) \end{cases}, \qquad (17)$$

where *j* is the parameter index; $x_{i_j}(t)$ stands for the *i*th individual of *j*th real-valued vector; $z_{i_j}(t)$ stands for the *i*th individual of *j*th real-valued vector of a mutant vector; $u_{i_j}(t)$ stands for the *i*th individual of *j*th real-valued vector after crossover operation; *randb*(*j*) is the *j*th evaluation of a uniform random number generation with [0,1]; *CR* is a crossover rate in the range [0, 1].

Step 6: Selection operation. Selection is the procedure whereby better offspring are produced. To decide whether or not the vector $u_i(t+1)$ should be a member of the population comprising the next generation, it is compared with the corresponding vector $x_i(t)$. Thus, if *f* denotes the objective function under maximization, then

$$x_i(t+1) = \begin{cases} u_i(t+1) & \text{if } f(u(t+1)) > f(x_i(t)) \\ x_i(t) & \text{otherwise} \end{cases}. \qquad (18)$$

In this case, the value of objective function cost of each trial vector $u_i(t+1)$ is compared with that of its parent target vector $x_i(t)$. If the objective function *f* of the target vector $x_i(t)$ is upper than that of the trial vector, the target is allowed to advance to the next generation. Otherwise, the target vector is replaced by a trial vector in the next generation.

Step 7: Verification of the stopping criterion. Set the generation number for *t*=*t*+1. Proceed to Step 3 until a stopping criterion is met, usually a maximum number of iterations (generations), $t_{\max}$. The stopping criterion depends on the type of problem.

5 Simulation Experiments

To verify the validity of disruption management and algorithm discussed in this chapter, we conduct the experiments using C++ for encoding, and employ the multi-mode scheduling test packs J20 and J30 in project scheduling standard question bank PSPLIB[29], to make test on the PC with Intel® Core™2 Duo 2.4GHz CPU. Here J20 and J30 include 20 and 30 tasks respectively (not including virtual tasks); for each task, there are 3 types of execution modes for option, in each of which, the task duration is 1-10 time units, consuming 2 kinds of renewable resources and 2 kinds of non-renewable resources. For test questions of each group, randomly generate a question. Therefore, there are 640 questions for each group; however, there exists no feasible solutions for some of these test questions. Because of this, eliminate those insolvable questions, there are 554 project cases left in J20, and 552 left in J30. During the experiment, we make discussion and analysis on the minimal disruption cost of the projects with and without project delivery time limits respectively.

5.1 PARAMETERS SETTING

The cases in question bank J20 and J30 are both about project scheduling under certain circumstance. However, as project scheduling is in the complex environment during the execution, some disruption parameters in J20 and J30 needs to be set to make the project scheduling disruption management close to reality. Firstly, in dynamic environment, the completion time of project scheduling is no shorter than that of baseline scheduling at planning stage. For the baseline project completion time generated by the Certain Scheduling Algorithm solution in J20, we extend it by 20%, 10% and 0% respectively as the expected completion time; for that in J30, we extend it by 40%, 30% and 20% respectively as the expected completion time. The extension of expected completion time will increase the amount of scheduling schemes, thereby affecting the time cost of solution directly. Secondly, we add two tasks to J20 and J30 respectively, which stand for the disruption of project structure caused by the requirement change by clients at the execution stage. Besides, we define the 10%-extension of completion time of two random tasks of the project caused by disruption, and 10% more resource consumption in the execution of two random tasks than the original. In addition, the resource consumption may be 5% fewer at some period of execution than the original supply. Lastly, the parameter values are defined for delay in start-up weight w_i, mode-switching cost, delayed delivery punishment, resource punishment coefficient and the time of delay in delivery. The detailed question parameter settings are shown in Table 1.

Here are the parameter settings in DE algorithm, the maximum evolutionary generation T is set to 100, the population size N is 40, the crossover probability P_c is 0.9, and the scaling factor F is 0.7. The proposed algorithms adopt penalty function method to deal with constraints using following equations:

$$f(w) = F(w_i, u_{hj}, s_{ih}) + M(\sum_{i=1}^{m} w_i - 1)^2 , \tag{19}$$

where M is the penalty impact. In the numerical experiments, M is set to 105 for the J30 and J60, and 106 for the J60 and J120.

TABLE 1 Parameter settings

Benchmark	J20, J30
Question bank size	554, 552
Project structure disruption	2 new tasks
New task 1	Execution duration:3, resource consumption, precedence task, subsequent task
New task 2	Execution duration:3, resource consumption, precedence task, subsequent task
Execution duration extension	2 random tasks, 10% extended
Resource consumption	2 random tasks, random resource 10% more consumed
Resource disruption	5% fewer resource, resource recovery time randomly got from [1, 5]
Delay in start-up punishment weight w_i	Uniform distribution between [1,10]
Mode switching cost	Uniform distribution between [0,5]
Delivery delay punishment	20
Resource punishment coefficient	3
Delayed time in delivery	No longer than 40% of original execution duration

5.2 COMPUTING RESULTS AND ANALYSIS

For the computing process, at first project scheduling method is employed to get the optimal/suboptimal scheduling scheme as the planning one, and then computation is implemented on the minimal disruption cost of projects with and without delivery time limit, based on the question parameters and algorithm parameters discussed above. Meanwhile, to simulate the real situation in project execution and analyse the influence of different disruption strategies on project execution, single and combined disruption strategies are taken respectively in simulative computations.

1) No delivery time limit

Table 2 shows the computing results for minimal disruption cost without delivery time limit with different disruption strategies.

TABLE 2 Computing results

Benchmark	Strategy	Average disruption cost	Average execution duration	Scheduled average execution duration	Real deviation rate (%)
J20	Task execution mode	238.46	34.52	27.71	24.58%
	Resource substitution	295.73	33.29	27.71	20.14%
	Combined	172.39	32.12	27.71	15.91%
J30	Task execution mode	484.24	46.85	33.38	40.35%
	Resource substitution	566.73	44.27	33.38	32.62%
	Combined	401.96	42.43	33.38	27.11%

As can be seen in Table 2, size of question bank exerts great influences on disruption in project execution. The deviation rate of average disruption cost in benchmark J20 from original scheduling is superior to that in J30, which indicates that the more complex the project structure is, the greater disruptions received in dynamic environment, thus the more difficult to implement disruption recovery.

Meanwhile, different disruption recovery strategies also have enormous influences on the disruption recovery cost. With respect to this cost, that of projects with combined strategies is the minimal, followed by that with Task Execution Mode Substitution Strategies, while that with Resource Substitution Strategies is the largest. Concerning the deviation rate of total real execution duration from the original scheduling that with Combined Disruption Recovery Strategies is the least, and next is the one with Resource Substitution Strategies, while that with Task Execution Mode Substitution Strategies is the greatest.

Based on the results above, for the disruption during project execution, resource supply increase can shorten the total real execution duration which entails dramatic cost rise, thus not the optimal scheme; given the resource condition, change of task execution mode can bring down the disruption cost to some degree at the expense of execution duration, thus leading to delay in delivery; combined strategies can reduce the risk of project execution and disruption cost to the largest extent by changing the execution mode with the permission of resource substitution.

2) Given delivery time limit

Table 3 and 4 reveals the computing results for the questions in J20 and J30 on the minimal disruption cost with delivery time limit with different disruption strategies.

TABLE 3 Computing results

Benchmark	Strategy	Average disruption cost	Average execution duration	Scheduled average execution duration	Real deviation rate (%)
≤40%	Task execution mode	253.61	38.73	27.71	39.77%
	Resource substitution	314.03	37.49	27.71	35.29%
	Combined	199.47	36.53	27.71	31.83%
≤30%	Task execution mode	217.46	35.64	27.71	28.62%
	Resource substitution	279.83	34.76	27.71	25.44%
	Combined	164.35	33.98	27.71	22.63%
≤20%	Task execution mode	249.79	33.18	27.71	19.74%
	Resource substitution	306.02	32.39	27.71	16.89%
	Combined	193.72	31.83	27.71	14.87%
≤10%	Task execution mode	268.45	30.42	27.71	9.78%
	Resource substitution	329.78	29.87	27.71	7.80%
	Combined	228.75	29.12	27.71	5.09%

TABLE 4 Computing results

Benchmark	Strategy	Average disruption cost	Average execution duration	Scheduled average execution duration	Real deviation rate (%)
≤50%	Task execution mode	512.33	49.63	33.38	48.68%
	Resource substitution	589.37	48.47	33.38	45.21%
	Combined	433.92	46.81	33.38	40.23%
≤40%	Task execution mode	473.51	46.71	33.38	39.93%
	Resource substitution	543.39	45.26	33.38	35.59%
	Combined	403.58	43.98	33.38	31.76%
≤30%	Task execution mode	525.46	43.18	33.38	29.36%
	Resource substitution	587.24	42.39	33.38	26.99%
	Combined	436.69	40.43	33.38	21.12%
≤20%	Task execution mode	566.87	39.82	33.38	19.29%
	Resource substitution	609.35	38.87	33.38	16.45%
	Combined	475.43	37.12	33.38	11.20%

It can be seen in Table 3 and 4 that similar to Table 1, size of benchmark and different disruption recovery strategies still have significant influence on the disruption during project execution. With regards to deviation of different delivery time limits, the deviation rate of average disruption cost from original scheduling in question bank J20 is superior to that in J30. As for the solution results of each question bank, different disruption recovery strategies is also greatly influential on recovery cost. The recovery cost with combined strategies has the least recovery cost, followed by that with Task Execution Mode Substitution Strategies, while that with Resource Substitution strategies is the largest. Judging from the deviation rate of total real execution duration from the original scheduling, that with combined strategies is the minimal, and that with Resource Substitution Strategies comes next, while that with Task Execution Mode Substitution Strategies is the largest.

For the same benchmark, the delivery time limit also has enormous influence on the recovery cost and total execution duration. As we can see from Table 3, in benchmark J20, when the delivery time limit is not 30% longer than the original scheduling, the recovery cost becomes the highest as the delay time in delivery shortens; when the delivery time limit is not 30% longer than the original delivery schedule, the recovery cost is the least, followed by that when delivery time limit is not 20% longer, while that when delivery time limit is not 10% longer is the highest. This indicates that the bigger the deviation of delivery time limit from the scheduled total execution duration, the lower the disruption recovery cost is. That is, loose delivery time limit may bring low execution cost. The same conclusion could be got form the result in Table 3.

However, the delivery time limit could not be extended without boundaries. We can see from Table 2 and Table 3 that when the delivery time limit is not 40% and 50% longer than the scheduled execution duration, the disruption execution cost increases. This is because despite of the decline of cost of mode switching and resource substitution in project execution, the degree of this kind of decline could not compensate the punishment cost caused by delayed delivery. Therefore, in project management, proper project completion schedule helps reduce the disruption cost that the dynamic environment entails in the project execution process.

Based on the computing results in single and combined strategies respectively with and without delivery time limit, to deal with disruption in execution process of project scheduling, the selection of disruption recovery strategies needs to be in accordance with delivery time and cost of the project. Within the delivery time limit, the total project execution duration can be extended to the largest extent and the execution mode can be switched to ensure a somehow soft execution and to reduce the disruption cost of the project. Meanwhile, within the permission of developing cost, Resource Substitution Strategies could be employed to shorten the overall execution duration and lower the disruption cost. This is one of the reasons why developing task outsourcing is employed currently in a large number of products development projects.

6 Conclusions

This chapter analyses the disruption types in the execution process of project scheduling, namely task delay disruption, task duration and resource consumption disruption, resource available disruption and delivery time disruption, and raises two corresponding disruption recovery strategies, Task Execution Mode Substitution Strategy and Resource Substitution Strategy. On the basis of this, we establish a disruption management model in dynamic environment, whose target function is to minimize disruption recovery cost.

At the end, we use DE to solve the computation of minimal disruption cost problem and analyse the computing results, with and without delivery time limit in multi-mode scheduling test pack J20 and J30 from the standard question bank in PSPLIB. The analysis indicates that within the permission of resource substitution, the change of task execution mode can reduce the executive risk and decrease the disruption cost to the greatest extent. This partly explains why task development outsourcing is widely used for lots of product development projects.

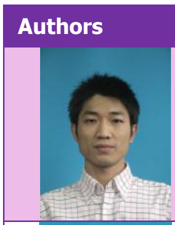
Through the model and computations in this paper, we can draw the following conclusion for practical engineering applications: task outsourcing for product development and proper scheduling of project execution duration can facilitate the disruption-resistance and risk-resistance capability of product development project, thus improving the product developing efficiency and quality.

Acknowledgments

This study was provided by China Postdoctoral Science Foundation funded project (Grant no. 2013M542091), and the Fundamental Research Founds for National University, China University of Geosciences (Wuhan) (Grant no. CUG120824).

References

[1] Colin E, et al 2000 *Journal of the Operational Research Society* **51**(3) 291-300
[2] Clausen J, Hansen J, Larsen J, Larsen A 2001 *Disruption management* **28** 40–3
[3] Yu G, Qi X 2004 *Disruption Management: Framework, Models, Solutions and Applications* World Scientific Publishers: Singapore
[4] Yang, Jian, Xiangtong Qi, Gang Yu 2005 *Naval Research Logistics* **52**(5) 420-42
[5] Bean J C, et al 1991 *Operations Research* **39**(3) 470-83
[6] Xia Yusen, et al 2004 *IIE transactions* **36**(2) 111-25
[7] Qi Xiangtong, Bard J F, Gang Yu 2006 *International Journal of Production Economics* **103**(1) 166-84
[8] Herroelen W, Roel L 2005 *European journal of operational research* **165**(2) 289-306
[9] Zhu Guidong, Bard J F, Gang Yu 2005 *Journal of the Operational Research Society* **56**(4) 365-81
[10] Xu Minghui, et al 2003 *Journal of Systems Science and Systems Engineering* **12**(1) 82-97
[11] Qi Xiangtong, Bard J F, Gang Yu 2004 *Omega* **32**(4) 301-12
[12] Al-Fawzan, Mohammad A, Mohamed Haouari 2005 *International Journal of production economics* **96**(2) 175-87
[13] Hur Daesik, Mabert V A, Bretthauer K M 2004 *Production and Operations Management* **13**(4) 322-39
[14] Herroelen W, Roel L 2005 *European journal of operational research* **165**(2) 289-306
[15] Howick S 2003 *Journal of the Operational Research Society* **54**(3) 222-9
[16] Van de Vonder Stijn, et al 2007 *Computers & Industrial Engineering* **52**(1) 11-28
[17] Lambrechts O, Demeulemeester E, Herroelen W 2007 International *Journal of Production Economics* **111**(2) 493-508
[18] Lambrechts O, Demeulemeester E, Herroelen W 2008 *Journal of scheduling* **11**(2) 121-36
[19] Deblaere F, Demeulemeester E, Herroelen W 2011 *Computers & Operations Research* **38**(1) 63-74
[20] Taghizadeh N, Neirameh A 2012 *International Journal of Computing Science and Mathematics* **3**(4) 332-40
[21] Storn R, Price K 1997 *Journal of global optimization* **11**(4) 341-59
[22] Coelho L S, Mariani V C 2006 *IEEE Trans Power System* **21**(2) 989-96
[23] Coelho L S, Mariani V C 2006 *IEEE Trans Power System* **21**(3) 1465
[24] Liu B, Wang L, Jin Y H, Huang D X, Tang F 2007 *Chaos, Solutions & Fractals* **34**(2) 412–9
[25] Chang Wei-Der 2007 *Chaos, Solutions & Fractals* **32**(4) 1469-76
[26] Coelho L S, Mariani V C 2007 *Energ. Convers Manage* **48**(5) 1631-39
[27] Peng Bo, et al 2009 *Chaos, Solutions & Fractals* **39**(5) 2110-8
[28] Kolisch R, Sprecher A 1996 *European Journal of the Operational Research* **96** 205-16

Authors	
	**Weiming Chen, born on April 21, 1981 at Hanchuan City, China** **Current position, grades:** Lecturer at Faculty of Engineering at China University of Geosciences, Wuhan, China **University studies:** PhD in Industrial Engineering from Huazhong University of Science and Technology in 2011. **Scientific interest:** modern design theory and methods, complex project management and product development. **Publications:** 5 papers
	Xiaoyang Ni, born on January 15, 1970 at Shen Yang City, China **Current position, grades:** vice-professor at Facuity of Engineering at China University of Geosciences. **University studies:** PhD in China University of Geosciences (Wuhan) in 2006. **Scientific interest:** safety engineering and management science **Publications:** 8 papers
	Hailin Guo, born on September 10, 1973 at Tangshan City, China **Current position, grades:** Lecturer at Faculty of Engineering at China University of Geosciences, Wuhan, China. **University studies:** PhD in Geological Engineering from China University of Geosciences in 2005. **Scientific interest:** risk analysis theory and methods **Publications:** 19 papers

The optimal dynamic robust portfolio model

Xing Yu[*]

Department of Mathematics & Applied Mathematics, Hunan University of humanities, science and technology, Loudi, 417000, P.R. China

Received 1 March 2014, www.tsi.lv

Abstract

This paper is concerned with the optimal dynamic multi-stage portfolio of mean- dynamic var based on high frequency exchange data with the constraint of transaction costs transaction volume. The proposed solution approach is based on robust optimization, which allows us to obtain a worst best but exact and explicit problem formulation in terms of a convex quadratic program. In contrast to the mainstream stochastic programming approach to multi-period optimization, which has the drawback of being computationally intractable, the proposed setup leads to optimization problems that can be solved efficiently.

Keywords: dynamic portfolio, mean-var, robust, high frequency exchange

1 Introduction

Markowitz [1] is the first to formulate the model for maximizing the expected return and minimizing its risk. He only considered the case of a single-period investment. However, investment behaviour, especially the investment behaviour of institutional investors are often long. For a long-term investor, he will adjust portfolio positions with the investment environment changes timely, rather than portfolio immutable and frozen early construction to keep to the investment plan period. With the development of computer, obtaining high frequency data more convenient is benefit to establish of dynamic investment strategy. So in order to reformulate the single stage asset allocation problem, a multi-period framework is a decision process has been developed by using multi-period stochastic programming, see for example [2-4] Kall (1976), Wallace (1994) and Breton (1995). The academics modelled the mean or the variance of total wealth at the end of the investment horizon as either linear stochastic programming or quadratic stochastic programming in Gulpinar et al. (2002, 2003) [5-6]. Inoue, and Wang (2010) [7] proposed a dynamic portfolio selection optimization with bankruptcy control for absolute deviation model.

Decision problems arising in engineering, finance, logistics etc. are usually dynamic and affected by uncertainty. However, in optimal portfolio model, it requires the knowledge of both mean and covariance matrix of the asset returns, which practically are unknown and need to be estimated. The standard approach, ignoring estimation error, simply treats the estimates from the history data as the true parameters and plugs them into the optimal portfolio optimization model derived under the mean–variance framework. That is, using known information replace the unknown information. Moreover, just as everyone knows, all the factors mentioned above are affected by human’s subjective intention. Thus, in these cases, it is impossible for us to predict the probability distributions of the returns of risky assets. To solve the problem of uncertainty, there are two routes. One is fuzzy theory. Seyed Jafar Sadjadi [8] considered several portfolio selection problems including probabilistic future returns with ambiguous expected returns assumed as random fuzzy variables. Yong etc. [9] deal with multi-period portfolio selection problems in fuzzy environment by considering some or all criteria, including return, transaction cost, risk and skewness of portfolio. Similar literatures see to [10-15]. Another is robust optimal. Robust optimization is another approach towards optimization under uncertainty. Anna [16] deal with a portfolio selection model in which the methodologies of robust optimization are used for the minimization of the conditional value at risk of a portfolio of shares. Yongma Moon [17] constructed the robust portfolio model represented portfolio risk by the return standard deviation, avoiding large computing problems. Seyed Jafar Sadjadi etc. [18] presented a new portfolio selection model for uncertain information, and solved the model with different robust method. Nalan [19] proposed the multi-period mean–variance portfolio optimization model with worst-case robust decisions. The mentioned literatures only pay attention to low frequency data, that is, the trading is controlled in day or even a month or longer time range. It is not reasonable. For example, there may exist a good chance in a day, also called intra-daily the investor should pursuit the most favourable business opportunities rather than continue to wait for the final trading deadline.

The rest of the paper is organized as follows. In Section 2, the multi-period mean variance optimization problem is described. In Section 3, we introduce the robust optimal methodology. Section 4 focuses on an empirical research of a optimal portfolio model with five risk assets in Chinese market. Conclusions are given in section 5.

[*] ***Corresponding author*** e-mail: hnyuxing@163.com

2 Mean- dynamic var multi-stage portfolio model

Suppose there are n alternative risk assets. $R_{it}, i=1,2\cdots n; t=1,2\cdots T$ are the yields of asset i at stage t, with expectation $r_{it}=E\left(R_{it}\right)$ and covariance matrix $\Sigma_t=\left(\sigma_{ij}\left(t\right)\right)_{n\times n}$, $\sigma_{ij}\left(t\right)=COV\left(R_{it},R_{jt}\right)$. At the beginning of stage t, the portfolio share is x_{it}. And at stage t, buying and selling respectively are b_{it},s_{it} .So, the portfolio share at stage t is $x_{i,t-1}+b_{it}-s_{it}$. The purchase and sale transactions need transaction costs, suppose the cost are $C\left(b_{it}\right)$ and $C\left(s_{it}\right)$. Another factor to consider is wealth constraint. Suppose an investor have the initial wealth S_0, then at stage t, its transition equation S_t: $S_t=\left(1+I_{pt}\right)S_{t-1}$, where I_{pt} is the portfolio earnings, $I_{pt}=r_{pt}-\sum_{i=1}^{n}\left[C\left(b_{it}\right)-C\left(s_{it}\right)\right]$, in which the expected rate of return $r_{pt}=\sum_{i=1}^{n}r_{it}\left(x_{it}+b_{it}-s_{it}\right)$.

At stage t, the VaR of the portfolio is $f_t=\Phi^{-1}\left(p\right)\sqrt{x_t{}^T\Sigma_t x_t}-r_t^T x_t$, $x_t=\left(x_{1t},x_{2t}\cdots x_{nt}\right)^T$, $r_t=\left(r_{1t},r_{2t}\cdots r_{nt}\right)^T$.

The optimal dynamic multi-stage portfolio of mean-dynamic VaR is $\min f_t$

$$s.t\begin{cases} S_t=\prod_{l=1}^{t}\left(1+I_{pl}\right)S_0 \\ 0\le b_{it}\le\sum_{j=1,j\ne i}^{n}x_{jt} \\ 0\le s_{it}\le x_{it} \\ b_{it}s_{it}=0\end{cases}\qquad\text{(Model 1)}$$

Generally, invest strategy aims to the future stage that at stage t according to stage $t-1,t-2\cdots1$, which are the given information. In model 1, the objective is to minimax the portfolio risk at stage t, and the first constraint means the state transition equation of wealth. The second constraint express not to short buy. In addition, not to short sell in the third constraint. Buying and selling at the same time is limited.

We should describe the relation of the state transition equation of wealth and the initial wealth.

$$\begin{aligned}S_t&=\left(1+I_{pt}\right)S_{t-1}\\&=\left(1+I_{pt}\right)\left(1+I_{p,t-1}\right)S_{t-2}\\&=\cdots\\&=\prod_{l=1}^{t}\left(1+I_{pl}\right)S_0\end{aligned}.$$

3 The classical portfolio model and its robust counterpart

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):

(M1) $\min_X X\Sigma X'$

$$\text{subject to}\begin{cases}\sum x_i=1 & (1)\\ \sum E\left(r_i\right)x_i=r_p & (2)\\ 0\le x_i\le1 & (3)\end{cases}$$

where x_i are portfolio weights, r_i is the rate of return of instrument i, and Σ 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. We consider the uncertain set for return mean. we define r' is the estimation of real value r, the uncertainly set I as $I=\left\{r:r_i'-s_i\le r_i\le r_i'+s_i\right\}$ for mean. According to Anna [16], the robust counterpart: $\min\sum r_i x_i\ge r_p$ can be transferred to the following form:

$$\begin{cases}\sum r_i'x_i-\sum s_i m_i\ge r_p\\ m_i\ge x_i\\ s_i\ge0\end{cases}$$

The model handling process reflects the robust optimization, the worst best solution.

4 Empirical research

4.1 THE MODEL DESCRIBE

In our model, for simply, let the confidence level $c=97.5\%$, so $\Phi^{-1}(p)=1.96$. We suppose an investor wish pursuit no less than 7% profit. So the first constraint is reformed as $S_t\ge(1+8\%)S_0$ or $\prod_{l=1}^{t}(1+I_{pl})\ge1.08$. Transaction costs function is $C(b_{it})=0.008b_{it}$ and $C(s_{it})=0.008s_{it}$. It reforms the model step as

(i) select the high frequency minute data, divide them in terms of a given minutes(such as 5 minutes)and get the investment stages n.

(ii) calculate the sample mean and covariance of each stock in every stage, we express them as r_{it},Σ_{it} approximately.

To avoid the risk effectively, we consider the robust optimization. It means if we make a strategic decision at one stage, we change the sample mean and covariance in an interval whose left is the minimax and right is the maximum of the stages before current stage.

(iii) solve the optimal portfolio model. The solutions show that at each stage, in order to seek the objective of minimax risk under the constraints, how to adjust the invest share.

For simple, given the original respective share are $\frac{1}{n}$. We special the stage is 2 and 3. $x_1 = (x_{11}, x_{21} \cdots x_{n1})$, $b_1 = (b_{11}, b_{21} \cdots b_{n1})$, $s_1 = (s_{11}, s_{21} \cdots s_{n1})$. The construction and solution steps are

$$\min \quad f_1 = 1.96\sqrt{x_1^T \Sigma_1 x_1} - r_1^T x_1 .$$

The constraints are

$$\begin{cases} S_1 = (1 + I_{p1}) S_0 \geq 1.08 S_0 \;\; or \;\; 1 + I_{p1} \geq 1.08 \\ 0 \leq b_{i1} \leq \sum_{j=1, j \neq i}^{n} x_{j1} \\ 0 \leq s_{i1} \leq x_{i1} \\ b_{i1} s_{i1} = 0 \\ r_{p1} = r_1 (x_1 + b_1 - s_1)^T \\ I_{p1} = r_{p1} - 0.008 b_1 I^T + 0.008 s_1 I^T \end{cases},$$

where $r_1 = (r_{11}, r_{21} \cdots r_{n1})$ is the return vector and Σ_1 is the covariance at stage 1 from the data, I^T is the transpose of n-dimensional unit vector. To solve this model, we obtain the adjustment strategy b_1 and s_1. So then, the portfolio shares are updated.

For stage 2, the objective is

$$\min \quad f_2 = 1.96\sqrt{x_2^T \Sigma_2' x_2} - r_2^{\prime T} x_1 .$$

The objective is not the same as the stage 1 case. Where Σ_2' is not certain but change from $\min\{\Sigma_1, \Sigma_2\}$ to $\max\{\Sigma_1, \Sigma_2\}$, the same as the uncertain return r_2' changes from $\min\{r_1, r_2\}$ to $\max\{r_1, r_2\}$. The constraints are similar only except the first one is $(1 + I_{p1})(1 + I_{p2}) \geq 1.08$.

For stage 3, we also obtain the corresponding optimal model.

4.2 EMPIRICAL STUDY

In this part, it focuses on an empirical research of a optimal portfolio model with five risk assets in Chinese market who show a good momentum in 2013. We choose the high frequency data in 5 minutes of five stocks NO 002583600694, 600089, 601166 and 600276 from 2012-01-04 09:01:00 to 2013-9-4 15:00:00,58320 data, which is from GATA database. However, for simply, we construct the optimal portfolio model only suppose a stage conclude five moths. That is, there is 5 stages. From solving the model, the dynamic portfolio is:

At stage 1, $a_{11} = 0.8034, a_{21} = a_{31} = 0, a_{41} = 0.164, a_{51} = 0$ and $s_{11} = s_{21} = s_{31} = 0,\; s_{41} = s_{51} = 0$.

At stage 2, $a_{12} = 0, a_{22} = 0.775, a_{32} = 0, a_{42} = 0.083, a_{52} = 0$ and $s_{21} = s_{22} = s_{32} = 0,\; s_{42} = 0, s_{52} = 0.561$.

At stage 3, $a_{31} = a_{32} = a_{33} = a_{34} = a_{35} = 0$ and $s_{31} = s_{32} = 0, s_{33} = 0.11, s_{34} = s_{35} = 0$.

At stage 4 not do any tradings.

At stage 5, $a_{51} = 0.381, a_{52} = 0.042, a_{53} = a_{54} = a_{55} = 0$ and $s_{51} = 0, s_{52} = 0,\; s_{53} = 0.0844, s_{54} = s_{35} = 0$

5 Conclusion

Dynamic portfolio is superior than a single static model. And in reality, the return and covariance change with some factors of future, it is not inappropriate to use the history features to describe the future case. However, for an investor, how to control risk is the first factor that he should consider when making an investment. This paper construct the dynamic portfolio model under robust counterpart, in which it focus minimax the risk under the constraints. It also considers the wealth budget and transaction costs, which is very important in dynamic investing. Because it is not a sensible stuff to trade frequently ignoring commission. At last, an empirical study choosing five stocks from Chinese market to test validity of the models, giving the strategies. In fact, the solution this paper mentioned can be extended to higher frequency in the data model, investment strategy curve (including the sale point in time and quantity) could provide investors better suggestions.

Acknowledgment

This research is supported by Department of Education Topics of Hunan Province (NO 12C0749) and scientific research fund of Hunan provincial education department (NO 2A077).

References

[1] Markowitz H M 1952 Portfolio selection *Journal of Finance* **7** 77-91

[2] Kall P 1976 *Stochastic Linear Programming* Springer: Berlin

[3] Kall P, Wallace S W 1994 *Stochastic Programming* Wiley: New York

[4] Breton M, El Hachem S 1995 Algorithms for the solution of stochastic dynamic minimax problems *Computational Optimization and Applications* **4** 317–45

[5] Gulpinar N, Rustem B, Settergren R 2002 Multistage stochastic programming in computational finance *Computational methods in decision making economics and finance: Optimization Models* **11** 33–45

[6] Gulpinar N, Rustem B, Settergren R 2003 Multistage stochastic mean–variance portfolio analysis with transaction cost *Innovations in Financial and Economic Networks* **3** 46–63

[7] Takahashi H Inoue, S Wang 2010 Dynamic portfolio optimization with risk control for absolute deviation model *European Journal of Operational Research* **201**(2) 349-64

[8] Seyed Jafar Sadjadi, Mohsen Gharakhani, Ehram Safari 2012 Robust optimization framework for cardinality constrained portfolio problem *Applied Soft Computing* **12** 91-9

[9] Yong-Jun Liu, Wei-Guo Zhang, Wei-Jun Xu 2012 Fuzzy multi-period portfolio selection optimization models using multiple criteria *Automatica* **48** 3042–53

[10] Takashi Hasuike, Hideki Katagirib, Hiroaki Ishiia 2008 Portfolio selection problems with random fuzzy variable returns *Fuzzy Sets and Systems* **160**(18) 2579–96

[11] Wang S, Zhu S 2002 On fuzzy portfolio selection problem *Fuzzy optimization and decision making* **4** 361-77

[12] Deschrijver G 2007 Arithmetic operators in interval-valued fuzzy set theory *Information Sciences* **177** 2906-24

[13] Li D-F 2010 TOPSIS-based nonlinear-programming methodology for multiattribute decision making with interval-valued intuitionistic fuzzy sets *IEEE Transactionson Fuzzy Systems* **18** 299-311

[14] Li D-F 2010 Linear programming method for MADM with interval - valued intuitionistic fuzzy sets *Expert Systems with Applications* **37** 5939-45

[15] V Lakshmana Gomathi Nayagam, Geetha Sivaraman 2011 Ranking of interval-valued intuitionistic fuzzy sets *Applied Soft Computing* **11** 3368-72

[16] Quaranta A G, Zaffaroni A 2008 Robust optimization of conditional value at risk and portfolio selection *Journal of Banking & Finance* **32** 2046–56

[17] Yongma Moona, Tao Yao 2011 A robust mean absolute deviation model for portfolio optimization *Computers & Operations Research* **38** 1251-8

[18] Seyed Jafar Sadjadi, Mohsen Gharakhani, Ehram Safari 2012 Robust optimization framework for cardinality constrained portfolio problem *Applied Soft Computing* **12** 91–9

[19] Nalan Gu¨lpınara Rustem 2007 Worst-case robust decisions for multi-period mean–variance portfolio optimization *European Journal of Operational Research* **183** 981–1000

Authors

Xing Yu, born on February 15, 1981, in Xianning City of Hubei province

Current position, grades: Hunan university of humanities, Science and technology (China); Lector
Professional interests: Applicate mathematics; Finance model
Research interests: the optimal portfolio model; mathematical model

Travel route choice model based on regret theory

Baohui Jin[*]

College of Transportation & Logistics, Southwest Jiaotong University, North 1st section of 2nd Ring Road 111, 610031 Chengdu, China

Received 1 March 2014, www.tsi.lv

Abstract

Travel route choice behaviour research is a hot issue in the field of urban traffic planning, and it mainly researches the traveller's route choice decisions under uncertainty conditions, which theory includes such as expected utility theory, prospect theory, and regret theory. Based on the analysis of expected utility theory and prospect theory's applicable condition and the insufficiency, this paper establishes a travel route choice model according to regret theory. Study shows that people always try to avoid occur that other options is better than that selected option, and the properties of selected option cannot be replaced each other, which fits regret minimization of regret theory. The travel route choice model based on regret theory is simpler than others, and it is suitable for describing traveller's route choice behaviour under uncertainty conditions.

Keywords: urban traffic, travel route choice, regret theory, Bayesian updating

1 Introduction

Travel route choice is a major decision for a traveller in the process of travelling or before travel. Research on travel route choice behaviour is a hot issue in the fields of urban traffic planning and navigation. Because of traffic network's complexity and time-varying characteristic, and traveller's own differences, this makes traveller's route choice behaviour is uncertain in some degree.

At present, the travel route choice behaviour research mainly references expected utility theory and prospect theory. Expected utility theory has been widely applied in traffic value comparison, and it assumes that people are perfectly rational. Travellers will choose the option which has the largest expected utility according to the complete information those travellers mastered. However, it is difficult to fully master the accurate traffic information for the travellers in fact, and the travellers' preferences and attitudes are not entirely rational, the practical behaviour of travellers' route choice doesn't fully respect the axiomatic system of expected utility theory [1, 2].

Considering travellers' limited rationality as "economic man", Kahneman and Tversky in 1979 raised the prospect theory on the basis of Simon's limited rationality theory, and improved the theoretical model in 1992 [3,4]. Compared with the utility function in expected utility theory, prospect theory introduced the concept of value function; there is a reference point concept for the value function. Reference point is the demarcation point to distinguish the gains and losses. When the reference point uses a different index, the same thing may produce a different result. Wang Yan and Zhang Li proved that under the uncertainty conditions of road network "prospect theory" is more appropriate to describe travellers' decision-making behaviour [5]. Zhang Yang's empirical research also shows the behaviour that people choose vehicles' travel time and paths are consistent with prospect theory under uncertain environment [6]. Luo applied the prospect theory to travel route choice and proved its effectiveness through a example [7].

Prospect theory describes a two stages decision-making model using a two dimensions model of evaluation: valuing outcomes function and weighting of probabilities function. In addition, it involves reference point's definition and using, to distinguish gain or loss. Although the reference point in economics domain is usually unique, in the traffic fields single reference point is not sufficient to solve practical problems. Jou and Kitamura assumes two reference points when they studied travel route choice problem, including the earliest acceptable arrival time and work start time [8]. Schwanen and Ettema set three reference points in the study of route choice behaviour of parents to transfer their children, including the possible departure time, the probability of arriving on time, and late penalties [9]. De Moraes Ramos researches the diversity of reference point in prospect theory [10]. It restricted the application of prospect theory in the travel route choice fields because of the reference point's diversity and complexity.

For a normal traveller, the positive effect is a factor of travel route choice, but the negative consequences that may occur also must be accepted. The researchers tried to looking for a more realistic theory to explain and describe the travellers' route choice behaviour. Among them, Loomes and Sudgen in 1982 [11], Bell in 1982 [12] independently proposed a "regret theory", and they pointed out that the single factor's utility function cannot explain the behaviour of non-rational decision well. People will compare the actual situation and possible

[*] ***Corresponding author*** e-mail: jetame@163.com

situation according to their decision-making. If they find their choosing can get better results than other options, they will rejoice. On the contrary, they will feel regret. On the basis of regret theory Casper proposed a random regret minimization model(RRMM), and applied it to travel route choice [13, 14]. RRMM supposed that the satisfaction degree of a travel route not only depends on the utility of selected travel route, but also on the regret of other options' possible better utility. Giselle, etc. compared the expected utility theory, prospect theory and regret theory in travel route choice behaviour prediction, and pointed out that regret theory can be more truly to describe the route choice behaviour than the expected utility theory, and has a simpler form than prospect theory algorithms [15]. Regret theory has only one parameter, namely a regret aversion parameter. When simulating travel route choice behaviour we only need to determine the regret aversion parameter and it is easy to identify according to the foregone experience.

This paper builds a traveller route choice model based on regret theory, and using Bayesian theory to update traveller's regret utility, analyses the travellers' route choice behaviour under uncertain conditions. Then we give a simple numerical example based on a three-link network. At last, we studies the characteristic of regret theory in travel route choice problem.

2 Model based on regret theory

2.1 REGRET UTILITY FUNCTION

Regret theory thinks that travellers' decisions depend not only on the selected route's expected utility, but also on the unselected route's expected utility. If the decision maker finds other unselected routes can produce better result he will feel regret. Conversely, if the selected route's expected result is superior to those unselected routes he will feel delighted [11]. Regret theory is the alternative methods to study expected utility of risk and uncertain selection, and it is mainly used to compare the options' results and options' attributes in practice.

Here we suppose there are two routes for traveller's choosing, as shown in Figure 1. Moreover, travel time is the only evaluation index. Travel times on both routes are uncertain. For travellers they cannot master the exact value of the both routes, but they can know the status' probability p_s of both routes and the travel times in different status (t_s^1, t_s^2).

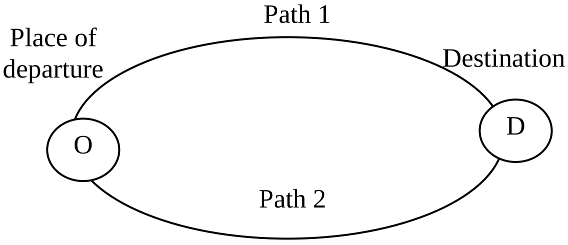


FIGURE 1 Schematic of travel route choice

In expected utility theory, the travellers will choose the route with the largest expected utility (EU). For one route, we can calculate its expected utility:

$$EU = \sum_s \left[p_s \bullet U\left(t_s\right) \right], \tag{1}$$

where $U(t_s)$ is a utility function, and it can be represented in multiple forms. Generally, we can use time and cost as index, using linear function or Exponential function. Following two equation is the frequently-used form.

$$U(t_s) = [1 - \exp(\theta \bullet t_s)] / \theta, \tag{2}$$

$$U(t_s) = \alpha \bullet t_s + \beta \bullet c_s + \varepsilon, \tag{3}$$

where θ is a risk aversion parameter, and α , β are the factors of travel time t_s , travel cost c_s . ε is the dimensionless parameter.

Compared with expected utility theory, regret theory supposes that traveller would feel regret or joyful when he finished choosing. The anticipated feeling can be introduced to the utility function as follow.

$$RU(t_s^1) = U(t_s^1) + \varphi[U(t_s^1) - U(t_s^2)], \tag{4}$$

$$RU(t_s^2) = U(t_s^2) + \varphi[U(t_s^2) - U(t_s^1)], \tag{5}$$

where $RU(*)$ is a regret utility function. When $R(0) = 0$, its mean that traveller will not feel regret or joyful. So according to the performance of regret theory, Chorus [14] described it as follows:

$$RU(*) = 1 - \exp(-\lambda \bullet [*]), \tag{6}$$

where $\lambda \in [0, +\infty)$, and it is regret aversion parameter, which reflect the importance of the variable [*]. When λ increases the regret becomes more and more important than joy, and when λ approaches zero regret utility function becomes expected utility function.

Figure 2 is the relation between time or other [*] index and regret utility with parameter λ. Here λ1<λ2<λ3.

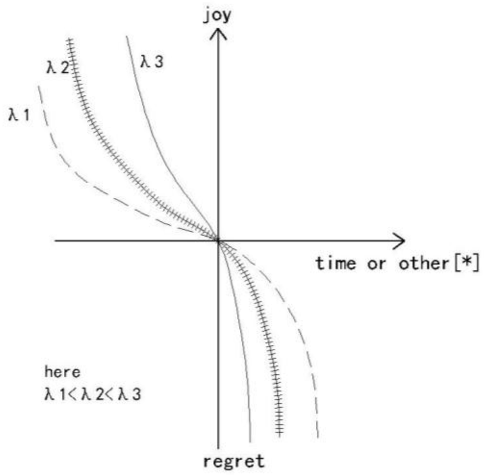


FIGURE 2 The relation between time or other [*] index and regret utility with parameter λ

Regret theory assumes that travellers may have expected utility about each option in regret and joy, and they can add the regret or joy together, and then produce each option's expected regret. In a given status, expected regret is the function of selected option's properties.

$$ERU_1 = \sum_s [p_s \bullet RU(t_s^1)], \tag{7}$$

$$ERU_2 = \sum_s [p_s \bullet RU(t_s^2)]. \tag{8}$$

Travellers will always have an optimum desired arrival time when they go to work from home to companies every day, and too early or to late arriving will suffer some loss. Before travelling them will prediction travel time of each path based on previous travel experience. Here we suppose that the traveller's perception travel time of the two paths T_i^P obey normal distribution, describes as:

$$T_i^P \sim f_i^P(t_i) = N(\hat{t}_i^P, \sigma_i^{P^2}), \tag{9}$$

where, $i = 1,2$. $\hat{t}_i^P$ is the perceived travel time average value of path Y_i, and $\sigma_i^{P^2}$ is the perceived variance.

We suppose the two paths are independent and their travel times are also independent. Traveller will choose path according to minimum regret decision strategy, namely they will choose the path, which has the lowest regret to avoid regret emotions. Chorus gave a regret computational formula of travel route Y_1, as follow:

$$ERU(Y_1) = \int_-^+ \int_-^+ r_1 \bullet [RU(t_2) - RU(t_1)] \bullet f_1^P(t_1) \bullet f_2^P(t_2) dt_1 dt_2 . \tag{10}$$

Here, $RU(t_i)$ is the utility of travel route Y_i. $f_i^P(t_i)$ is the probability density function of perception of travel route Y_i 's time. r_1 is the determining factor, $r_1 = 1$ when $RU(t_2) > RU(t_1)$, otherwise $r_1 = 0$.

Similarly, we can calculate the regret $ERU(Y_2)$ of route Y_2.

Assuming travellers select the path, which has the smallest regret finally, and then the selected path's regret is shown in Equation (11).

$$ERU = \min_{i=1,2}(ERU(Y_i)). \tag{11}$$

2.2 BAYESIAN UPDATING FUNCTION

Before travelling traveller will compare the alternative options according to foregoing travel experience and current traffic information gained, and then choose the best travel route. After travelling travellers will evaluate and modify the regret value according to the actual result of the selected option, and form the updated travel experience for the next travel at the same conditions. Because of the uncertainty of choice condition and the limitation of the traveller's knowledge, they cannot estimate the travel time very accurately. Here we suppose that the perceived travel time is continuous random variable and obeys the normal distribution, after received the newest traffic or travel information the travellers will update the perceived travel time distribution of each alternative option, and we can use Bayesian updating method to simulate the updating process or result.

After obtaining the newest traffic, information travellers will update expected travel time distribution for each route combined with previous travel experience. Because traffic information is time varying, there is a deviation between traveller's perceivable travel time T_i^I and actual travel time T_i. Therefore, we suppose that perceivable travel time obeys normal distribution, namely $T_i^I \sim f_i^I(t_i)$. Standard deviation σ_i^I is the perception about travel time of travellers. According to Bayesian updating theory, the updating travel time T_i^u also obeys normal distribution.

$$T_i^u \sim f_i^u(t_i) = N(\hat{t}_i^u, \sigma_i^{u^2}). \tag{12}$$

Then we can obtain Equation (13) by Bayesian updating inference [16].

$$f_i^u(t_i) = f(t_i^P \mid t_i^I) = \frac{f(t_i^I \mid t_i^P) \bullet f_i^P(t_i)}{f_i^I(t_i)}, \tag{13}$$

$$f_i^I(t_i) = \int_-^+ f(t_i^I \mid t_i^P) \bullet f(t_i^P) dt_i^P . \tag{14}$$

T_i^u 's mean value and variance are Equation (15) and Equation (16) respectively.

$$\hat{t}_i^u = \frac{(1/\sigma_i^P)^2 \bullet \hat{t}_i^P + (1/\sigma_i^I)^2 \bullet t_i^I}{(1/\sigma_i^P)^2 + (1/\sigma_i^I)^2}, \tag{15}$$

$$\sigma_i^u = \sqrt{\frac{\sigma_i^{P^2} \bullet \sigma_i^{I^2}}{\sigma_i^{P^2} + \sigma_i^{I^2}}}. \tag{16}$$

After updating the distribution of the expected travel time, travellers will make decisions again. Now the regret of travel route choice Y_1 can be calculate by Equation (17).

$$ERU^I(Y_1) = \int_-^+ \int_-^+ r_1 \bullet [RU(t_2) - RU(t_1)] \bullet f_1^u(t_1) \bullet f_2^u(t_2) dt_1 dt_2 . \tag{17}$$

In a similar way, we can calculate the regret of Y_2.

Travellers still select the route which has the smallest regret. At the same time considering the probability of received information, the regret of the desired route which traveller selected can be gained through Equation (18).

$$ERU^{I}=\int_{-}^{+}\int_{-}^{+}[\min_{i=1,2}(ERU^{I}(Y_i))]\bullet f_1^{I}(t_1)\bullet f_2^{I}(t_2)dt_1dt_2\ . \quad (18)$$

2.3 MODEL BASED ON REGRET THEORY

After travellers determined all routes' regret, they will carry out route choice according to regret. There are two methods of travellers' route choice.

Firstly, they will choose the route, which has the smallest regret directly without considering the impact of the route factors, and this is named as deterministic selection model.

$$P(r,s,n)=\min\{R_i,R_j,R_k\}\ . \quad (19)$$

Secondly, we can establish discrete choice model with considering the impact of the route factors. The probability of each route can be calculated by Logit model as Equation (20).

$$P(r,s,n)=\exp[-\sigma t(n)/\bar{t}]/\sum_{i=1}^{m}\exp[-\sigma t(i)/\bar{t}]\ , \quad (20)$$

where, $P(r,s,n)$ is the assignment probability which OD T(r,s) in route n. $t(n)$ refers the regret value of the route n, and $\bar{t}$ is the average regret value of all routes. σ is the assignment parameter. m is the number of valid travel routes.

3 Examples

Here we reference the example in the literature 17 and replace part of the data. There are three routes i, j, k between the place of departure and destination for traveller n. So he or she has three travel routes or options, including alt_i, alt_j and alt_k. The property of travel routes includes travel time x (min) and travel cost y (yuan). Here we suppose that the time and the cost are fixed for every route. And $alt_i=\{x_i=75,y_i=1\}$, $alt_j=\{x_j=45,y_j=3\}$, $alt_k=\{x_k=30,y_i=20\}$.

The regret of one route or option equals to the compare between it and other two options, namely:

$$R_i=\max\{R_{ij},R_{ik}\}\ ,$$
$$R_j=\max\{R_{ji},R_{jk}\}\ , \quad (21)$$
$$R_k=\max\{R_{ki},R_{kj}\}\ .$$

Here the regret value of one route is the sum of the inter-comparison of two options' every property. For example, we can define the binomial regret as follow:

$$R_{ij}=\varphi_x(x_i,x_j)+\varphi_y(y_i,y_j)\ . \quad (22)$$

And $\varphi_x(x_i,x_j)$, $\varphi_y(y_i,y_j)$ are the regret function of property. They can be described as follow:

$$\varphi_x(x_i,x_j)=\max\{0,\beta_x(x_j-x_i)\}\ ,$$
$$\varphi_y(y_i,y_j)=\max\{0,\beta_y(y_j-y_i)\}\ . \quad (23)$$

Here β is the parameter of property regret, and it reflects the relative importance between different properties. Refer to actual perceived value, we suppose that the β of travel time is -1.0/min, and the β of travel cost is -0.5/yuan.

Equation (22) and Equation (23) reflect that the regret function of property is the linear function about property difference.

And we can calculate the regret compared of alt_i, alt_j.

$$R_{ij}=\max\{0,-1.0\times(45-75)\}+\max\{0,-0.5\times(3-1)\}=30\ .(24)$$

The regret of traveller is travel time, and not travel cost.

In a similar way, we can gain:

$R_{ik}=\{0,-1.0\times(30-75)\}+\{0,-0.5\times(20-1)\}=45$,

$R_{ji}=1$, $R_{jk}=15$, $R_{ki}=9.5$, $R_{kj}=8.5$.

Thus, it is concluded that: $R_i=45$, $R_j=15$, $R_k=9.5$. With regret theory, model traveller preference structure is $alt_k>alt_j>alt_i$, travel will choose alt_k. This reflects that travel time is the main influence factor. However, in fact some person will choose route according travel cost when they have enough time. Moreover, it is different to expected theory that it cannot be totally compensation or replace between properties. When certain property beyond the ability of traveller's expense, he or she will not choose the option.

Under the framework of utility maximization, the poor performance of the travel time can be replaced by the good performance. Nevertheless, in regret theory framework, people prefer to choose the option without regret in all the properties.

Now we use Equation (20) to analysis the probability of every route according to Logit distribution. Here σ =3.5, and

$P(i)=\exp(-3.5\times 45/23.16)/[\exp(-3.5\times 45/23.16)$
$+\exp(-3.5\times 15/23.16)+\exp(-3.5\times 9.5/23.16)]$
$=0.003$

$P(j)=0.302$, and $P(k)=0.695$.

In order to further study of regret features and applications, we suppose that the time of three routes is varying every day and they are independent of each other. Each has two status, including normal and congestion.

The normal travel times of alt_i , alt_j and alt_k respectively are 55, 40 and 30 min. The congestion travel times respectively are 75, 55 and 60 min. For every route the probability of congestion is the 0.50. Therefore, we can gain 8 status having equal probability. The travel costs of each route respectively are 1, 3 and 20. Here we do not update regret. Table 1 shows the properties of each route in different status.

TABLE 1 Properties of each route in different status

Status	**alt$_i$**		**alt$_j$**		**alt$_k$**	
	x	***y***	***x***	***y***	***x***	***y***
1	55	1	40	3	30	20
2	55	1	40	3	60	20
3	55	1	55	3	30	20
4	55	1	55	3	60	20
5	75	1	40	3	30	20
6	75	1	40	3	60	20
7	75	1	55	3	30	20
8	75	1	55	3	60	20

Table 2 shows the regret value of each route in different status.

TABLE 2 Regret value of each route in different status

status	**R(alt$_i$)**	**R(alt$_j$)**	**R(alt$_k$)**	**choose**
1	25	10	9.5	alt$_k$
2	15	1	28.5	alt$_j$
3	25	25	9.5	alt$_k$
4	0	1	14.5	alt$_i$
5	45	10	9.5	alt$_k$
6	35	1	28.5	alt$_j$
7	45	25	9.5	alt$_k$
8	20	1	13.5	alt$_j$
average	26.25	9.25	15.375	alt$_j$

From table 1 and table 2, we can know that under different status conditions travellers will choose different route based on regret theory. Totally under 8 status travellers choose alt_i 1 times, alt_j 3 times, and alt_k 4 times. However, using average regret value, travellers' choosing order is $alt_j > alt_k > alt_i$. Travellers would like to alt_j . From the comparison of the properties of the routes, we can know route alt_j has a relative balanced index in travel time and travel cost.

If we use expected utility function to analysis the travel route choice, we can gain the result with little significant difference. Based on travel time traveller will choose alt_k , however based on travel cost they would like to choose alt_i . In addition, if the two index has different weight, travellers would have varied choose, among them some may be similar to model based on regret theory.

In travel route choice domain based on regret theory, it is the regret relative to the best options playing an important role in the decision-making, not joy. This shows the character of properties of options in choosing process, namely bad utility of option's property cannot be directly compensated by another good utility of option's property. Travel route choice rules are to choose the option with the smallest regret. When the regret value is greater than a certain threshold, its value is beyond of the ability of traveller's ability to pay, and this will lead to cannot make decisions and they will delay choose through looking for more information available.

Now we use Equation (20) to analysis the probability of every route according to Logit distribution. Here σ =3.5, and

$$P(i) = \exp(-3.5\times 26.25/16.96)/[\exp(-3.5\times 26.25/16.96)$$
$$+\exp(-3.5\times 9.25/16.96)+\exp(-3.5\times 15.38/16.96)]$$
$$=0.023$$

$P(j)=0.762$, and $P(k)=0.215$.

Using regret theory or expected utility theory to measure the relative attraction of options for selection result is different, under the framework of discrete choice within econometrics, attract degree's difference between alternative options means that the predict choice probability will be different.

4 Conclusions

In this paper, firstly compared expected utility theory, prospect theory and regret theory in travel route choice study, and then a model based on regret theory and Bayesian updating method is established. The conclusion shows that when travellers choose travel routes they will try to avoid those unselected routes are better than the selected route. At the same time they think one worse property of route cannot be replaced by another better property. All of these are corresponding with regret minimization of regret theory. In addition, travel route choice model based on regret theory is relatively simple.

When each property of one option is the same or better than the other option's each property, the regret minimization model with generic selecting aggregation and multiple attribute decision making can be simplified as the utility maximization model.

Since the travel route choice problem involves not only the establishment of the utility function, it also involves the travellers' information updates mechanism and the travellers' choice characteristics; we should carry further study through actual travel route choice data.

References

[1] Manski C F 1977 *Theory and Decision* **8**(3) 229-54

[2] Ettema D, Timmermans H 2006 *Transportation Research Part C* **14**(5) 335-50

[3] Kahneman D, Tversky A 1979 *Econometrica* **47**(2) 263-91

[4] Tversky A, Kahneman D 1992 *Journal of Risk and Uncertainty* **5**(4) 297-323

[5] Wang Yan, Zhang Li 2011 *Journal of Xihua University - Natural Science* **30**(2) 12-16 *(in Chinese)*

[6] Zhang Yang, Jia Jian-min, Huang Qing 2007 *Journal of Management Sciences in China* **22**(5) 78-85 *(in Chinese)*

[7] Luo Qing-Yu, Yang Yin-Sheng, Sun Bao-Feng 2011 *Proc. 2011 Int. Con. on Transportation, Mechanical, and Electrical Engineering* 1822-25

[8] Jou R C, Kitamura R, Weng M C,d Chen C C 2008 *Transportation Research Part A* **42**(5) 774-83

[9] Schwanen T, Ettema D F 2009 *Transportation Research Part A* **43**(5) 511-25

[10] De Moraes R G, Daamen W, Hoogendoorn S 2013 *Journal of Choice Modelling* **6** 17-33

[11] Loomes G, Sugden R 1982 *The Economic Journal* **92**(368) 805-824

[12] Bell D E 1982 *Operations Research* **30**(5) 961-81

[13] Chorus C G, Arentze T A, Timmermans H J P 2008 *Transportation Research Part B* **42**(1) 1-18

[14] Chorus C G 2012 *Transportmetrica* **8**(4) 291-305

[15] Giselle de M R, Winnie D, Serge H 2011 *Transportation Research Record* **n2230** 19-28

[16] Yan Zhen-zhen, Liu Kai, Wang Xiao-guang 2013 *Journal of Transportation Systems Engineering and Information Technology* **13**(4) 76-83 *(in Chinese)*

[17] Luan kun, Juan Zhicai, Ni Anning 2012 *Journal of Transport Information and Safety* **30**(6) 77-80 *(in Chinese)*

Authors

Baohui Jin, born on August 30, 1975, Liaoning, China

Current position, grades: doctoral candidate and senior engineer
University studies: Transportation Planning & Management
Scientific interest: travel behaviour research, transportation planning
Publications: 4 articles
Experience: Jin Baohui is a doctoral candidate of College of Transportation & Logistics, Southwest Jiaotong University, Chengdu, China. And his specialized subject is transportation planning & management. He worked at the Planning and Design Institute of Chengdu, and his main research area includes urban planning and transportation planning. In 2004, he joined the team, which developed the railway demand prediction software. He finished many urban traffic planning and traffic impact evaluation. His current research interests are travel behaviour theories and models.

Optimal contracts of production personnel's innovation based on slack resources

Heping Zhong[*]

School of Management, Xuchang University, Xuchang, 461000, Henan, China

Received 1 March 2014, www.tsi.lv

Abstract

Based on the analysis frames of the multi-task principal-agent model, this paper establishes a principal-agent model of production personnel's innovation based on slack resources and obtains the optimal incentive contracts for production personnel while they are engaged in "production task" and "slack innovation" through the analysis of the model. In order to improve the performance of production personnel's "slack innovation", on one hand, the firm can reward their "slack innovation" according to the optimal incentive contracts; on the other hand, the firm can optimize the incentive contracts for their "production task" according to the interdependence of the cost functions of "production task" and "slack innovation" to promote indirectly the performance of "slack innovation". The originality of this paper is not only examining the multi-task problems of the compensation incentives for production personnel's "slack innovation" but also considering the impacts of the firm's active actions to support the production personnel's "slack innovation" on incentive contracts.

Keywords: contract, incentive, optimization, multi-task agent model, innovation, slack resources, production personnel

1 Introduction

Production personnel are the production line staffs who are mainly engaged in relatively simple and procedural work such as product processing, the maintenance of equipment, repetitive daily management work and so on. They complete the operation of the specific technics process and produce specific products by use of professional skills, knowledge and experience in their specific jobs. They have a wealth of operational knowledge, experience and high On-site skills on product technics, quality, cost control and equipment capacity. Production personnel's position characteristics are practical and operational. They overcome difficulties to realize the drawings and programs what R&D personnel design.

Production personnel produce qualified products by virtue of their production skills, at the same time; they also participate in the research and development and process improvement by virtue of their skills. Because of their substantial experience, they have the potential of innovation and process innovation advantage, and play an irreplaceable role at least in the key links such as the product design, production and processing of technological innovation. Therefore, production personnel are the key strength in the success of the product innovation and process innovation.

Although production personnel play an important role in technological innovation activities, the existing literatures mainly focus on R&D personnel's innovation incentives (e.g. Wang, 2008; Pan & Wan, 2010; Zhong, 2012), which is lack of research on the production personnel's innovation incentives. Some scholars research the incentive problems for skilled workers combing with China's current situation lacking of skilled personnel (e.g. Pan, 2011), but these studies are mainly the macro-policy researches, lacking of the research on incentives for the production personnel's firm practice, especially lacking of the research on incentives for the production personnel's technological innovation.

Theory and practice show that production personnel's technology innovation activities are the fundamental means to improve the skills of production personnel. The innovation of a modern firm focuses on all-involvement innovation in a firm, and everyone is the source of innovation in a firm. The firm only improves the sense, capacity and efficiency of innovation of all staff in order to successfully innovate, improve innovation performance and achieve the best operation results (Xie et al., 2005). Therefore, researching incentives for production personnel's technological innovation and improving their innovation performance have great significance to achieve all-involvement innovation and improve the competitiveness of the technological innovation of a firm.

Based on the above, this paper constructs an incentive contract model of production personnel's innovation based on slack resources through applying multi-task principal-agent theory in the base of comprehensive drawing on previous researches. The novelty of this model is not only examining the multi-task problems of the compensation incentives for production personnel's

[*] ***Corresponding author*** e-mail: Hepingz2004@163.com

technological innovation based on slack resources but also considering the impacts of the principal's active actions to support the agent's innovation based on slack resources on incentive contracts for the agent.

2 Analysis of production personnel's innovation behaviour

Production personnel bear the heavy production and processing tasks. In continuous repetitive production process, production personnel master the more knowledge of production equipment, materials, products, processes etc. and have a wealth of operational experience, so as to continuously improve the production and processing skills.

With the enhancement of production personnel's knowledge and skills, production personnel have the in-depth understanding of the performance and use of production equipment, and gradually grasp the knowledge and skills on saving materials, improving operation methods and processes, enhancing work efficiency and product quality. Thus, in the production process, they hold some slack resources such as the improvement of device performance and versatility, saving materials (raw materials, semi-finished products, and finished products), the secret of high-quality and high-efficiency operation methods, which provide a resource base for the next technological innovation activities. Driven by the innovative driving force imposed by the managers, production personnel will make use of slack resources at their disposal to engage in technological innovation activities if they have affluent time and effort after the completion of the formal production tasks. These innovative activities may be initiated by themselves or may also be pushed into the innovative team of other innovation personnel.

Production personnel's full-time work is to complete the production tasks. Different specialization makes the firm difficult to allocate appropriate resources for the production personnel's technological innovation activities. Even if production personnel's innovative projects have a larger value, the firm only configures rarely some key resources for their innovative projects, accordingly, production personnel can only use their own slack resources such as various real-time information, expertise knowledge, and a variety of personal relationships and relatively idle resources in the firm for technological innovation. Therefore, the production personnel's technology innovation activities are mainly technology innovation based on slack resources at the same time as production personnel complete production tasks.

Production personnel's technological innovation activities based on slack resources are mainly showed in the following areas:

(1) Improvement of the existing process methods. R&D personnel's knowledge about equipment and technics is limited, but production personnel have accumulated more equipment knowledge and skills in the repetitive operation of equipment, making it very easy to find product problems in the process and suggest improvement program. The "unique skill" and "unique technique" they grope in the production process greatly improved work efficiency and product quality.

(2) Design of new tools or improvements in equipment. Innovation comes from practice; the practice of the front-line production personnel is the source of innovation. In the long-term production process, production personnel will find the deficiencies or defects in the performance of existing equipment and some inefficient process methods. Accordingly, they will produce creative ideas of designing new tools or improving equipment performance, and use slack resources to test these ideas until new tools and improved equipment to improve the work efficiency are designed.

(3) Improvement of product. Any product needs to constantly be improved and perfected. In the production process, production personnel will naturally reflect product problems. With the increasing of product knowledge, they will find the problems of product design, especially the matching problems between products design and process technology, thereby, innovative proposals or programs to improve products may be made.

Obviously, production personnel's innovation activities based on slack resources are very useful for the development of a firm, which not only reduces the risk and cost of the innovation and improves the innovation performance, but also improves the efficiency of resources, especially what is important is that it trains a large number of high-skilled talents.

3 Incentive contracts

3.1 MODEL ASSUMPTION

Suppose the firm as a principal, production personnel as an agent. Then, on one hand, the agent is engaged in the formal production tasks arranged by the principal (for short "production task"); on the other hand, he is engaged in the technological innovation activities by use of slack resources (for short "slack innovation"). Thus, the model can be regarded as a multi-task agent model (Holmstrom & Milgrom, 1991; Zhang, 2004; Zhong, 2012). In order to make the analysis easier, we may make the following assumptions:

Assumption 1: Denote $a_i, i=1,2$ the level of the agent's effort, a_1 is the level of effort spent on the "production task" action, a_2 is the level of effort spent on the action of "slack innovation"; denote a_3 the level of the principal's effort spent on the active action of supporting the agent's "slack innovation" (for short "support slack"). With $C(a_3)$ denoting the cost of the principal's effort a_3, we assume $C(a_3)=\frac{b_3}{2}a_3^2$, b_3 is the cost coefficient of the principal's effort a_3, $b_3 > 0$ (Zhang, 2004); with $B(a_1, a_2, a_3)$ denoting the expected gross benefits of the technological innovation activities, in order to conveniently analyse the question, we may assume $B(a_1, a_2, a_3) = f_1 a_1 + f_2 a_2 + f_3 a_2 a_3$ (Bernardo, 2001; Gibbons,2005; Zhong, 2012), where f_1 is the benefit coefficient of the output of the agent's effort a_1, $f_1 > 0$; f_2 is the benefit coefficient of the output of the agent's effort a_2, $f_2 > 0$; f_3 is the benefit coefficient of the output of the principal's effort a_3, $f_3 > 0$. This assumption is reasonable. Obviously, if the principal takes the active action to support the agent's "slack innovation", such as to enable the firm to maintain much slack resources, to actively promote the production personnel to coordinate with other departments and to actively encourage the production personnel translate slack resources into innovation output, then, the production personnel can easily possess some slack resources and access to the slack resources required for innovation, and is easier to use slack resources to produce the desired innovation output, that is, the principal's "support slack" action contribute greater to the agent's "slack innovation" action. Therefore, the principal's a_3 has an active influence on the agent's a_2, and thus generate positive impact on innovation output, that is $f_3 > 0$, and, the higher the principal's effort a_3 is, the greater the contribution to the innovation is.

Assumption 2: With $C(a_1, a_2)$ denoting the cost of the agent's effort, we assume that the function $C(a_1, a_2)$ is strictly convex. In order to conveniently analyse the question, we may assume $C(a_1, a_2) = \frac{b_1}{2} a_1^2 + b_{12} a_1 a_2 + \frac{b_2}{2} a_2^2$, $b_1 > 0, b_2 > 0$ (Zhong, 2012), where b_1 is the cost coefficient of the effort spent on the "production task" action, marginal cost change rate $\frac{\partial^2 C}{\partial a_1^2} = b_1 > 0$; b_2 is the cost coefficient of the effort spent on the "slack innovation" action, marginal cost change rate $\frac{\partial^2 C}{\partial a_2^2} = b_2 > 0$; b_{12} is the interdependent cost coefficient of "production task" and "slack innovation", $\frac{\partial^2 C}{\partial a_1 \partial a_2} = \frac{\partial^2 C}{\partial a_2 \partial a_1} = b_{12}$, when $b_{12} < 0$, the cost functions of the efforts of the two actions are complementary; when $b_{12} = 0$, the cost functions of the efforts of the two actions are independent; when $b_{12} > 0$, the cost functions of the efforts of the two actions are substitute (Holmstrom & Milgrom, 1991; Laffont & Martimort, 2002).

Assumption 3: The principal cannot observe the level of the agent's effort, but the result of the agent's effort $x_i, i = 1,2$ can be observed, suppose $X = x_1 + x_2$, $x_i = g_i a_i + \varepsilon_i$, where g_i is the output of per unit of effort in every action, $g_i > 0$, ε_i is exogenous random variable which is normally distributed with mean vector zero and variance σ_i^2 (on behalf of all the factors that the agent cannot control), ε_1 and ε_2 are independent (Gibbons, 2005).

Assumption 4: The principal takes the linear function for the **c**ompensation rule to pay the agent, namely, the compensation rule is $\mathrm{w(X)} = \mathrm{w}_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2$, where w_0 is the fixed income of the agent, β_1, β_2 is the incentive factor of x_1, x_2 respectively, $\beta_1 \geq 0, \beta_2 \geq 0$.

Assumption 5: The principal is risk-neutral, the agent is risk aversion. And further assume that the agent has the utility function of unchanged absolute risk aversion, his preferences are represented by a negative exponential utility function $u(w) = -e^{-\rho w}$, where ρ measures the agent's absolute risk aversion, w is his compensation minus personal cost. Denote CE the agent's "certainty equivalent" money payoff, then CE meets $u(CE) = Eu(w)$, so one could utilize the exponential form to deduce that the agent's certainty equivalent is $CE = \mathrm{w}_0 + \beta_1 \mathrm{g}_1 a_1 + \beta_2 \mathrm{g}_2 a_2 - \frac{1}{2} \rho(\beta_1^2 \sigma_1^2 + \beta_2^2 \sigma_2^2) - \frac{b_1}{2} a_1^2 - b_{12} a_1 a_2 - \frac{b_2}{2} a_2^2$,

where $\mathrm{w}_0 + \beta_1 \mathrm{g}_1 a_1 + \beta_2 \mathrm{g}_2 a_2$ is the expected payoff of the agent, $\frac{1}{2} \rho(\beta_1^2 \sigma_1^2 + \beta_2^2 \sigma_2^2)$ is the risk premium of the agent, $\frac{b_1}{2} a_1^2 + b_{12} a_1 a_2 + \frac{b_2}{2} a_2^2$ is the cost of the agent's effort (Zhang, 2004).

3.2 BASIC MODEL

The principal's expected profit is:

$$Y_P = f_1 a_1 + \mathrm{f}_2 a_2 + \mathrm{f}_3 a_2 a_3 - \mathrm{w}_0 - \beta_1 \mathrm{g}_1 a_1 - \beta_2 \mathrm{g}_2 a_2 - \frac{b_3}{2} a_3^2$$

Because the principal is risk-neutral, therefore, the principal's expected profit is his certainty equivalent. Denote $\overline{w}$ the agent's reservation wage, so the constraint of the agent's participation is: $CE = \mathrm{w}_0 + \beta_1 \mathrm{g}_1 a_1 + \beta_2 \mathrm{g}_2 a_2 - \frac{1}{2} \rho(\beta_1^2 \sigma_1^2 + \beta_2^2 \sigma_2^2) - \frac{b_1}{2} a_1^2 - b_{12} a_1 a_2 - \frac{b_2}{2} a_2^2 \geq \overline{w}$ the agent's incentive compatibility constraint is: $(a_1, a_2) \in \arg\max CE$. It is equivalent to $(a_1, a_2) \in \arg\max(\beta_1 \mathrm{g}_1 a_1 + \beta_2 \mathrm{g}_2 a_2 - \frac{b_1}{2} a_1^2 - b_{12} a_1 a_2 - \frac{b_2}{2} a_2^2)$.

The principal-agent model is:

$$\max Y_P = f_1 a_1 + \mathrm{f}_2 a_2 + \mathrm{f}_3 a_2 a_3 - \mathrm{w}_0 - \beta_1 \mathrm{g}_1 a_1 - \beta_2 \mathrm{g}_2 a_2 - \frac{b_3}{2} a_3^2, (1)$$

s.t.

$$\mathrm{w}_0 + \beta_1 \mathrm{g}_1 a_1 + \beta_2 \mathrm{g}_2 a_2 - \frac{1}{2}\rho(\beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2) - \frac{b_1}{2}a_1^2, \quad (2)$$
$$- b_{12}a_1a_2 - \frac{b_2}{2}a_2^2 \geq \overline{w}$$

$$(a_1, a_2) \in \arg\max(\beta_1 \mathrm{g}_1 a_1 + \beta_2 \mathrm{g}_2 a_2 - \frac{b_1}{2}a_1^2 - b_{12}a_1a_2 - \frac{b_2}{2}a_2^2) \cdot (3)$$

3.3 SOLVING MODEL

By (3), the incentive compatibility constraint becomes

$$a_1 = \frac{\beta_1 g_1 b_2 - \beta_2 g_2 b_{12}}{b_1 b_2 - b_{12}^2}, \quad (4)$$

$$a_2 = \frac{\beta_2 g_2 b_1 - \beta_1 g_1 b_{12}}{b_1 b_2 - b_{12}^2}. \quad (5)$$

By (4) and (5),

$$\frac{\partial a_1}{\partial \beta_1} = \frac{g_1 b_2}{b_1 b_2 - b_{12}^2} \qquad \frac{\partial a_1}{\partial \beta_2} = \frac{-g_2 b_{12}}{b_1 b_2 - b_{12}^2}$$
$$\frac{\partial a_2}{\partial \beta_1} = \frac{-g_1 b_{12}}{b_1 b_2 - b_{12}^2} \qquad \frac{\partial a_2}{\partial \beta_2} = \frac{g_2 b_1}{b_1 b_2 - b_{12}^2}$$

Obviously, from the practical significance, $\frac{\partial a_1}{\partial \beta_1} > 0$, thus $b_1 b_2 - b_{12}^2 > 0$. (6)

By (1), (2), (4), (5), the optimal incentive factors are:

$$\beta_1 = \frac{f_1 g_1 g_2^2 (b_3 Q - f_3^2 b_1) + (f_1 b_2 - f_2 b_{12}) g_1 \rho \sigma_2^2 b_3 Q}{M}, \quad (7)$$

$$\beta_2 = \frac{g_2 g_1^2 (f_2 b_3 Q - f_1 f_3^2 b_{12}) + (f_2 b_1 - f_1 b_{12}) g_2 \rho \sigma_1^2 b_3 Q}{M}, \quad (8)$$

where

$$M = g_1^2 g_2^2 (b_3 Q - f_3^2 b_1) + \rho \sigma_1^2 g_2^2 (b_1 b_3 Q - f_3^2 b_1^2)$$
$$+ \rho \sigma_2^2 g_1^2 (b_2 b_3 Q - f_3^2 b_{12}^2) + \rho^2 \sigma_1^2 \sigma_2^2 b_3 Q^2,$$
$$Q = b_1 b_2 - b_{12}^2$$

Obviously, from the practical significance, β_2 is an increasing function of f_2, namely, $\frac{\partial \beta_2}{\partial f_2} > 0$, thus $M > 0$.

The optimal condition of the level of effort of principal's "support slack" is:

$$a_3 = \frac{f_3(\beta_2 g_2 b_1 - \beta_1 g_1 b_{12})}{b_3(b_1 b_2 - b_{12}^2)}. \quad (9)$$

3.4 MODEL ANALYSIS AND DISCUSSION

Assumption 6: Production personnel are engaged in "production task" and "slack innovation" at the same time. Production personnel's "production task" action can be observed directly.

This assumption is reasonable in most cases for production personnel who are engaged in "production task" and "slack innovation" at the same time. Because production personnel's jobs are procedural and are easy to observe, his "production task" action can be observed directly.

Proposition 1: If the cost functions of "production task" and "slack innovation" of production personnel are independent $(b_{12} = 0)$, then, the optimal incentive contracts of the every action under the conditions of the incentive compatibility are independent each other, and the optimal incentive factors for "production task" are unaffected by the principal's "support slack" action. If the cost functions of "production task" and "slack innovation" of production personnel are interdependent and production personnel's "production task" action can be measurable directly, then, under the conditions of the incentive compatibility, on one hand, the firm can directly award his "slack innovation" according to the optimal incentive contracts to improve the performance of "slack innovation"; on the other hand, the firm can optimize the incentive for his "production task" to promote indirectly the performance of "slack innovation", and when the cost function of these two actions are complementary ($b_{12} < 0$), the firm should strengthen the incentive for the "production task", and the strengthening degree is increased as the complementary degree of the cost function of these two actions and the marginal value of the principal's "support slack" contribution to the agent's "slack innovation" increases; when the cost function of these two actions are substitute, the firm should weaken the incentive for the "production task", and the weakening degree is increased as the substitute degree of the cost function of these two actions and the marginal value of the principal's "support slack" contribution to the agent's "slack innovation" increases.

Proof: From assumption 6, because the production personnel's "production task" and "slack innovation" are not related, the cost functions of these two actions are independent, that is $b_{12} = 0$, by (7) and (8):

$$\beta_1 = \frac{f_1 g_1}{g_1^2 + b_1 \rho \sigma_1^2}, \quad (10)$$

$$\beta_2 = \frac{f_2 g_2 b_2 b_3}{g_2^2 b_2 b_3 - f_3^2 g_2^2 + b_2 b_3 \rho \sigma_2^2}. \quad (11)$$

Obviously, the optimal incentive factor β_1 and β_2 are independent. By formula (10), β_1 and f_3 are not related, namely, the optimal incentive factors for “production task” are unaffected by the principal’s “support slack” action.

If production personnel’s “production task” action can be observed directly, then $\sigma_1 = 0$, by (7) and (8):

$$\beta_1 = \frac{\begin{array}{c} f_1 g_1 g_2^2 [b_3(b_1 b_2 - b_{12}^2) - f_3^2 b_1] + \\ (f_1 b_2 - f_2 b_{12}) g_1 b_3 (b_1 b_2 - b_{12}^2) \rho \sigma_2^2 \end{array}}{\begin{array}{c} g_1^2 g_2^2 [b_3(b_1 b_2 - b_{12}^2) - f_3^2 b_1] + \\ \rho \sigma_2^2 g_1^2 [b_2 b_3 (b_1 b_2 - b_{12}^2) - f_3^2 b_{12}^2] \end{array}}, \tag{12}$$

$$\beta_2 = \frac{g_2 g_1^2 [f_2 b_3 (b_1 b_2 - b_{12}^2) - f_1 f_3^2 b_{12}]}{\begin{array}{c} g_1^2 g_2^2 [b_3(b_1 b_2 - b_{12}^2) - f_3^2 b_1] + \\ \rho \sigma_2^2 g_1^2 [b_2 b_3 (b_1 b_2 - b_{12}^2) - f_3^2 b_{12}^2] \end{array}}. \tag{13}$$

Obviously, from the practical significance, β_2 is an increasing function of f_2, namely, $\frac{\partial \beta_2}{\partial f_2} > 0$, Accordingly, $\frac{\partial \beta_2}{\partial f_2} = \frac{g_2 g_1^2 b_3 (b_1 b_2 - b_{12}^2)}{\begin{array}{c} g_1^2 g_2^2 [b_3(b_1 b_2 - b_{12}^2) - f_3^2 b_1] + \\ \rho \sigma_2^2 g_1^2 [b_2 b_3 (b_1 b_2 - b_{12}^2) - f_3^2 b_{12}^2] \end{array}} > 0$.

By (6):

$$\begin{array}{l} g_1^2 g_2^2 [b_3(b_1 b_2 - b_{12}^2) - f_3^2 b_1] + \\ \rho \sigma_2^2 g_1^2 [b_2 b_3 (b_1 b_2 - b_{12}^2) - f_3^2 b_{12}^2] > 0 \end{array}, \tag{14}$$

$$\begin{aligned} \delta_b &= \beta_1 \big|_{b_{12} \neq 0} - \beta_1 \big|_{b_{12} = 0} \\ &= \frac{\begin{array}{c} f_1 g_1 g_2^2 [b_3(b_1 b_2 - b_{12}^2) - f_3^2 b_1] + \\ (f_1 b_2 - f_2 b_{12}) g_1 b_3 (b_1 b_2 - b_{12}^2) \rho \sigma_2^2 \end{array}}{\begin{array}{c} g_1^2 g_2^2 [b_3(b_1 b_2 - b_{12}^2) - f_3^2 b_1] + \\ \rho \sigma_2^2 g_1^2 [b_2 b_3 (b_1 b_2 - b_{12}^2) - f_3^2 b_{12}^2] \end{array}} - \frac{f_1}{g_1} \\ &= \frac{-b_{12} \rho \sigma_2^2 g_1^2 [f_2 b_3 (b_1 b_2 - b_{12}^2) - f_1 f_3^2 b_{12}]}{\begin{array}{c} g_1 \{ g_1^2 g_2^2 [b_3(b_1 b_2 - b_{12}^2) - f_3^2 b_1] + \\ \rho \sigma_2^2 g_1^2 [b_2 b_3 (b_1 b_2 - b_{12}^2) - f_3^2 b_{12}^2] \} \end{array}} \end{aligned}$$

By (13): $\frac{g_1^2 [f_2 b_3 (b_1 b_2 - b_{12}^2) - f_1 f_3^2 b_{12}]}{\begin{array}{c} g_1^2 g_2^2 [b_3(b_1 b_2 - b_{12}^2) - f_3^2 b_1] + \\ \rho \sigma_2^2 g_1^2 [b_2 b_3 (b_1 b_2 - b_{12}^2) - f_3^2 b_{12}^2] \end{array}} = \frac{\beta_2}{g_2}$.

Accordingly,

$$\delta_b = \beta_1 \big|_{b_{12} \neq 0} - \beta_1 \big|_{b_{12} = 0} = \frac{-b_{12} \rho \sigma_2^2 \beta_2}{g_1 g_2}. \tag{15}$$

By (13):

$$\frac{\partial \beta_2}{\partial f_3} = \frac{\begin{array}{c} 2 f_3 g_2 g_1^4 b_3 (b_1 b_2 - b_{12}^2) [g_2^2 (f_2 b_1 - f_1 b_{12}) \\ - \rho \sigma_2^2 b_{12} (f_1 b_2 - f_2 b_{12})] \end{array}}{\begin{array}{c} \{ g_1^2 g_2^2 [b_3 (b_1 b_2 - b_{12}^2) - f_3^2 b_1] + \\ \rho \sigma_2^2 g_1^2 [b_2 b_3 (b_1 b_2 - b_{12}^2) - f_3^2 b_{12}^2] \}^2 \end{array}}. \tag{16}$$

When $b_{12} < 0$,

By (15):

$\delta_b > 0$, δ_b increases as $|b_{12}|$ increases and increases as β_2 increases.

Thus $\beta_1 \big|_{b_{12} < 0} > \beta_1 \big|_{b_{12} = 0}$, the firm should strengthen the incentive for β_1, and β_1 increases as $|b_{12}|$ increases and increases as β_2 increases.

By (16) and (6): $\frac{\partial \beta_2}{\partial f_3} > 0$, β_2 increases as f_3 increases.

Because β_1 increases as β_2 increases and β_2 increases as f_3 increases, β_1 increases as f_3 increases.

Accordingly, when $b_{12} < 0$, the firm should strengthen the incentive for β_1, and β_1 increases as $|b_{12}|$ increases and increases as f_3 increases.

By (5) and (6): $a_2 = \frac{\beta_2 g_2 b_1 + \beta_1 g_1 |b_{12}|}{b_1 b_2 - b_{12}^2}$, a_2 increases as β_1 increases.

By assumption 3, x_2 increases as a_2 increases. Thus, x_2 increases as β_1 increases, namely, to strengthen the incentive for the “production task” can improve the performance of “slack innovation”.

Therefore, when the cost functions of the two actions are complementary, the principal should strengthen the incentive for the “production task”, and the strengthening degree is increased as the complementary degree of the cost function of these two actions and the marginal value of the principal’s “support slack” contribution to the agent’s “slack innovation” increases.

When the cost functions of the two actions are complementary, the harder one action works, the lower the marginal cost of the other action (Holmstrom & Milgrom, 1991). The principal strengthens the incentive for the agent’s “production task” will prevail the agent on “production task” to work harder, but it has lowered the marginal cost of “slack innovation” action, thereby reduces the risk of “slack innovation” and thus can improve the performance of “slack innovation”. In the same way, if the marginal value of the principal’s “support slack” contribution to the agent’s “slack innovation” is greater, and the principal’s effort of “support slack” is higher, the production personnel can easily maintain some slack resources and access to the slack resources required for innovation activities, the more slack resources the agent controls, the greater enthusiasm for innovation the agent has, and the more efforts the agent takes (Wang & Pu, 2005). On one hand, it promotes the performance of "slack innovation", on the other hand, it reduces the marginal cost of "production task" and thus improve the performance of "production task".

When $b_{12} > 0$,

By (15): $\delta_b < 0$, δ_b decreases as $|b_{12}|$ increases and decreases as β_2 increases.

Thus $\beta_1\big|_{b_{12}>0} < \beta_1\big|_{b_{12}=0}$, the firm should weaken the incentive for β_1, and β_1 decreases as $|b_{12}|$ increases and decreases as β_2 increases.

By (16) and (6):

If $g_2^2(f_2b_1 - f_1b_{12}) > b_{12}(f_1b_2 - f_2b_{12})\rho\sigma_2^2$, then $\frac{\partial\beta_2}{\partial f_3} > 0$, β_2 increases as f_3 increases.

Because β_1 decreases as β_2 increases and β_2 increases as f_3 increases, β_1 decreases as f_3 increases.

Accordingly, when $b_{12} > 0$, if $g_2^2(f_2b_1 - f_1b_{12}) > b_{12}(f_1b_2 - f_2b_{12})\rho\sigma_2^2$, the firm should weaken the incentive for β_1, and β_1 decreases as $|b_{12}|$ increases and decreases as f_3 increases, namely, the weakening degree for β_1 increases as $|b_{12}|$ and f_3 increases.

If $g_2^2(f_2b_1 - f_1b_{12}) < b_{12}(f_1b_2 - f_2b_{12})\rho\sigma_2^2$, then $\frac{\partial\beta_2}{\partial f_3} < 0$, β_2 decreases as f_3 increases.

Because β_1 decreases as β_2 increases and β_2 decreases as f_3 increases, β_1 increases as f_3 increases.

Accordingly, when $b_{12} > 0$, if $g_2^2(f_2b_1 - f_1b_{12}) < b_{12}(f_1b_2 - f_2b_{12})\rho\sigma_2^2$, the firm should weaken the incentive for β_1, and β_1 decreases as $|b_{12}|$ increases and increases as f_3 increases, namely, the weakening degree for β_1 increases as $|b_{12}|$ increases and decreases as f_3 increases.

Generally speaking, the marginal value of the production personnel's innovation results is much higher than the "production task". Production personnel's "slack innovation" is mainly "five small" activities of "innovation and improvement of performance" combined with "production task", so these two actions are complementary.

If the production personnel is busy in his production tasks, the cost function of these two actions are substitute because of time and effort limitations. However, production personnel is not very busy in practice, the substitution of these actions is very little. In an addition, the agent's risk aversion and profit-driven decide that the risk of production personnel's "slack innovation" is smaller and the marginal cost change rate of production personnel's "slack innovation" is not more than that of his "production task". Thereby, in most cases, $g_2^2(f_2b_1 - f_1b_{12}) > b_{12}(f_1b_2 - f_2b_{12})\rho\sigma_2^2$, that is, generally, the weakening degree for β_1 increases as $|b_{12}|$ and f_3 increases.

By (5) and (6): $a_2 = \frac{\beta_2 g_2 b_1 + \beta_1 g_1 |b_{12}|}{b_1 b_2 - b_{12}^2}$, a_2 increases as β_1 decreases.

By assumption 3, x_2 increases as a_2 increases. Thus, x_2 increases as β_1 decreases, namely, to weaken the incentive for the "production task" can improve the performance of "slack innovation".

Therefore, when the cost functions of the two actions are substitute, the principal should weaken the incentive for the "production task", and the weakening degree is increased as the substitute degree of the cost function of these two actions and the marginal value of the principal's "support slack" contribution to the agent's "slack innovation" increases.

When the cost functions of the two actions are substituting, the harder one action works, the higher the marginal cost of the other action is (Holmstrom & Milgrom, 1991). The principal weakens the incentive for the agent's "production task" will prevail the agent on "slack innovation" with more energy, thus can improve the performance of "slack innovation". In the same way, if the marginal value of the principal's "support slack" contribution to the agent's "slack innovation" is greater, and the principal's effort of "support slack" is higher, the production personnel can easily maintain some slack resources and access to the slack resources required for innovation activities, the more slack resources the agent controls, the greater enthusiasm for innovation the agent has, and the more efforts the agent takes (Wang, Pu, 2005). On one hand, it promotes the performance of "slack innovation", on the other hand, it increases the marginal cost of "production task". Because the agent is risk-aversion, the effect of strengthening the incentive for the "production task" is not so much at this time that the firm should weaken the incentive for the "production task" to save costs.

Assumption 7: When production tasks of the production personnel are full, the cost functions of "production task" and "slack innovation" of production personnel are substitute, that is $b_{12} > 0$; when production tasks of the production personnel are not full, the cost functions of these two actions are complementary, that is $b_{12} < 0$.

This assumption is reasonable. If the agent's cost with the two tasks at the same time is more than the total of the cost the agent is engaged in one task respectively, the cost functions of these two tasks are substitute. In other words, after a task is executed, another task will be harder to implement. If the agent's cost with the two tasks at the same time is less than the total of the cost the agent is engaged in one task respectively, the cost functions of the two tasks are complementary. In other words, when a task is executed, another task will be easier to implement (Laffont & Martimort, 2002). In the case that a person's energy is certain, the more efforts he spends on one work, the higher the marginal cost of another work is (Holmstrom & Milgrom, 1991). Production personnel's full "production task" take up his major energy, then production personnel's slack innovation activities would be reduced, that is, when a task is executed, another task will be harder to implement, so the cost functions of the two actions are substitute, that is $b_{12} > 0$ (Holmstrom & Milgrom, 1991); but when "production tasks" of the production personnel are not full, production personnel have wealthy effort for "slack innovation" activities, generally speaking, production personnel's "slack innovation" activities and the expertise knowledge and the technology of the "production task" are related, in this way, the more in-depth production personnel's understanding about "production task" is, the more easily the innovations about "production task" generate, in the same way, production personnel's "slack innovation" activities around the "production task" are beneficial to improve the performance of "production task", in other words, when a task is executed, another task will be easier to implement, the cost functions of these two actions are complementary, that is $b_{12} < 0$ (Holmstrom & Milgrom,1991).

Proposition 2: When production personnel's production tasks are full, the firm should weaken the incentive for his "production task" to induce him to do some "slack innovation" activities; when production personnel's production tasks are not full, the firm should strengthen the incentive for his "production task" to encourage him to do some "slack innovation" activities.

Proof: by assumption 7, if production personnel's production tasks are full, the cost functions of "production task" and "slack innovation" of production personnel are substitute, that is $b_{12} > 0$; if production personnel's production tasks are not full, the cost functions of these two actions are complementary, that is $b_{12} < 0$. Therefore, by Proposition 1, when production personnel's production tasks are full, the firm should weaken the incentive for his "production task" to induce him to do some "slack innovation" activities; when production personnel's production tasks are not full, the firm should strengthen the incentive for his "production tasks" to encourage him to do some "slack innovation" activities.

This also meets with the fact. When production personnel's production tasks are full, the firm often give relatively low piecework wage or hourly wage; in this way, some skilled workers who have high skills and strong innovation abilities will take advantage of the slack resources mastered by themselves for innovation activities, and find ways to improve the work efficiency of the "production task" through improved technic to increase the personal income. When production personnel's production tasks are not full, the firm should strengthen the incentive for their "production task" to encourage them find ways to improve the performance of "production task". Obviously, only through innovation activities around the "production task", the firm can improve work efficiency and quality of the "production task", thus contributing to the production personnel's "slack innovation" activities.

4 Conclusions

The paper uses multi-task principal-agent model to research the coordination incentive problems for production personnel to be engaged in daily "production task" and "slack innovation" at the same time under the conditions of the information asymmetric. The results show that the optimal incentive contracts of "slack innovation" have nothing to do with the interdependence of the cost functions of the two actions, and when the cost functions of the two actions are complementary, the firm should strengthen the incentive for "production task"; when the cost functions of the two actions are substitute, the firm should weaken the incentive for "production task".

Therefore, in order to improve the performance of production personnel's "slack innovation", on one hand, the firm can reward their "slack innovation" according to the optimal incentive contracts; on the other hand, the firm can optimize the incentive contracts for their "production task" according to the interdependence of the cost functions of "production task" and "slack innovation" to promote indirectly the performance of "slack innovation".

In general, if production personnel's production tasks are full, the cost functions of "production task" and "slack innovation" of production personnel are substitute, then, the firm should weaken the incentive for their "production task" to prevail them to do some "slack innovation" activities, and the weakening degree is increased as the substitute degree of the cost function of these two actions and the marginal value of the principal's "support slack" contribution to the agent's "slack innovation" increases. if the production personnel's production tasks are not full, the cost functions of these two actions are complementary, then, the firm should strengthen the incentive for their "production task" to encourage them to do some "slack innovation" activities, and the strengthening degree is increased as the complementary degree of the cost function of these two actions and the marginal value of the principal's "support slack" contribution to the agent's "slack innovation" increases.

This paper has not considered the impact of the slack level and newly-added resources of a firm on the incentive contracts for production personnel's innovation based on slack resources, which provides opportunity for future research efforts.

Acknowledgments

This paper is supported by the Humanities and Social Science Research Project of the Ministry of Education of China (Grant No. 12YJA630198), the Natural Science Foundation of China (Grant No. 71240023)

References

[1] Wang Yanmei, Zhao Xinan 2008 Incentive and Monitoring for R&D Staff in Software Enterprises *Journal of Northeastern University (Natural Science)* **29**(5) 750-2 *(In Chinese)*

[2] Pan Yingwen, Wan Difang 2010 Study on Influence of Uncertainty on R&D Researchers' Incentive Contract Design, *Chinese Journal of Management* **7**(4) 525-8 *(In Chinese)*

[3] Heping Zhong 2012 Incentive contracts for R&D personnel's technological innovation based on organizational slack *International Review on Computers and Software* **7**(4) 1803-11

[4] Pan Ming 2011 Incentive mechanism for highly skilled talent of SME in Zhejiang Province *China Adult Education* **2011**(19) 144-6 *(In Chinese)*

[5] XIE Zhang-shu, YANG Zhi-rong, XU Qing-rui 2005 A Study on the ALL-involvement Innovation of Enterprise and Organizational Mechanisms *R&D MANAGEMENT* **17**(5) 7-13 *(In Chinese)*

[6] Holmstrom B, Milgrom P 1991 Multi-task principal-agent analyses: Incentive contracts, asset ownership and job design *Journal of Law, Economics and Organization* **7** 24-52

[7] Zhang Weiying 2004 *Game and Information Economics* Shanghai: Shanghai People's Publishing House *(in Chinese)*

[8] Bernardo A E, Cai H, Luo J 2001 Capital Budgeting and Compensation with Asymmetric Information and Moral Hazard *Journal of Financial Economics* **61**(3) 311-44

[9] Gibbons R 2005 Incentives Between Firms (and Within) *Management Science* **51**(1) 2–17

[10] Laffont J J, Martimort D 2002 *The theory of incentives I: the principal-agent model* Beijing, China: China Peoples University Press *(In Chinese)*

[11] WANG Chang-lin, PU Yong-jian 2005 The Research of the Control Rights Incentive Mechanism in Technological Innovation *Journal of Industrial Engineering & Engineering Management* **19**(3) 52-5 *(In Chinese)*

Authors

Heping Zhong, born on August 28, 1966, Chongqing, China

Current position, grades: Associate Professor
University studies: Xuchang University, China
Scientific interest: Technology Innovation, Game Theory
Publications: 48
Experience: Heping Zhong received the Bachelor degree in Engineering from Huazhong University of Science and Technology in 1991, MBA degree from Wuhan University in 2003, Ph.D. degree in Business Administration from Chongqing University in 2009. He is currently an associate professor at Xuchang university, China. His research interests include firm strategy and innovation, high-tech industry development and technological innovation, information technology application. Dr. Zhong is a member of the International Association of Computer Science and Information Technology, and has got several scientific or technical awards of Henan provincial level.

A facilities state-based evaluation method on level of service in subway station

Zhou Huijuan[1]*, Zhao Huan[1], Liu Baoxun[1], Fan Qinglan[2]

[1] *Beijing Key Lab of Urban Intelligent Traffic Control Technology, North China University of Technology, 100144, Beijing, China*

[2] *Key Laboratory of Intelligent Transportation Systems Technologies, Research Institute of Highway, Ministry of Transport, 100088, Beijing, China*

Received 1 March 2014, www.tsi.lv

Abstract

Subway station is the key node in the urban rail transit system. Its level of service affects directly the subway's operation efficiency and traveller's choice of track traffic way to travel. Considering the facility characteristics in subway station and pedestrians perspective, on the basis of a large number of survey data, this paper identifies the facilities which impact the level of service in subway station mainly, takes safety, comfort and smoothness as evaluation index, and evaluates respectively from the entrance, channel and platform area. The judgment matrix of facilities condition influence in each region on pedestrians is constructed and the evaluation model of level of service in subway station has been built based on the facility state. Finally taking PingGuoYuan subway station as instance for analysis, the result verifies that the evaluation method is effective.

Keywords: urban transportation, facilities state, level of service, judgment matrix, pedestrian experience

1 Introduction

Subway has been chosen as the main daily traffic trip with more and more cities starting to build. Thus, it becomes a priority in rail transit to make sure the LOS (level of service) of subway station, which depends on good, reasonable and effective running of facilities in subway station. Accurate evaluation of the LOS could promote the operational efficiency, make passengers safety and provide guidance for equipment maintenance in subway station.

There are many researches on LOS of subway station. Gong X Land Li D P et al studied the LOS of a particular area in subway station. Cao S H and Liu P X et al put forward a method for evaluating LOS, which puts people first. Wang B et al came up with a method to evaluate the LOS of a certain kind of facility in subway station. Many researches on LOS of subway station mainly depend on questionnaire and simulation. M Mori et al put forward a way to evaluate the LOS which depends on advices and behaviours of pedestrian. F Kaakai et al evaluated the LOS through simulation on facilities. There are also some people taking satisfaction of travellers to evaluate the LOS (SA, Lei, 2010). Based on our study of present researches, there are few researches, which take consideration of both passengers' feelings and station facilities.

The LOS of subway station depends on many factors, which result from the interaction of passengers and facilities. Based on existing research results, this paper considers both passengers' feelings and facility characteristics. On basis of the influence of facility on flow density, the main factors which influence LOS of subway station including passengers' feelings, safety, comfort and smoothness are proposed, and the method to evaluate the LOS of subway station which based on facility state is put forward.

2 Data Collection

2.1 INVESTGATION CONTENT

PingGuoYuan subway station, which is the terminal station of Line 1 in Beijing, has apparent traffic tide phenomenon with large passenger flow. We investigated its facility configuration, facility service capability and its aisle access space. The site investigation content is showed in Table 1.

TABLE 1 Investigation content

Survey area	Investigation content
Entrance area	Number of ticket machine Queuing length of pedestrians Available walking area in access area Queuing length on front of safety check (Period of time statistics, flat and peak) Number of automatic gate machine, check time at flat and peak(count the number of people in 15 minutes)
Passage area	Stairs: step height, step length, the max delivery capacity Record the number of checked people per unit time
Platform area	Platform area(length, width), available queuing area, queuing length in 15 minutes at flat and peak(select few platform random) Record the number of people who has arrived platform from passage (15 minutes).

**Corresponding author* e-mail: zhhjuan@sina.com

At the same time, an online questionnaire, covering the age and status of pedestrians, the important degree of auxiliary device, the influence of service facility, the choose wishes and satisfaction of pedestrians and frequency of auxiliary device, is conducted. The degree of importance of facility in this questionnaire has three levels, which is unimportant, general and very important.

2.2 INVESTGATION RESULT

Through the investigation, we can get that the standard equipment configuration in subway station include automatic ticket vending machine, automatic gate machine, security equipment, information display screen, self-service terminals, elevator/escalator, oriented facilities and assistance call. PingGuoYuan subway station has these characteristics: two entrances, outdoor artificial ticket booth, limited flow facility in entry, security equipment, six automatic gate machines in every entrance, and 24 queuing area in platform area.

In this survey, 100 valid questionnaires are taken for study. The proportion of age among the respondents is 1.47% under 18 years old, 80.88% between 18 and 25 years old, 17.65% between 26 and 55 years old. Nobody is above 55 years old. The occupations of respondents include students, office workers and freelancers. There are 47.06% student, 48.53% office workers and 4.41% freelancers. After analysis the collecting data, the factors influencing the LOS of subway station are passengers' age and identity, security equipment, automatic gate machine, automatic ticket vending machine, elevator/escalator, self-service terminals, oriented facilities and assistance call.

3 LOS index system of subway station

3.1 DATA ANALYSIS

The factors which can influence the LOS of subway station fall into three types: behavioural characteristics, traffic efficiency and equipment completeness. The weight of different equipment on LOS and the weight of passengers' age and identity was obtained through YaAHP, which is a software using AHP Method. All of the factors weight is shown in Table 2. The weight in Table 2 was just a reference to judge the main facilities influencing the LOS of subway station because the data is get according to people's subjective feeling. Combining the properties of impact of facilities on pedestrians with Table 2, the main facilities influencing the LOS are automatic gate machine, security equipment and elevator/escalator. The stair is a non-electric device, so it did not been involved though it has big effect on traffic efficiency and travel time. It only involves age and identity on respondent in this survey, so the weight of age and status in Table 2 are the same as the weight of the passengers' feeling. Self-service terminals and assistance call are excluded as secondary cause for their low weight. The index, which can influence the LOS, is confirmed by the characters of main factors including safety, comfort and smoothness.

TABLE 2 Influence factors weight

Influence factor	Weight
identity	0.0714
age	0.0714
security equipment	0.2834
automatic gate machine	0.2834
automatic ticket vending machine	0.0644
elevator/escalator	0.0831
information display screen	0.0356
self-service terminals	0.0138
oriented facilities	0.0797
assistance call	0.0138

3.2 DETERMINATION OF EVALUATION INDEX

Passengers are the service clients of facilities in subway station and they are with perceptual. Therefore, passengers' feeling is one of the factors affecting the LOS. The existing research (Lu C X, 2006) shows that crowded accident would not happen when the flow density is low and pedestrians walk freely. When the flow density reaches a critical point, the interaction of speed and density will lead to a crowded accident. There is a direct relationship between comfort level and pedestrian density. The pedestrian walking speed slows down and will feel uncomfortable as the pedestrian density increases. Meanwhile the safety, comfort and smoothness will get lower. Therefore, this paper will evaluate the safety, comfort and smoothness with the pedestrian density as the core.

The interaction among pedestrians brings crowded feeling and either does the interaction between facilities and pedestrians. The crowded will increase when some facility does not work well. Since the degree of congestion in subway station has different feature in different area, this paper divides the transfer area into three parts for study, including entrance, passage and platform. There are security equipment, automatic gate machine and ticketing equipment in entrance, where has relatively concentrated crow flow and will produce queues. Elevator, escalator and stairs are in passage, which has the characteristics of uniform flow density and speed. The platform area's main characteristics are scattered distribution of crow flow and most of the crow flow distributed in entrance area of the train.

(1) Safety: this paper use U_1 to show the safety degree of facility. It represents the ratio of crowd flow in the area to the max crowd flow it can bear (Ran L J, Liu M, 2007). U_1 is calculated using Equation 1 below.

$$U_1 = 1 - \frac{p}{p_{\max}}, \qquad (1)$$

where p is crowd flow in the area and $p_{\max}$ is the max crowd flow it can bear.

(2) Comfort: this paper use U_2 to show the degree of comfort, that is the passengers' feeling about the subway station service. It is the ratio of average flow density in the area to the max crowd flow, which people can endure.

$$U_2 = 1 - \frac{k}{k_{\max}}, \qquad (2)$$

where k is average flow density of the facility, $k_{\max}$ is the max flow density people can bear.

(3) Smoothness: U_3 was used to show the level of smoothness. The measure method is using the ratio of average delay time of getting through the area to the average time of transfer in the subway station. U_3 is calculated according to the following formula.

$$U_3 = 1 - \frac{t}{T}, \qquad (3)$$

where t is average delay time of pedestrians at peak and the T is average time of passing the subway station.

4 Comprehensive LOS evaluation of subway station

4.1 ESTABLISHING THE HIERARCHY STRUCTURE OF EVALUATION

The LOS of subway station is evaluated by these factors including comfort level, safety and smoothness of pedestrians in subway station. The safety, comfort and smoothness will be evaluated in three areas and can be estimated by the configuration of facilities and the influence of facilities on people in each region. Figure 1 shows the evaluation structure of LOS.

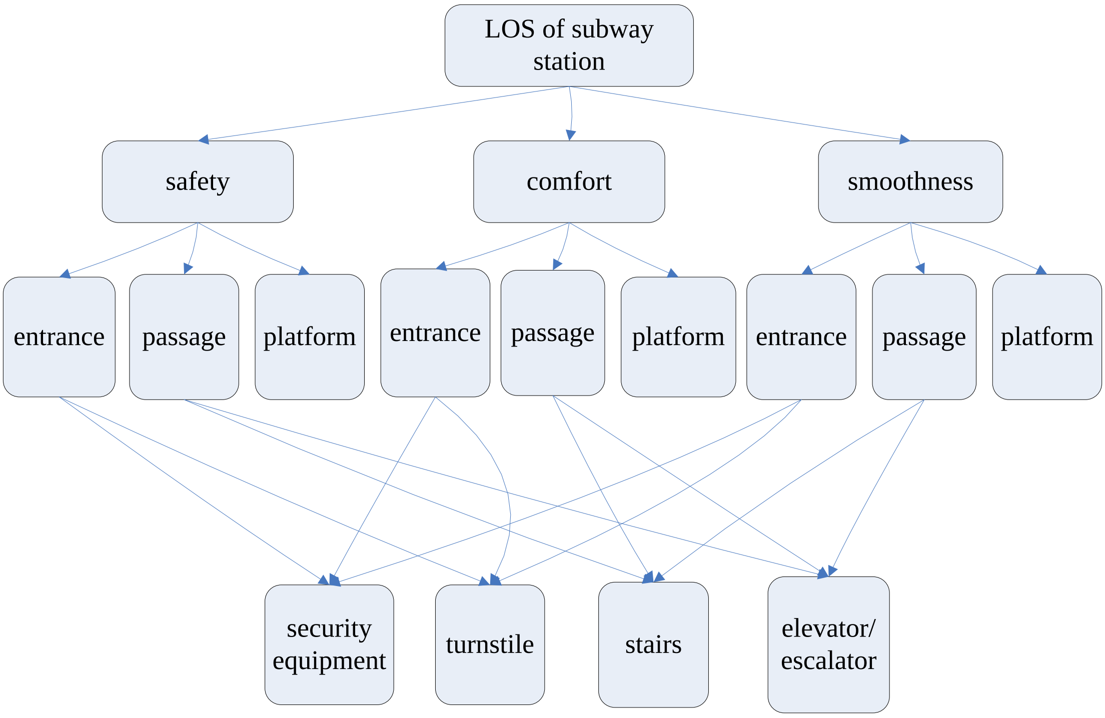


FIGURE 1 Hierarchy structure of evaluation

4.2 BUILDING THE JUDGMENT MATRIX

For requiring the influence weight of each region, the AHP method is used to analyse the impact of the facilities in each region on safety, comfort and smoothness according to 1-9 degree calibration method and get the judgment matrix.

(1) The basis of judging the influence weight of facilities on safety in each region. As the definition of safety, the region safety is related to passenger flow and the possible accident when the passenger flow exceeds the carrying capacity of service facility. So the weight of each facility can be judged by the transport capacity of the facility, and the transfer capacity will go down when the facility does not work. The declining degree of safety caused by the change of operating status of each facility can be calculated by the following formula.

$$\Delta U_1 = \left(1 - \frac{S\rho_0}{P_{\max}}\right) - \left(1 - \frac{\rho_1}{P_{\max}}\right), \qquad (4)$$

where ρ_0 is the average flow density when the facilities in this area is good. ρ_1 is the average flow density when the facilities' operating status changed. $P_{\max}$ is the standard flow density of safety in China which is 9 per/m². The influence weight of each facility on region safety can be got by comparing ΔU_1 of each facility. The influence weight of each region safety on the safety of subway station can be measured by hidden danger and the flow density of each region.

(2) The basis of judging the influence weight of facilities on comfort in each region. Region comfort is related with average passenger flow in the area, which would increase when the facility does not work well. Therefore, the weight of each facility can be judged by the facility influence on average flow density. The declining degree of comfort caused by the change of operating status of each facility can be calculated by the following formula.

$$\Delta U_2 = \left(1-\frac{k_0}{k_{\max}}\right)-\left(1-\frac{k_1}{k_{\max}}\right), \tag{5}$$

where k_0 is the average flow density when facilities in the area are available. $k_{\max}$ is the max flow density people can bear which can be get by questionnaire. k_1 is the average flow density when the status of facility changes. Here k_0 is the same as ρ_0 and k_1 is the same as ρ_1 in formula (4). The influence weight of each region on the comfort of subway station can be measured by the duration of crowded.

(3) The basis of judging the influence weight of facilities on smoothness in each region. Region smoothness is related to average delay time. The weight of each facility can be judged by the facility efficiency. The crowded will get worse once the operating status of facility gets bad. That is to say, the flow density will increase, the pedestrian speed will slow down, and the average delay time will increase. Hence, the declining degree of smoothness caused by the change of operating status of each facility can be calculated by the following formula.

$$\Delta U_3 = \left(1-\frac{t_0}{T}\right)-\left(1-\frac{t_1}{T}\right), \tag{6}$$

where t_0 is the average delay time at peak when all the facilities is available. t_1 is the average delay time when the operating of facility changes. T is the time of passing the subway station. The influence weight of each region on the smoothness of subway station can be measured by the time of staying in the area.

The weight of safety, comfort and smoothness of each facility can be obtained by 1-9 point calibration and judgment matrix according to the formulas (4), (5) and (6).

Then it is turn to calculate the weight and check the consistency using the software YaAHP, which can simplify the procedure. Importing the judgment matrix directly to YaAHP, and then it will check the consistency automatically. If the matrix has good consistency, the weight it calculated is the one we need, otherwise the elements in the matrix need be adjusted to until good consistency and the weight should be calculated again.

4.3 EVALUATION MODEL

The service level of subway station can be evaluated by the degree of comfort, safety and smoothness. The comfort of entrance can be calculated as the formula $C_1 = U_{21}(\alpha_1 z_1 + \alpha_2 z_2)$, U_{21} is the comfort of entrance, z_1 is the operating status of security equipment, z_2 is the operating status of automatic gate machine, α_1, α_2 respectively shows the influence weight of security equipment and automatic gate machine on the entrance. The comfort of passage region can be calculated as the formula $C_2 = U_{22}(\alpha_3 z_3 + \alpha_4 z_4)$, U_{22} is the comfort of passage area, z_3, z_4, is the operating status of stair and escalator, α_3, α_4 respectively shows the influence weight of stair and escalator on the passage area. There is no facility which can effect the comfort in platform area, so the comfort of platform is $C_3 = U_{23}$. The influence weight of comfort in entrance, passage and platform on the whole subway station are ω_1, ω_2 and ω_3. The comfort of subway can be calculated as the following formula.

$$C = \omega_1 U_{21}\left(\alpha_1 z_1 + \alpha_2 z_2\right) + \omega_2 U_{22}\left(\alpha_3 z_3 + \alpha_4 z_4\right) + \omega_3 U_{23}. \tag{7}$$

The safety of subway can be calculated as:

$$S = \alpha_1 U_{11}\left(s_1 z_1 + s_2 z_2\right) + \alpha_2 U_{12}\left(s_3 z_3 + s_4 z_4\right) + \alpha_3 U_{13}, \tag{8}$$

α_1, α_2 and α_3 respectively shows the influence weight of entrance, passage and platform on the safety of the subway station. U_{11}, U_{12} and U_{13} respectively shows the safety of entrance, passage and platform. s_1 and s_2 respectively shows the influence weight of security equipment and automatic gate machine on the safety in the entrance area. s_3 and s_4 respectively shows the influence weight of stair and escalator on the safety in the passage area.

Similarly, the degree of smoothness can be calculated as:

$$F = \beta_1 U_{31}\left(m_1 z_1 + m_2 z_2\right) + \beta_2 U_{32}\left(m_3 z_3 + m_4 z_4\right) + \beta_3 U_{33}. \tag{9}$$

The influence weight of comfort, safety and smoothness on the LOS of subway station respectively is α_2 and α_3, so the LOS of subway station can be calculated as the following formula.

$$L = \alpha_1 C + \alpha_2 S + \alpha_3 F \,. \tag{10}$$

The LOS of subway station can be calculated according to the value of *L* and showed in Table 3.

TABLE 3 Evaluation standard

LOS	Score
A excellent	1.0
B very good	0.9
C good	0.8
D qualified	0.6
E bad	0.4
F very bad	0.2

5 Case study

PingGuoYuan subway station is taken as an example to evaluate the LOS of subway station. We survey this subway station for a whole week and select the peak time 7:00-9:00 as the survey time. The passenger flow and transfer time during the peak time are recorded. Then we establish judgment matrix according to the survey data and use the YaAHP software to get judgment matrix and weight.

(1) The first level judgment matrix and weight

The ratio of matrix consistency is 0.0825, and the general objective weight is 1.0 (Table 4).

TABLE 4 LOS judgment matrix

LOS	Comfort	Safety	Smoothness	Weight
Comfort	1	1/5	1/3	0.1007
Safety	5	1	4	0.6738
Smoothness	3	1/4	1	0.2255

(2) The second level judgment matrix and weight

The ratio of matrix consistency is 0.0000 and the general objective weight is 0.1007 (Table 5)

TABLE 5 Comfort judgment matrix

Comfort	Entrance	Passage	Platform	Weight
Entrance	1	3	3	0.6
Passage	1/3	1	1	0.2
Platform	1/3	1	1	0.2

The ratio of matrix consistency of safety is 0.0236, and the weight to the general objective is 0.6738 (Table 6)

TABLE 6 Safety judgment matrix

Safety	Entrance	Passage	Platform	Weight
Entrance	1	2	1/4	0.1998
Passage	1/2	1	1/5	0.1168
Platform	4	5	1	0.6833

The ratio of matrix consistency of smoothness is 0.0236, and the general objective weight 0.2255 (Table 7)

TABLE 7 Smoothness judgment matrix

Smoothness	Entrance	Passage	Platform	Weight
Entrance	1	4	5	0.6833
Passage	1/5	1	2	0.1998
Platform	1/5	1/2	1	0.1168

According to the influence of security equipment, automatic gate machine, stair and escalator on pedestrians, the influence weight of comfort, safety and smoothness are got in its region respectively. The comfort weight of the equipment to its region is 0.75, 0.25, 0.25, 0.75, the safety weight respectively is 0.83, 0.17, 0.75, 0.25 and the smoothness weight respectively is 0.75, 0.25, 0.25, 0.75.

We use Thursday's data to compare difference of LOS when facilities' status changes, for an automatic gate machine does not work on that day. Because of its stabilized passengers flow, Friday's data is selected to compare with the LOS on Thursday. According to the survey, there are six automatic gate machines in PingGuoYuan subway station; five is good and one does not work on Thursday, all of them are well on Friday. The passengers flow of entrance is 5860 per/h on Thursday and 5560 per/h on Friday.

Suppose z=1 when the facility is well, z=0 when facility's status changes. Then z_1 =5/6=0.83, z_2 =1, z_3=1, z_4 =0. On the basis of the survey data, the safety at peak time is U_{11} =0.89, U_{12} =0.75, U_{13} =0.78; the comfort is U_{21} =0.8, U_{22} =0.95, U_{23} =0.54, and the degree of smoothness is U_{31}= 0.94, U_{32} =1, U_{33}=1.

The LOS of subway station on Thursday and Friday can be calculated as the formula (7), (8), (9) and (10). The results are showed in Table 8.

TABLE 8 LOS of PingGuoYuan

Thursday	C=0.7168	S=0.773	F=0.877	L=0.79
Friday	C=0.778	S=0.78	F=0.96	L=0.82

Contrast with the Table 3, the LOS of PingGuo Yuan subway station on Thursday is between C and D, the LOS is close to good. The service level of PingGuoYuan subway station on Friday is good. Contrast with it on Thursday, the status of facility can influence the LOS of subway station, and the LOS will decline from good to qualified when an automatic gate machine is bad. The LOS of PingGuoYuan subway station has been evaluated accurately by this method and it corresponds to the actual situation.

6 Conclusion

On the basis of site investigation and online questionnaire, every equipment influence to passengers is calculated and the key facilities affect the LOS of subway station are required. Then the paper calculates the influence weight of facility status on comfort, safety and smoothness according to the influence of facilities status on pedestrian density. An evaluation model about LOS of subway station is established and the judgment matrix in the model is got by using YaAHP software. The evaluation model and method can provide reference to the configuration of key facilities in subway station and improve the LOS during the peak time.

Acknowledgements

This paper was supported by the National High Technology Research and Development Program of China (No. 2012AA112401), Beijing Municipal Science and Technology Project (KM20140009002), and the Opening Project of Key Laboratory of Intelligent Transportation Systems Technologies, Ministry of Communications, China.

References

[1] Gong X L, Wei Z H 2009 A method for evaluation of level of service in the area for waiting in line *Journal of Beijing University of Technology* **35**(10) 1373-7 *(In Chinese)*

[2] Li D P, Yan K F, Xu M M, Liu H 2012 Evaluation Analysis of Arriving Level of Service for Urban Rail Station Facilities *Journal of Tongji University(Natural Science)* **38**(10) 1458-62 *(In Chinese)*

[3] Liu P X, Yi J 2012 The evaluation model of service level of urban rail transit based on idea of Humanization *Science & Technology Information* (3) 232-3 *(In Chinese)*

[4] Cao S H, Yuan Z Z, Zhang C Q, Zhao L 2009 LOS Classification for Urban Rail Transit Passages Based on Passenger Perceptions *Journal of Transportation Systems Engineering and Information Technology* **9**(2) 99-104 *(In Chinese)*

[5] Wang B, An S Z, Li X X 2007 Evaluation System for the Transfer Infrastructure of Urban Rail Transit *Urban Rapid Rail Transit* **20**(4) 40-3 *(In Chinese)*

[6] Mōri M, Tsukaguchi H 1987 A new method for evaluation of level of service in pedestrian facilities *Transportation Research Part A: General* **21**(3) 223–34

[7] Kaakai F, Hayat S, Moudni A E 2007 A hybrid Petri nets-based simulation model for evaluating the design of railway transit stations *Simulation Modelling Practice and Theory* **15**(8) 935-69

[8] Sa L 2010 Evaluation model of passenger satisfaction degree of service quality for urban rail transit *Integrated Transportation Systems: Green, Intelligent, Reliable - Proceedings of the 10th International Conference of Chinese Transportation Professionals,* 1464-73

[9] Lu C X 2006 Analysis on the Wave of Pedestrians *China Safety Science Journal* **16**(2) 30-4 *(In Chinese)*

[10] Ran L J, Liu M 2007 Effects of crowded people density on crushing fatalities *Journal of Safety and Environment* **27**(4) 135-8

Authors	
	Huijuan Zhou, born on April 4, 1975, Hunan province, China **Current position, grades:** Assistant professor at Beijing Key Laboratory of Urban Intelligent Traffic Control Technology, North China University of Technology. **University studies:** her M.Sc. in Cartography & Geographic Information System (2002) from Beijing Normal University and PhD in System Analysis & Integration (2011) from Beijing Jiaotong University. **Scientific interest:** intelligent transportation system and traffic safety.
	Huan Zhao, born on October 20, 1990, Hubei province, China **Current position, grades:** graduate student at Beijing Key Laboratory of Urban Intelligent Traffic Control Technology, North China University of Technology. **University studies:** Bachelor's degree in automation (2012) from Xi'an Polytechnic University. **Scientific interest:** intelligent transportation control.
	Baoxun Liu, born on April 29, 1987, Hebei province, China **Current position, grades:** graduate student at Beijing Key Laboratory of Urban Intelligent Traffic Control Technology, North China University of Technology **University studies:** Bachelor's degree in Industrial Engineering (2011) from ShenYang Ligong University. **Scientific interest:** His current research interest involves intelligent transportation control.
	Qinglan Fan, born on May 17, 1983, Beijing, China **Current position, grades:** Assistant Researcher at Key Laboratory of Intelligent Transportation Systems Technologies, Research Institute of Highway, Ministry of Transport. **University studies:** M.S. in Control Engineering (2012) from Beijing University of Technology. **Experience:** She has worked on intelligent transportation planning and transportation information for nine years.

Resource management modelling and simulating of construction project based on Petri net

Hailing Li^1, Kejian Liu^{2*}

[1] *School of Architecture and Civil Engineering Xihua University, Chengdu, 610039, China*

[2] *School of Mathematics and Computer Engineering Xihua University, 610039 Chengdu, China*

Received 1 March 2014, www.tsi.lv

Abstract

This paper establishes a model to exactly express the resource configuration, task duration and information transmission during the project execution phase. Based on the resources' properties in the projection execution phase and the hierarchical timed coloured Petri net (HTCPN), this hierarchical model exactly express the information required for project resource management, such as the task dependencies, resource demands and the task durations by defining a non-empty colour set as coloured tokens to represent the classes and combinations of the resources. This model is then simulated and analysed on the model structure, resource conflicts and run time using CPN Tools to verify the correctness and effectiveness of the HTCPN modelling of the project resources in the project execution phase.

Keywords: Construction Project, Petri Net, Resource, Modelling, Simulation

1 Introduction

The execution of a project has various constraints such as the time, resources and information, in which the resources are the key constraint having direct influence on the project cycle and economics [1]. There are three existing methods of scheduling and optimizing the project resources: peak shaving and least variance methods to optimize resources configuration. (They do not consider the random factors that may occur in the project execution and the consequence); mathematically modelling based on resource demands to optimize resources configuration using the heuristic algorithm or the meta-heuristic algorithm to seek a globally optimized solution. (The application of this method is limited due to the modelling difficulty); simulation method to optimize the resources configuration. CYCLONE (Cycle Operation Network) is the first simulation technique wildly used in project execution [2]. It integrates the queuing theory and the simulation technique into the network scheduling to consider the random factors that may occur in a cycle operation. However, this method cannot characterize the synchronous, asynchronous, parallel, and resource-sharing properties between tasks so well.

The Petri net is a modelling tool with the properties of strict formal expression, matured mathematical analysis and visual graphic representation [3]. A model established for project management using the Petri net not only has the advantage of the one by CYCLONE to characterize the random in system but also visually and in more detail depicts the parallel, synchronous and resource sharing properties between tasks. By defining, a non-empty colour set as coloured tokens to represent the classes and combinations of the resources, this paper establishes a resource management model for a project in execution phase based on the hierarchical timed coloured Petri net. It conveniently and exactly expresses the information needed for resource management, such as the task dependencies, resource demands and task durations. This resource management model is then simulated and analysed on the model structure, resource conflicts and run time using CPN Tools.

2 Resource Properties of a Construction Project in Execution Phase

In execution phase, a construction project involves a large number of resources, in which the human resource, materials and machines are the key resources. The resources can be divided into consumable resources (e.g. the materials) and non-consumable resources (e.g. human resource and machines). This paper discusses the management of the non-consumable resources. In the construction project execution phase, the non-consumable resources have the following properties [4].

a) Reusability. Subject to certain constraints, the non-consumable resources can be reused. In contrast, the consumable resources are not reusable.

b) Shareability. Subject to certain constraints, resources of one class can be shared by various tasks.

c) Exclusivity. Under certain conditions, a resource can only be used by one task. In a finite set of resources,

**Corresponding author* e-mail: liukejian@gmail.com

the reusability and exclusivity of one resource may bring to resource conflict.

3 Definition of a Hierarchical Timed Coloured Petri Net

HTCPN=(S,SN,SA,PN,PT,PA,FS,FT,PP) [5][6], where S is a finite set of pages, belonging to which each page s ($s \in S$) is a non- hierarchical timed coloured Petri net, TCPN=(P,T,A,C,W,M,G,I,I_0) [7] and $\forall s_1,s_2 \in S: s_1 \neq s_2 \Rightarrow (P_{s1} \cup T_{s1} \cup A_{s1}) \cap (P_{s2} \cup T_{s2} \cup A_{s2}) = \varnothing$.

SN is a set of substitution nodes ($SN \subseteq T$).

SA is the page assignment function, through which SN maps to S.

PN is a set of port nodes ($PN \subseteq P$). The inter-page message flow functions via these port nodes.

PT is a port type function, through which PN maps to {in,out,i/o,general}.

PA is the port assignment function that is a binary relation to SN and subject to the following constrains.

$\forall t \in SN: PA(t) \subseteq X(t) \times PN_{SA(t)}$

$\forall t \in SN$, $\forall (p1,p2) \in PA(t)$:

$[PT(p_2) \neq general \Rightarrow ST(p1,t)=PT(p2)]$

$\forall t \in SN, \forall (p1,p2) \in PA(t)$: $\square [C(p_1)=C(p_2) \wedge I(p_1)=I(p_2)]$

FS is a finite set of unit type, each element of it has the same colour and initial expression, i.e.

$\forall fs \in FS, \forall p_1,p_2 \in fs: [C(p_1)=C(p_2) \wedge I(p_1)=I(p_2)]$

FT is a unit type function. It maps FS to [global, page, instance]. The set of page union nodes and the set of local nodes are included in the same page, i.e.

$\forall fs \in FS$: $\square [FT(fs) \neq global \Rightarrow \exists s \in S: fs \subseteq Ps]$

PP is a multi-set defined in the root page, which is the top-most page.

4 Modelling of Resources in Construction Project Execution Phase

4.1 DEFINING THE NON-EMPTY COLOR SETS

C is the non-empty coloured set of the HTCPN. Before modelling the resources, it is required to define a number of data types, that is to say, define coloured tokens to represent the classes and combinations of various resources. In this paper, the definition of coloured tokens is focused on the non-consumable resources.

The colour sets are defined through CPN ML. For modelling of the resources of the project execution phase, the following colour sets are defined.

a) colset Pro = int timed; Timed Project.

b) colset P = int timed; Timed Partial Project (Task).

c) colset S = bool; Acceptance Criteria of Project. Its token in place is of Boolean type.

d) colest res = int; Resource Classes, e.g. 1 represents the civil work engineers, 2 the piping engineers and 3 the electrical engineers.

e) colset ress = list res; List of Resources.

f) colset req = record n:INT*m:INT*tim:INT*rn1:INT*rn2:INT*rn3:INT Resource Request,

where

n is the number of the project item that sends the resource request.

m is the number of the project sub-item that sends the resource request.

tim is the time when the request is sent.

rn1, rn2, rn3 are the requested amount of the three classes of resource respectively.

g) colset reqs = list req; List of Resource Requests.

h) colset mtres = record m: INT*t:INT*r: ress; List of Resources that the Project Sub-item m Received at the Time t.

i) fun FuzzyTime (a,b,c); Function Declaration. It is the fuzzy time when the task completes, where a is the minimal duration, b the most possible duration and c the maximal duration, of the task.

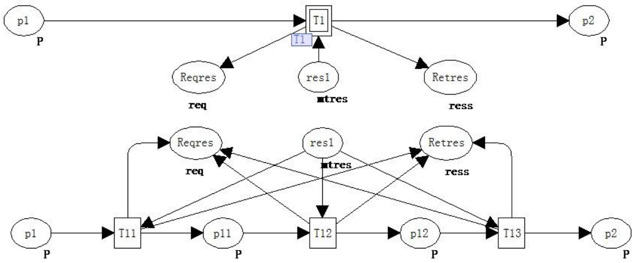


FIGURE 1 Hierarchical characterization of resource management through the substitution transitions

4.2 MODELING THE RESOURCE MANAGEMENT IN HIERARCHY

A project consists of a number of project items that are to be performed in sequence, in parallel or synchronously. As soon as a project item starts, it inevitably requests the resources, receives them and returns the (non-consumable) resources when the task ends. A project item divides into several sub-items. As project items, the sub-items are processes to consume resources (consumable resources) or to use non-consumable resources. That is to say, requests, assignment and return of resources of a project item are finally those of the divided sub-items. Figure 1 shows the hierarchy of the resource management. Through the substitution transitions, the resource management is characterized in a hierarchical net. T1 is start transition of the project item. It requests, receives and returns resources from the places Reqres, res1 and Retres. T1 is a substitution transition, which belongs to the page of its parent net. The net (page) in the dotted box is the subnets of T1. The start transitions of several project sub-items, T11, T12 and T13 request resources from the place Reqres and one sub-item starts when it receives the resources from the place res1 and returns the resources to the place Retres when it completes. In spite of the transitions of parent net and its subnet(s) relates to the same set of places, the hierarchical characterization simplified the net model. Furthermore, sub-items may be modelled as common models, which can be repeatedly used in modelling of different projects to enhance the modelling efficiency.

4.3 MODELING OF RESOURCE ASSIGNMENT AND RETURN FOR RESOURCE MANAGEMENT

Figure 2 shows a basic model of resource assignment and return. All the request tokens for resources sent by the tasks go into the place Reqres. When the amount of resources in the places pool1, pool2 and pool3 satisfies the requests from the place Reqres, the transition R1 assigns the requested resources to the resource output places res1, res2, res3, res4 for the tasks respectively. If the amount of resources does not satisfy the requests from the place Reqres, resources confliction occurs.

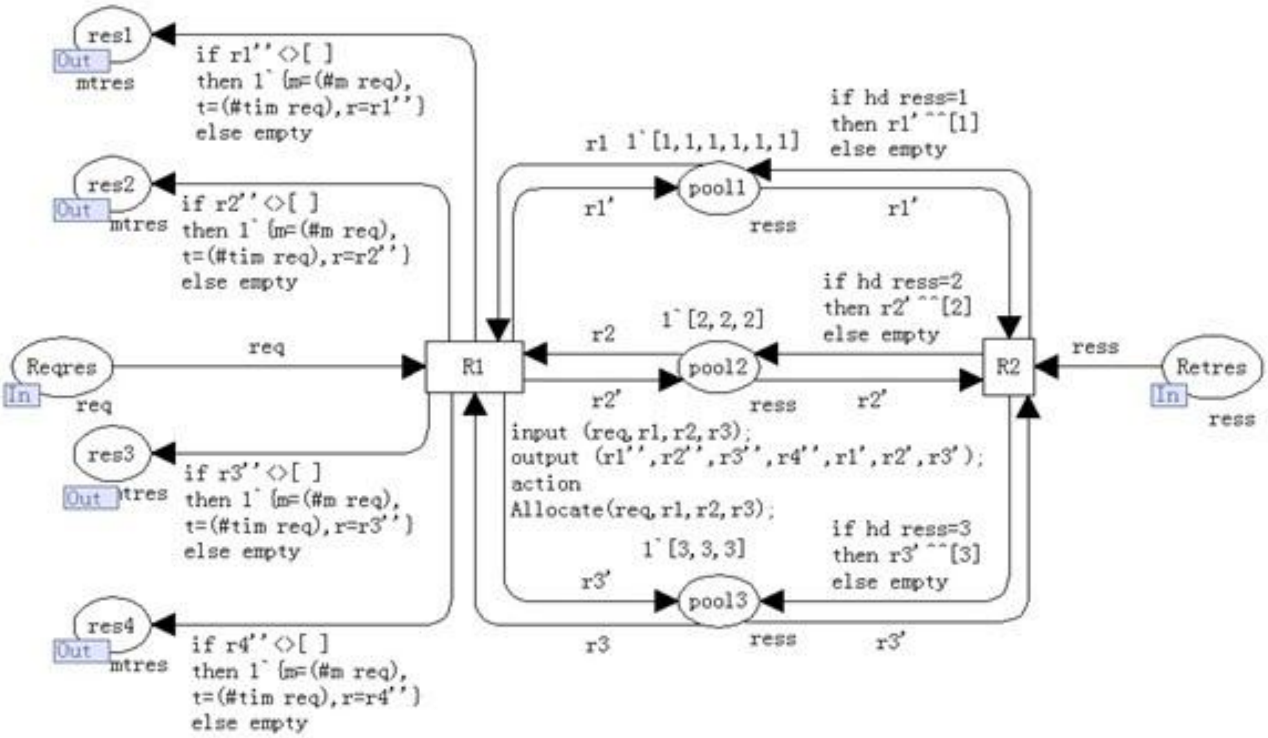


FIGURE 2 The basic model of resource assignment and return

5 Examples of Modelling and Simulation

In this paper, a project is taken as an example to model the source management for simulation. The CPN tools from Aarhus University are used as the modelling and simulation platform.

5.1 MODELLING

The example project consists of the construction word p1, installation work p2 and the furnishing work p3. To perform the works in parallel, the work p1 is divided into two phases, t1 and t3. t4 is the execution of p2 and t5 the execution of p3. They are partially parallel with t1 and t3 through multi-dependencies. t1 and t3 are substitution transitions, which represent the construction work of the underground foundation and the structure under the Floor 1, and the structure of and above Floor 1, respectively. Also, t4 and t5 are substitution transitions. t1, t3, t4 and t5 send resource requests, receive the resources assigned to them and return the resources at the tasks' ends. The substation transition Res globally controls the resource movement of t1, t3, t4 and t5, as shown in Figure 3. Of t1, t3, t4 and t5, the sub-item works have their own subnets for characterizing the resource movement. Figure 4 shows the subnets of t1. For simplification, other subnets are not shown. Figure 5 shows the subnets of Res, which are modelled base on the basic model of resource assignment and return (as shown in Figure 2) and added the timer transition TIMER to generate the time list poolt for assigning the resource(s) under any confliction condition. It sequentially assigns the resources in the order of request time. In the event that a request waits for one resource for some time, the time stamp of the resource will be changed at the time of assigning this resource. When the resource is returned, the time stamp of the resource is changed to the value the waiting time adding the task duration, as shown in Figure 5. For simplification, the definition of variables and functions and the description of transitions and places are not discussed in this paper.

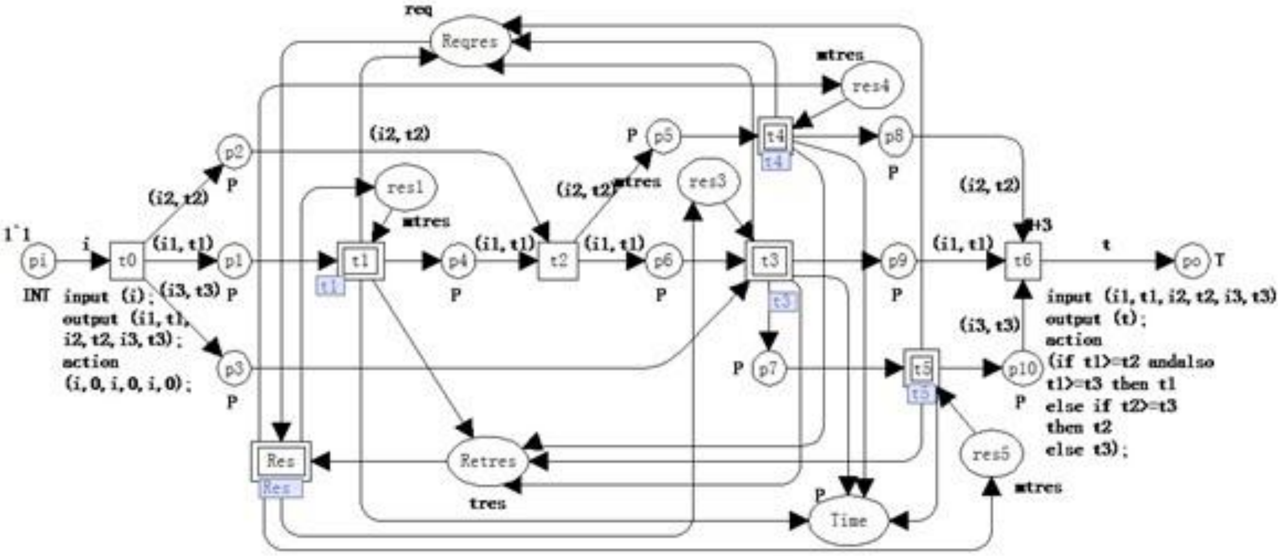


FIGURE 3 Resource management model (top)

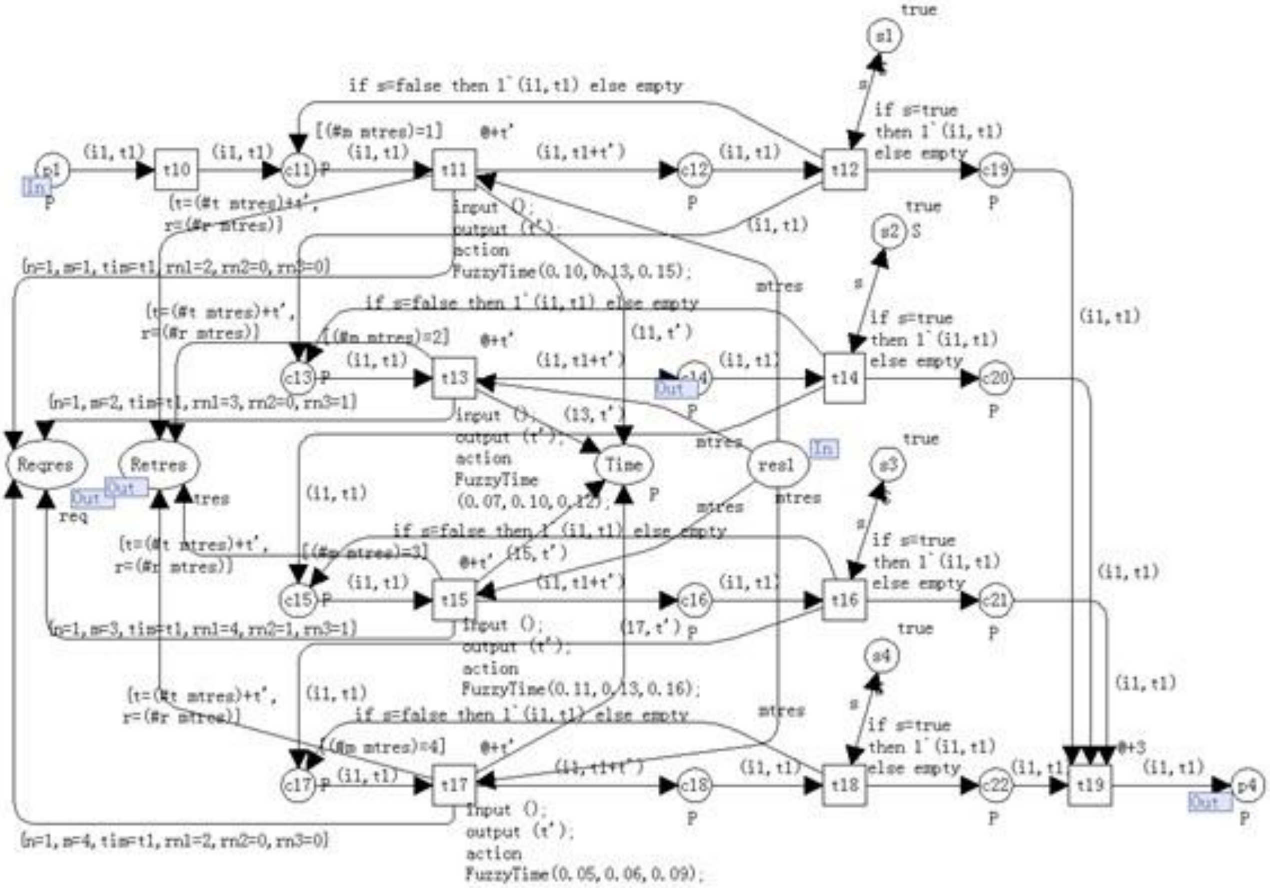


FIGURE 4 Resource management model (the subnets oft1)

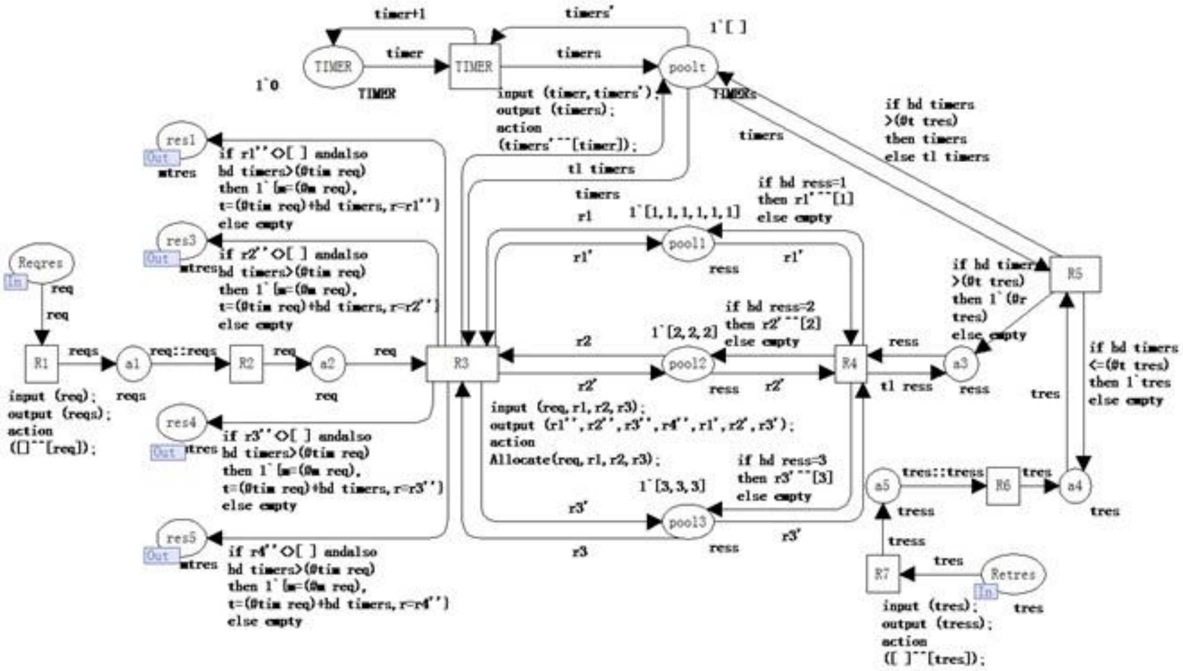


FIGURE 5 Resource management model (Resource Layer)

5.2 SIMULATION

5.2.1 Setting the Parameters

Table 1 shows the basic parameter settings for the model simulation, in which r1, r2, r3 represent construction engineers, piping engineers and electrical engineers respectively. In Table 1, the individual quantities of the three types of engineers, the quantities of engineers that various tasks requires and the fuzzy duration of the tasks are also shown.

TABLE 1 Resource list of tasks in the execution phase

	r1	**r2**	**r3**	**Time Required (d)**
	4	2	2	
t11	2	0	0	@+(FuzzyTime(10,13,15))
t13	3	0	1	@+(FuzzyTime (7,10,12))
t15	4	1	1	@+(FuzzyTime (11,13,16))
t17	2	0	0	@+(FuzzyTime (5,6,9))
t31	4	1	1	@+(FuzzyTime (90,110,120))
t33	3	1	1	@+(FuzzyTime (35,40,45))
t35	2	1	1	@+(FuzzyTime (15,20,29))
t41	1	2	2	@+(FuzzyTime (110,120,130))
t43	0	2	2	@+(FuzzyTime (25,30,35))
t51	2	1	1	@+(FuzzyTime (55,60,65))
t53	2	0	0	@+(FuzzyTime (50,60,90))

5.2.2 Simulating the Model

a) General Analysis of the Petri Net Established.

The feature of State Space of CPN Tools is used for checking the syntax and analysing the structure of the resource management model established to get the state space report, which shows the reachability, boundedness and liveness properties of the established model. By these properties, the model can be justified.

b) Conflictions

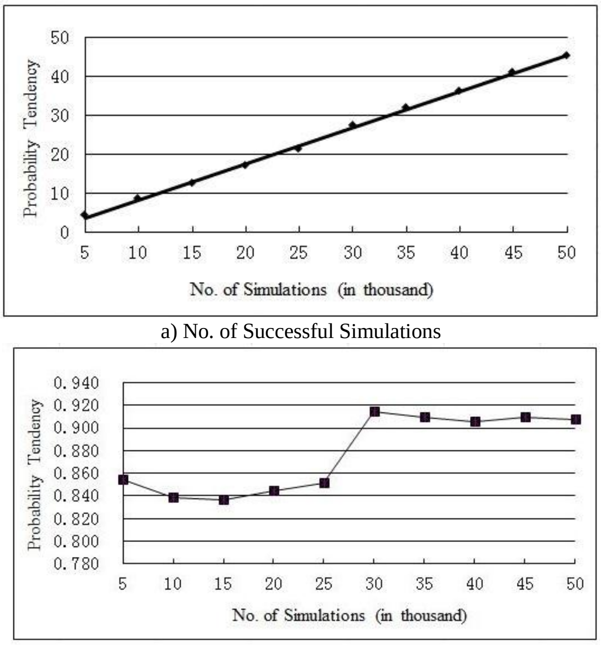
Due to the tasks performed in parallel and the exclusive property of the engineers, a confliction occurs when an engineer is requested by more than one task. To the model, a confliction occurs when the finite quantity of the coloured tokens in the resource place is short to fire two or more transitions (tasks) at one time. The real firing time of the task that lacks of resources will be delayed due to waiting for the resources and the time stamps of the tokens in the resource places are changed.

c) Analysis of Run Time

Under the resource constraints shown in Table 1, 5,000 times of simulation are made to the resource model using CPN Tools Simulation for the probability of the construction project ending in 280 days, as shown in Table 2. As the simulation progresses, the probability of this industrial and civil construction project ending in 280 days tends to 91%, as shown in Figure 6. The resources configured in the places r1, r2 and r3 satisfy the need of project schedule.

TABLE 2 Data sheet of resource management model simulation

No. of Simulations	**No. of Successful Simulations**	**Probability Tendency**
5000	4270	0.854
10000	8386	0.839
15000	12546	0.836
20000	16893	0.845
25000	21276	0.851
30000	27427	0.914
35000	31823	0.909
40000	36225	0.906
45000	40931	0.91
50000	45381	0.908



a) No. of Successful Simulations

b) Probability Tendency

FIGURE 6 Simulation for the Probability of Project Ending in 280 days: a) No. of Successful Simulations, b) Probability Tendency

6 Conclusions

This paper discussed the resource properties of the project in execution phase and established a model for resource management. CPN Tools are used for modelling the instance and simulation to check the model structure, find out any resource conflictions and analyse the run time so as to justify the model and verify its correctness. The model can be used to assess various project schedules through setting and adjusting the resource input and select one from them that has high resource utilization rate while keep the optimized project duration.

Acknowledgments

The author was supported by National Natural Science Foundation of China (61271413), the National Key Technology R&D program (No.2011BAH26B00) and the Key scientific research fund project of Xihua University (Grant No: Z1320607)

References

[1] CHENG Fei-fei, WANG Yao-wu 2009 *Journal of System Simulation*. **21** 7727731

[2] YANG Xue-hong, HU Zhi-gen 2001 *International Journal Hydroelectric Energy*. **1** 33-6

[3] Wakefield R R, Sears G A 1997 *Construction engineering and management*. **123** 105-12

[4] Wang Yaowu, Cheng Feifei 2009 *China Civil Engineering Journal*. **42** 189-96

[5] Zu Xu, Huang Hong-zhong, Zhou Feng, Gu Ying-kui 2005 *Journal of System Simulation*. **17** 1322-25

[6] CHEN Chun-liang, WANG Yan-lei, SUN Sheng-kun 2008 *Journal of System Simulation*. **20** 2746-9

[7] W M P van der Aalst 1998 *Journal of Circuits, Systems and Computers*. **8** 21-66

Authors

	Hailing Li, born on March 29, 1976, Sichuan, China **Current position, grades:** Associate Professor, Ph.D. **Research Interests:** Project optimization techniques, project management
	Kejian Liu, born on June 1, 1974, Hubei, China **Current position, grades:** associate professor **Research Interests:** Computer Network, Database and Information System

Study on distribution centre's location selection of internal supply chain for large group manufacturing companies

Lifei Yao, Ruimin Ma, Maozhu Jin*, Peiyu Ren

Business School, Sichuan University, 610064, Chengdu, Sichuan, China

Received 1 March 2014, www.tsi.lv

Abstract

The purpose of this paper is to study what distribution centre's location selection can bring to the internal supply main management for large group manufacturing companies. This paper chooses the analytic hierarchy process to select an optional location for internal distribution centre', and evaluate it through the simulation method. Internal distribution centre construction can effectively shorten the delivery time, reduce the logistics intensity, and improve utilization rate of transport equipment. Therefore, the distribution centre's location selection is necessary and reasonable. This paper simplified some information when running the simulation and it is not all the same as the actual situation. This paper provides a good internal supply main management method for large group manufacturing companies. This paper put forward the importance of internal supply chain for a large group manufacturing company and studied the internal distribution centre's location selection.

Keywords: manufacturing, internal supply chain, distribution centre, simulation

1 Introduction

Scientific and reasonable distribution centre's location plays a vital role for balancing the entire supply chain logistics, the logistics service reliability and management level of ascension. Appropriate location can effectively reduce the cost of supply chain, improve the overall efficiency of the supply chain and reduce the impact of the bullwhip effect. Therefore, the internal distribution centre's location is very important. The evaluation of distribution centre's location selection includes evaluation system and evaluation method and a lot of scholars at home and abroad have studied about the problem in the literatures [1-6].Nine basic location models were given by Alkenes, which included the simple incapacitated facility location model, the capacitated facility location model, the dynamic and stochastic capacitated facility location models etc. [7]. Holmberg [8] studied the exact solution method for the incapacitated facility location problem in which the transportation costs were nonlinear. An integer programming model for the plant location was presented by Badalona and Jensen [9]. It considered not only the fixed costs and transportation costs, but also the inventory costs, which had been solved by the Danzig-Wolfe (D-W) decomposition method. All the objective functions of these models were to minimize the transportation costs and fixed investment costs.

In the literature, many existing research studies for determining a multi-objective logistics distribution centre location have mainly focused on calculating the index weight of the influence factors by using particular approaches, such as barycentre theory, the osculating value method [10]. Barycentre theory is calculated on the minimum cost as the goal for best location, but the location is likely to be unable to use (e.g., in the middle of the river, streets, etc.). The osculating value method is to select a best solution from many available options under certain constraints. Its disadvantage is that if the situation is more complex, it is hard to establish a proper programming model.

In general, the overall goal these methods above was all transportation cost minimum or poor feasibility. This paper studied distribution centre's location selection of the internal supply chain in enterprise. Its main goal is to improve the distribution speed and reduce logistics intension, relatively low requirement on distribution cost, so the several methods above are not applicable for this study. Taking the confirmed factory layout into consideration, selection method of this paper is mainly to find an optimal location within a few eligible alternative locations in the factory area, so we just considered several special constraint conditions. Finally, the paper chose the analytic hierarchy process to score several alternatives locations, and evaluate it through the simulation.

2 Problem descriptions

In general, the internal supply chain is simple, only includes three parts: purchasing, product, and products sale. However, for large-scale manufacturing enterprises have big production scale, complex process and need a number of raw materials, spare parts and production

* ***Corresponding author*** email: jinmaozhu@scu.edu.cn

equipment, meanwhile, the complex process results in the logistics direction variously, And the large-scale logistics, such as production logistics and distribution logistics, are done internally, the scale of internal logistics is huge, the task of internal material supply logistics system is burdensome. In order to improve logistics efficiency and reduce delivery time, it is necessary to change the distribution mode, which is a passive mode to active distribution mode according to the plan.

Scientific and reasonable distribution centre's location plays a vital role for balancing the entire supply chain logistics, the logistics service reliability and management level of ascension. Appropriate location can effectively reduce the cost of supply chain, improve the overall efficiency of the supply chain and reduce the impact of the bullwhip effect. Therefore, the internal distribution centre's location is very important.

Distribution centre's location has some constraints: First, the site must locate inside the factory. For confidentiality reasons, the distribution centre construction within the scope of the company's existing plant area is very necessary. In addition, distribution centre locates inside the factory is convenient for unified management, real-time monitoring and real-time updates for logistics information can be more convenient, and also can shorten the distance of distribution, increase the speed of delivery. Second, the construction of the distribution centre cannot impact on the present layout. The change of present layout will damage the direction of logistics and improve the cost. Third, Transportation around the distribution centre is convenient. In order to facilitate the material transportation, roads surrounding the selected site should be built perfect, can lead to various professional plants. In addition, the surrounding roads can hold big enough traffic.

3 Evaluation indicators for selection of internal distribution centre's location

The internal supply chain distribution centre's location problem is different from the external distribution centre or the third party logistics distribution centre location problem, the selected location must be within the plant, only in accordance with the requirements of the warehouse construction alternatives in the screening, so when determining the selection indicator does not need to consider the cost of the land itself and its use (except for planning of land). And for the internal supply chain, logistics process is relatively fixed, so the relatively fixed on distribution route, too. The detail indicators are introduced as follows:

1) Supply speed: Special manufacturer requests, after their demand is sent out, distribution centre can deliver material in the shortest possible time. Especially when there is urgent demand, supply speed is particularly important.
2) Logistics intension: Logistics intension is the product of logistics capacity and the distance, Logistics intension of the selected location should as small as possible.
3) Degree of traffic convenience: Degree of traffic convenience directly affects the delivery speed.
4) Warehouse terrain: The site must is higher than the surrounding terrain and has good drainage, in case of a severe climate led to water.
5) Peripheral equipment: Mainly includes the roads surrounding site, communication and other public facilities complete, with plenty of power, water, gas, etc.
6) Size of location: Site must be large enough to meet the needs of its business development.

4 Case study

Company H is a large group aircraft manufacture enterprise. There exists large materials distribution among its warehouse and professional factories. These professional form factories an internal supply chain. However, its logistics system is not perfect and balanced due to the existing low level supply chain management. In order to improve the internal supply chain management, and realize the targets of process optimization, reducing costs, it becomes very important to construct an internal distribution centre.

4.1 DISTRIBUTION MANAGEMENT STATUS OF COMPANY H

Now raw materials and components of Company H are temporarily stored in the material warehouse. Once the production plan is issued, the warehouse will distribute all raw materials and spare parts to the temporary storage location according to the material requirements of each professional plant in a lump sum. However, the production plan is rough and some materials stack up to a long time. This often results in heavy inventory.

In such a distribution management, materials management and distribution of company H appeared to be very disorderly. Since distribution is unreasonable, the use of materials and logistics information cannot be effectively monitored.

4.1.1 Layout diagram

To select a good logistics distribution centre's location, we must know about the layout of company. Layout diagram of Company H is shown in Figure1. The shadow line represents rarely developed area and the solid line represents the existing workshop of each professional plant.

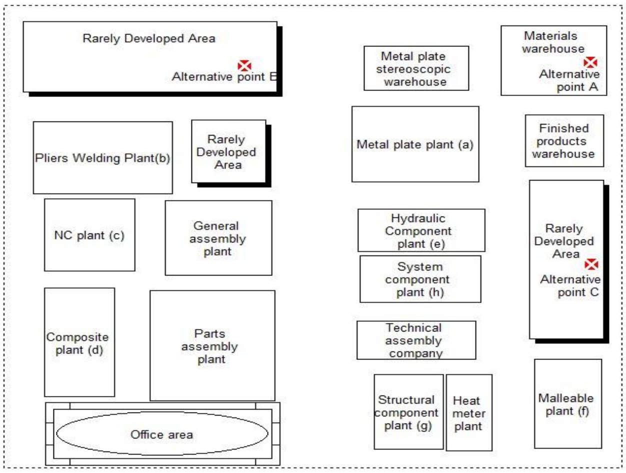


FIGURE 1 Layout diagram of Company H

4.1.2 Logistics processes

All required materials are distributed from the materials warehouse and their main distribution processes are shown in Figure 2.

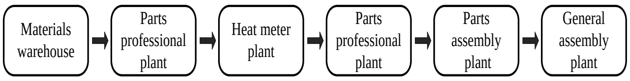


FIGURE 2 Materials main processes of Company H

4.1.3 Distribution condition of materials

According to the statistics data, the number and the average weight of parts and materials which are distributed to each professional plant from the materials warehouse are both shown in Table 1.

TABLE 1 Distribution condition from warehouse to professional plant/per year

Professional plant	Number of materials	Average weight of materials\kilo	Total weight of materials\kilo
Metal plate plant	38313	52.7	2020776
Pliers Welding Plant	46846	51.6	2429236
NC plant	6523	186.2	1214443
Composite plant	15700	68.1	1069840
Hydraulic Component plant	89459	12.7	1133690
Malleable plant	26440	30.5	805807
Structural component plant	23524	546.6	12858043
System component plant	583894	0.11	86228.34

4.2 SELECTION OF ALTERNATIVE LOCATIONS

4.2.1 Initially determine alternative locations

Selection criteria of alternative location is stated as follows: (a) It must be selected within the company; (b) Alternative location does not change the existing layout of the company; (c) The space is big enough to meet the storage requirement; (d) Surrounding transportation is convenient and material supply is convenient and quick; (e) It should be flat and has a good drainage system. Considering the five standards above, we determined three alternative locations initially which are signed as small rectangle in Figure 1.

Alternative location A: Namely, the existing warehouse. Do not need to move and the traffic is quite convenient. However, the space is small and it is unlikely to enlarge the warehouse area.

Alternative location B: The space is very large. B is close to the existing material so it is easy to move. Moreover, roads surrounding are developed and materials can be distributed quickly. However, the surrounding power facilities are penurious.

Alternative location C: The space is enough to use, but transportation around is not so convenient, and the terrain is not too flat. Besides, drainage facilities are poor.

4.2.2 Evaluation index weight calculation

Evaluation index has been written in Section3.Use AHP to calculate the evaluation index weight. Relevant tales are shown in Table 2-5.

TABLE 2 Classification of importance

Importance	Definition
1	A and B have the same importance
3	A is slightly important than B
5	A is important than B
7	A is much more important than B
9	A is extremely important than B

In order to reflect the importance of various factors more objectively and accurately, this paper adopted Delphi method. Select each logistics personnel respectively from 15 professional plants, 4 personnel who are responsible for the warehouse and distribution from Purchasing Department, and 1production department manager, a total of 20 people forma panel of experts. Let they fill out the importance of the various factors. Calculate the average value and get the factors comparison matrix of distribution centre location selection. Then normalize the matrix, and calculate the weight vector.

Let the factors comparison matrix of distribution centre location selection be B and weight vector be W, then

$$G = BW = (g_1, g_2, \ldots, g_n)\ . \qquad (1)$$

TABLE 3 Factors comparison matrix of distribution centre location selection

	Supply speed	**Logistics intension**	**Degree of traffic convenience**	**Warehouse terrain**	**Peripheral Equipment**	**Size of location**
Supply speed	1	3	3	5	7	3
Logistics intension	1/3	1	3	5	5	3
Degree of traffic convenience	1/3	1/3	1	3	5	3
Warehouse terrain	1/5	1/5	1/3	1	3	1/3
Peripheral Equipment	1/7	1/5	1/5	1/3	1	1/3

TABLE 4 Comparison matrix after normalization

	Supply speed	**Logistics intension**	**Degree of traffic convenience**	**Warehouse terrain**	**Peripheral Equipment**	**Size of location**	**Sort weight vector**
Supply speed	0.4268	0.5921	0.3814	0.2885	0.2917	0.2813	0.3769
Logistics intension	0.1423	0.1974	0.3814	0.2885	0.2083	0.2813	0.2498
Degree of traffic convenience	0.1423	0.0658	0.1271	0.1731	0.2083	0.2813	0.1663
Warehouse terrain	0.0854	0.0395	0.0424	0.0577	0.1250	0.0313	0.0635
Peripheral Equipment	0.0610	0.0395	0.0254	0.0192	0.0417	0.0313	0.0363
Size of location	0.1423	0.0658	0.0424	0.1731	0.1250	0.0938	0.1070

TABLE 5 Average random consistency table (RI)

Size of matrix	1	2	3	4	5	6	7	8
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41

Results can be calculated as follows after inquiring average random consistency table.

$$\lambda_{\max} = \frac{1}{n}\sum_{i=1}^{n} g_i / w_i = 6.4367 \,, \tag{2}$$

$$CI = \frac{\lambda_{\max} - n}{n-1} = 0.0873 \,, \tag{3}$$

$$CR = \frac{CI}{RI} = 0.0704 < 0.1 \,. \tag{4}$$

TABLE 6 Yearly Logistics intension of alternative locations

Professional plant	**Distance to A(meter)**	**Logistics intension to A(kilo*meter)**	**Distance to B(meter)**	**Logistics intension to B(kilo*meter)**	**Distance to C(meter)**	**Logistics intension to C(kilo*meter)**
Metal plate plant (a)	400	808310400	360	727479360	595	1202361720
Pliers Welding Plant (b)	890	2162020040	25	60730900	1570	3813900520
NC plant (c)	1680	2040264240	705	856182315	1180	1433042740
Composite plant (d)	2150	2300156000	1025	1096586000	1200	1283808000
Hydraulic component plant (e)	800	906952000	810	918288900	150	170053500
Malleable plant (f)	1430	1152304010	1630	1313465410	400	322322800
Structural component plant (g)	1340	17229777620	1420	18258421060	340	4371734620
System component plant (h)	850	73294089	860	74156372.4	145	12503109.3
Total	-	26673078399	-	23305310317	-	12609727009

It complies with the consistency judgment; therefore the judgment matrix can be thought to have satisfactory consistency. Thus the weight of each factor has also been determined, namely the importance of supply speed, logistics intension, degree of traffic convenience, warehouse terrain, peripheral equipment , size of location is respectively: 0.3769, 0.2498, 0. 1663,0.0635, 0.0363, 0.1070.

4.2.3 Score of alternative locations

Calculate logistics intension firstly. The logistics intension is the sum of logistics weight multiplied by the distance from alternative location to each plant.

From Table 6 we can see the logistics intension of alternative location C is the smallest and A is the largest. The logistics intension of alternative location A, B, C respectively occupies 42.6%, 37.2%, 20.2% of the sum of the three, and then the logistics intension score of A, B, C is respectively given 57.4、62.8、79.8.

Invite the panel of experts to grade the remaining five evaluation index to the alternative locations, full mark is 100 points and the score is the average value of the panel of experts scoring. Then calculate the total score of each alternative location according to the weight of each factor, which is shown in Table 7.

TABLE 7 Score of each alternative location (Full mark: 100)

Evaluation index	**Score of location A**	**Score of location B**	**Score of location C**
Supply speed(0.3769)	80.3	84.5	86.4
Logistics intension(0.2498)	57.4	62.8	79.8
Degree of traffic convenience(0. 1663)	86.7	92.4	79.4
Warehouse terrain(0.0635)	88.1	90.6	73.8
Peripheral equipment(0.0363)	84.9	80.3	81.3
Size of location(0.1070)	64.8	90.4	72.7
Total score	74.6	81.3	81.1

From Table 7 we find that alternative location B got the highest score, therefore select alternative location B to construct logistics distribution centre.

4.3 SIMULATION OF LOCATION SELECTION PLANNING

4.3.1 Simulation evaluation index

In order to evaluate the effect of after locating the distribution centre at B, simulate the processes running to the current distribution planning and the distribution centre respectively through Software Promodel. The goal of company H establishing distribution centre is to improve the distribution speed and reduce the logistics intensity. Therefore, distribution time and total logistics intensity are selected to be the simulation evaluation index.

4.3.2 The status quo simulation

Now raw materials and components of company H are temporarily stored in the material warehouse, shown in Figure 3. Once the production plan is issued, the warehouse will distribute all raw materials and spare parts to the temporary storage location according to the material requirements of each professional plant in a lump sum. Therefore, we collected part distribution record of materials warehouse in 2011 to carry on the simulation.

(1) Resource-transportation equipment. The warehouse of Company H has three kinds of transportation equipment-manual hydraulic carrier, internal combustion balance forklift and medium-sized truck. Their number is respectively 47, 35, 8, and maximum load is 2,8,18 (t). Carry on group distribution to the materials of each professional plant in the simulation, and the total weight of each group of materials must be less than the maximum load for the selected device.

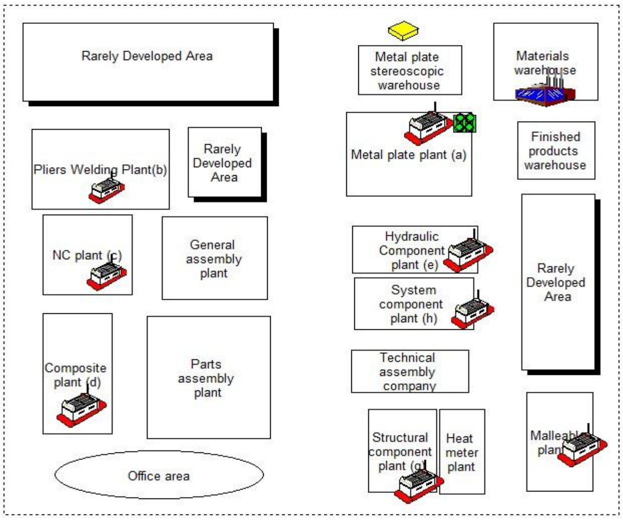


FIGURE 3 The status quo simulation model

(2) Relevant distribution data. Data related to distribution mainly includes distribution distance, quantity and weight of deliver materials, etc. This article only carried the simulation of the distribution from material warehouse to each professional plant, so the collected materials data are all the data from material warehouse to each professional plant, not including logistics information between various professional plants. It is shown in Table 8.

(3) Running the simulation model .With the data above, carry on the simulation according to the actual time of distribution tasks. Running process of the simulation model is shown as in the figure below.

TABLE 8 Relevant distribution data

Professional plant	Distance from material warehouse (meter)	Yearly average distribution material weight (Kg/year)	Average distribution material number/year (Actual data)	Average distribution material number/year (Simulation data)
Metal plate plant (a)	400	2020776	38313	3831
Pliers Welding Plant (b)	890	2247691	46846	4685
NC plant (c)	1680	1214443	6523	625
Composite plant (d)	2150	1069840	15700	1570
Hydraulic component plant (e)	800	1133690	89459	8946
Malleable plant (f)	1430	1065807	31440	3144
Structural component plant (g)	1340	16528043	33524	3352

4.3.3 Internal distribution centre simulation

Simulation data is based on the current plan. So the simulation of running effect for the distribution centre is based on the assumptions as follows: (a) all production plans and scheduling is precise and detailed. (b) The plan information and logistics information of distribution centre and each distribution point can be real-time monitored effectively. (c) All raw materials, spare parts, semi-finished products of distribution centre are enough, not existing out of stock problem. Setting up and running process of the simulation model is shown as in Figure 4.

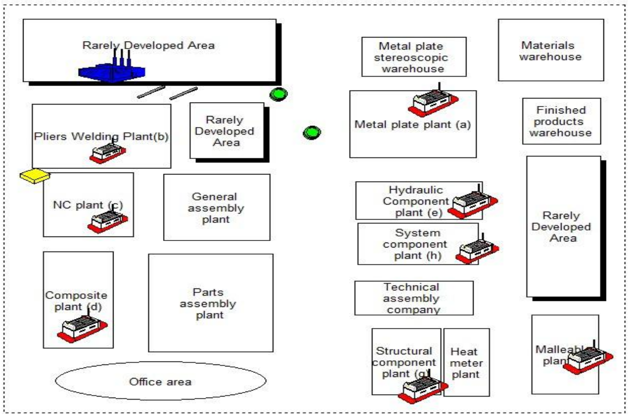


FIGURE 4 Internal distribution centre simulation model

4.4 SIMULATION RESULTS AND COMPARATIVE ANALYSIS

In order to guarantee the reliability of the simulation results, both simulation models run for five times and all the results are the average of five running results. According to the evaluation index, we select the following data for analysis.

(1) Conveyance\resource data

Table 9 and 10 show conveyance relevant data, which are respectively the running results of the status quo simulation and simulation of after constructing distribution centre. The Name, Units, Scheduled Time (MIN), Number Times Used, Average Time Per Usage (MIN), Total Time used (MIN), Utilization rate, Idle rate of each conveyance are all listed in the table above. After constructing a distribution centre at location B and adopting active distribution way of variable granularity, the usage of handling tools will be greatly reduced. What's more, the total use time of three kinds of transportation tools respectively shortened 18109 min, 5073 min and 42925 min.

TABLE 9 Resources for simulation-- The status quo (Avg. of 5 replications)

Name	Units	Scheduled Time	Number Times used	Avg. Time Per Usage	Utilization%
Pallet truck	47	5640000	18165.4	14.6	4.70%
Truck	8	960000	3692.6	6.31	2.43%
Forklift	35	4200000	1185.8	3549.14	91.75%

TABLE 10 Resources for simulation-- The status quo (Avg. of 5 replications)

Name	Units	Scheduled Time	Number Times Used	Avg. Time Per Usage	Utilization%
Pallet truck	47	5640000	9462.1	15.01	4.38%
Truck	8	960000	1872.4	6.35	1.90%
Forklift	35	4200000	385	4854.44	90.73%

(2) Logistics intension

Handling tools that all kinds of materials use are set in the simulation. Hydraulic parts and structural components use carts. Materials of Metal plate plant, Pliers Welding Plant and Composite plant use forklift. In addition, materials of NC plant, malleable plant and Structural component plant use truck. The speed of three handling tools is respectively set 50 m/min, 120 m/min and 300 m/min. According to the information, logistics intension can be calculated combining with the distribution time and weight of the materials in the tables.

The logistics intensity of distribution centre simulation is far more less than the logistics intensity of the status quo simulation from Table 11 and 12. The overall logistics intensity decreased by 17024145983 kg·m.

5 Conclusions

This paper studied distribution centre's location selection of the internal supply chain in enterprise. The selection method of this paper is mainly to find an optimal location within a few eligible alternative locations, the paper chose the analytic hierarchy process to score several alternatives locations, and selected the optional location. Then evaluated it through the simulation compared with the status quo.

TABLE 11 Logistics intension - The status quo simulation

Professional plant	Number	Time(min)	Speed(m/min)	Weight(kg)	Logistics intension (kg·m)
Metal plate plant (a)	3807	3.3	120	52.7	794490444
Pliers Welding Plant (b)	4672	7.38	120	51.6	2134961971
NC plant (c)	620	5.28	300	186.2	1828632960
Composite plant (d)	1557	17.88	120	68.1	2275016155
Hydraulic component plant (e)	8946	23.33	50	12.7	1325309643
Malleable plant (f)	3144	10.1	300	30.5	2905527600
Structural component plant (g)	3350	9.85	300	546.6	54109300500
System component plant (h)	58311	26.6	50	0.11	853089930

TABLE 12 Logistics intension - Internal distribution centre simulation

Professional plant	Number	Time(min)	Speed(m/min)	Weight(kg)	Logistics intension (kg·m)
Metal plate plant (a)	3822	11.8	120	52.7	2852098704
Pliers Welding Plant (b)	4674	15.01	120	51.6	4344105341
NC plant (c)	620	16.18	300	186.2	5603651760
Composite plant (d)	1563	37.77	120	68.1	4824300157
Hydraulic component plant (e)	8943	52.87	50	12.7	3002384204
Malleable plant (f)	3144	43.58	300	30.5	12536920080
Structural component plant (g)	3346	37.96	300	546.6	2.08278E+11
System component plant (h)	58343	37.49	50	0.11	1203003489

The results indicate that internal distribution centre construction can effectively shorten the delivery time, reduce the logistics intensity, and improve utilization rate of transport equipment. Therefore, the distribution centre's location selection is necessary and reasonable.

Acknowledgements

This work was funded by the National Science Foundation of China under Grant Number 71001075, 71131006 and 71020107027, and Central University Fund of Sichuan University No. skqy201112, and the National High Technology Research and Development Major Program of China (863 Program) (Grant No.2008AA04A107).

References

[1] Hodder J E, Dincer M C 1986 A multifactor model for international plant location and financing under uncertainty *Computer & operations research* **13** 610-9

[2] Stevenson W J 1993 *Production/operations management*

[3] Taniguchi, Noritake 1999 Optimal size and location planning of public logistics terminals *Transportation Research Part E* **35** 207-22

[4] Li H, Yang. J Q 2002 The application of fuzzy priority and heuristic algorithm in site choice of logistics centre *Transportation Science & Engineering* **26** 389-92

[5] Sun H J, Gao Z Y 2003 Bi-level programming model and solution algorithm for the location of logistics distribution centres based on the routing problem *China Journal of Highway and Transport* **16** 115-9 *(In Chinese)*

[6] Han L S, Li X H 2004 Multiperson and multicenter evaluation and selection of logistics centres with fuzzy analytic hierarchical process method *System Engineering Theory and Practice* **7** 128-34

[7] Aikens C H 1985 Facility location models for distribution planning *European Journal of Operations Research* **22** 263–79

[8] Holmberg 1995 Exact solution methods for incapacitated location problem with convex transportation costs *European Journal of Operations Research* **114** 127-40

[9] Baracoa F, Jensen D 1998 Plant location with minimum inventory *Math. Program* 83101–11

[10] Felix T S 2011 A study of distribution centre location based on the rough sets and interactive multi-objective fuzzy decision theory *Robotics and Computer-Integrated Manufacturing* **27** 426-33

Authors

Lifei Yao, born in December, 1988, Qinhuangdao, China

Current position, grades: A Graduate student
University studies: Sichuan University
Scientific interest: Industry Engineering , prediction, evaluation and decision control
Publications: 1
Experience: She is a Graduate student in Sichuan University for majoring in management science and engineering, and a member of Information and Business Management Institute of Sichuan University.

Ruimin Ma, born on September, 1989, Shaoyang, China

Current position, grades: A doctoral student
University studies: Sichuan University
Scientific interest: simulation, multi-objective decision, and vehicle scheduling
Publications: 2
Experience: He is a doctoral student in Sichuan University for majoring in management science and engineering, and a member of Information and Business Management Institute of Sichuan University. His researches mainly relate to simulation, multi-objective decision, and vehicle scheduling, etc.

Maozhu Jin, born on December, 1978, Nanchong, China

Current position, grades: Instructor of Business School, the tutor of MBA operations management and innovation and entrepreneurship management in Sichuan University
University studies: Huazhong University of Science and Technology
Scientific interest: operations management, organizational process reengineering, strategic management, service operations management, and platform-based mass customization.
Publications: 15
Experience: He has been engaged in the teaching of core curriculums such as operations management and management consulting. As a main researcher, he has participated in and completed three projects supported by National Natural Science Foundation of China and two surface projects. He has published two books and over ten research papers in authoritative journals of high quality both at home and abroad, and ten of them are retrieved by SCI and EI.

Peiyu Ren, born on December, 1952, Chongqing, China

Current position, grades: Professor, PhD Supervisor, currently acting as the Director of Information and Business Management Research Institute of Sichuan University.
University studies: Sichuan University
Scientific interest: Industrial Enterprise Management, scenic area management and Informatization integration management
Publications: 120
Experience: He has presided over and completed five surface projects of National Natural Science Foundation of China, being in charge of project research of Projects 863, 985 and 211, having published 15 books, monographs and more than 100 academic papers, including SCI, EI and CSSCI

Exploring Dynamic Performance improvement in Service SCM: the Lean Six Sigma's perspective

Ouyang Fang[1]*, Chih-hung Hsu[2]

[1] *Economic and Trade Department, Suzhou Institute of Industry Technology, China*

[2] *Industry Engineering, Hsiuping University of Science and Technology*

Received 1 March 2014, www.tsi.lv

Abstract

This paper defines the performance evaluation system of Service SCM. As service is intangible and heterogeneous, the paper is to develop a model that illustrates under which conditions Lean Six Sigma is deemed most appropriate according to the type of service delivered. It investigate Lean Six Sigma practice in service supply chain and show how the Lean Six Sigma improve the performance of Service SCM from the statistics perspective. Furthermore, it stresses the CTQ (critical to quality) to the customer and clarifying their demands in terms of value-added requirements.

Keywords: Service Supply Chain Management, Intangible, Lean Six Sigma

1 Introduction

Research in supply chain management (refer to SCM) is approached from different disciplines: logistics, operation management, distribution etc., which related to the physical movement of goods (tangible products) and the related Information flow, Business processes and Capital flow. However, SCM is also relevant for services (intangible products) (Ellram et al., 2004). The notion of Service SCM has garnered increased awareness in SCM field and has been realized the importance within the organization. Servitisation is even predicted as being a future significant research area within operations management (Taylor and Taylor, 2009). SCM in a service context is, like SCM in general, related with designing and managing supply chains, controlling its assets and uncertainties in order to meet the needs of the customers in a cost-effective manner (Ellram et al., 2004).

Service SCM has been defined in a way that differentiates it from a traditional SCM manufacturing centric focus. Ellram et al. (2004) define Service SCM as: “the management of information, processes, capacity, service performance and funds from the earliest supplier to the ultimate customer.” An important message in SCM is that a differentiation of tasks should take place. Such a differentiation can, for example, be practiced through different types of relationships with customers, as well as suppliers. The Service SCM framework was normally portrayed to seven service process (Ellram et al., 2004):

(a) Information flow (e.g. collaboration with customers and suppliers and information sharing).

(b) Capacity and skills management (e.g. capability to satisfy the customer, quick response to the market).

(c) Customer relationship management (CRM) (e.g. customer service and opportunity analysis on winning customer).

(d) Supplier relationship management (SRM) (e.g. supplier segmentation, supplier audit, supplier assessment, supplier selection).

(e) Service delivery management (e.g. service coordination on the delivery to customer, enabling service providers).

(f) Cash flow (e.g. flow of payments between parties).

(g) Demand management (e.g. forecasting market requirements).

Currently the functions of Service SCM have been of great strategic extension in the breadth and depth of field, this include the types of Service project, Service objects, and Service area. The difficulty of managing Service Supply Chain has been increased due to the Cross-border diversification regulations, differences individual needs, longer delivery time, increasing transport costs etc., the characteristics of Service SCM are to face a large number of customers, extensive customization requirements, while the fulfilment is mostly reply on external service conditions.

While the pure Services are intangible, and have a quality dimension, which is difficult to evaluate .Service quality evaluation has been critical for the Service SCM. The purpose of this paper is to find out the service assessment index and introduce the Lean Six Sigma practice to improve it in a Service SCM context. More specifically, the paper aims to answer two questions

RQ1. How to define the KPIs in the assessment of Service SCM?

RQ2. How Lean Six Sigma solution to improve the performance of Service SCM?

* ***Corresponding author*** e-mail: ouyangf@siit.edu.cn

2 Performance evaluation system of Service SCM

A. The principles of Service performance evaluation

The quality of Service supply chain is a primary factor to indicate if the organization has the ability to create time and space effectiveness of the scale to customer, to retain existing customers and attract new customers. Hence, another problem is how to measure the quality of Service supply chain. Although the concept of SCM has been developed for more than 30 years and there are many studies on performance evaluation of supply chain, there is lack of definition on Service supply chain. Thus, performance evaluation of Service supply chain is not defined consistently. The content is still incomplete and not systematic. It is an issue worthy of further study (Atkinson, 2004). For the Construction of performance evaluation system of Service supply chain, it should be based on overall strategy of Service SCM and aims to establish balance between short-term and long-term goals, financial and non-financial performance measures and internal and external performance compositions. Moreover, performance evaluation system should follow the principles as below (Youngdah, 2000):

a) Importance

The measures should be divided into different degrees. From measures of each degree, key points of evaluation are selected to analyse key performance evaluation measures.

b) Dynamics

Dynamic evaluation which reflects business process of Service supply chain is adopted and it is not limited to examination of static operating outcome.

c) Completeness

The measures can reflect operation of overall Service supply chain instead of operation of only one node.

d) Immediateness

Immediate evaluation and analysis can demonstrate immediate operation of Service supply chain. It is more valuable than after analysis.

e) General principle

It values evaluation on long-term benefit and long-term potential of Service supply chain.

f) Comparative principle

The evaluation system selected can be compared in terms of time. Moreover, it can horizontally compare different supply chains in the same industry.

g) Quantitative principle

The measures include quantitative and qualitative ones. However, qualitative measures should be quantified for comparison among different supply chains.

h) Economic principle

Scope of evaluation system should be appropriate. With too many measures, the evaluation will be difficult. Few measures will not reflect the performance of supply chain. Moreover, acquisition of measures should be economic and convenient.

B. The KPIs concerned with Service SCM

In literature, three scholars Parasuraman, Zeithaml and Berry (referred to as PZB) made the most representative assessment of service quality, PZB three scholars believe that the customer perception of the "Quality of Service" (Service Quality, abbreviated as SERVICEQAL) by the "expectations of service" and "cognitive services", the difference between the size and direction of the joint decision made by the five services assessed differences in quality of service constitute the general model, that "Tangibility", "Reliability", "Responsiveness", "Assurance", "Empathy" in five areas of service quality assessment table. It is the famous SERVICEQAL scale. In view of above principles and the specifications of Service SCM, the paper list the related KPIs as shown in (table 1, the answer of the RQ1).

TABLE 1 Related kpis in Service SCM

	Evaluate factors	Quantitative index(example)
Tangibility	Distribution process ability	Deal with the special demand (frequency)
	Order receiving	Reject customer order (times and amount)
	Transport capability	Transport speed and cost advantage
	Storage capability	Scrap amount
Reliability	On-time -delivery	OTD%
	Fulfil customer order capability	Satisfy the expedition and postpone requirement (times)
	Good intact rate	Damage %, customer return%
	Delivery error rate	Error %
Responsiveness	Response time	Reply customer within 24hrs
	Fulfil customer order speed	Lead time (purchasing time, production time, transporting time)
	Tracking system consistency	Missing information %
Assurance	Reputation	Industry authority ranking
	satisfaction about Service attitude	3'rd party satisfaction survey
	Employee trust degree	3'rd party satisfaction survey
Empathy	Satisfy the personal demand	Deal with the special demand (types and times)
	Awareness about customer demand	3'rd party satisfaction survey

It is important to improve the efficiency of measurement system as process capability improves; evaluate the use of control measurement systems (e.g., attributes, variables, destructive), and ensure that measurement capability is sufficient to evaluate service system. Performance evaluation system can tell how the overall Service SCM working timely and guide the organization to achieve the business goal.

3 Lean Six Sigma summary

Lean is about eliminating waste and creating customer value and consists of principles that constitute the backbone of the philosophy. The present review of the lean concept has identified that lean is concerned with waste reduction and value creation for ultimate customer. In addition to the diverse conceptualizations of the lean concept, σ is a statistical term for standard deviation, Six

Sigma's strengths is the ability to transfer a practical service problems into a statistical analysis problem, resulting in a statistical solution, then converted back to a practical solution through the DMAIC process that is: D-definition (Define), M-measurement (Measure), A-analysis (Analyse), I-improvement (Improve) and C-control (Control) to improve the existing Service SCM processes. DMAIC is a process cycle which can achieve to serve the customer as the "centre", continue to improve customer satisfaction and link the Service SCM closely to business objectives, it emphasis full use of quantitative analysis and statistical thinking .Six Sigma is a measurement scale upon which improvements can be gauged.

There is supreme advantage that can be gained from integrating Lean with Six Sigma. Lean is primarily about reducing waste and Six Sigma can provide certain problem solving abilities to waste elimination .Lean Six Sigma is an overall methodology, which aim to the continuous improvement. It was always taken into account the service quality during the implementation of Lean Six Sigma project ,as well as the efficiency and effectiveness, it obsolesce the interference rate, to get rid of the waste factors, and move out the Non-value-added process, it conduct a statistical analysis through numerous practical information and data, to dig out and break the outmoded ideas and explain the real changes in the new results, Finally it provide great support for innovative solutions of Service SCM.

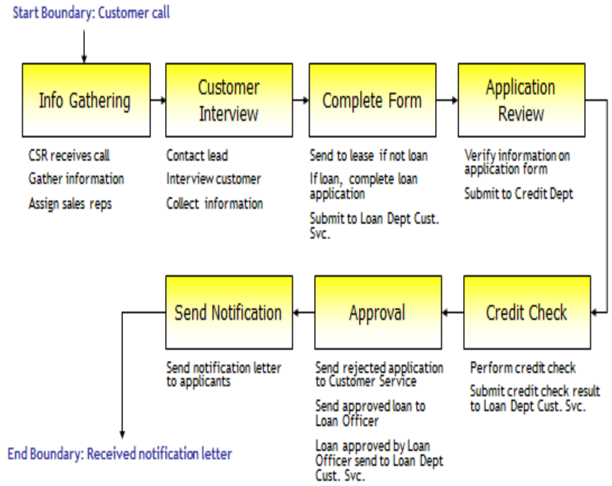
As a result, Lean Six Sigma should be fit well with the Service SCM concept since they are both concerned with creating customer value through cost-effective processes

4 Lean Six Sigma implementation in service supply chain

ABC Savings and Loans Bank is currently the 4th largest bank and plan to be the 2nd largest one in China in coming 2 years. Management studied their markets and determined that cycle time for loans and leases is a key competitive issue in all markets. The target cycle time is 8 days. While current process time is 9.2 days, Management decided to immediately attack this issue to alleviate both customer dissatisfaction and significant financial loss to the company. They made an agreement to introduce Lean Six Sigma to improve this service time.

The Lean Six Sigma team firstly plans to analyse the current application process. The best tool to analyse the process is an "as is" functional deployment map. They started by generating a SIPOC ((Suppliers, Inputs, Process, Outputs, Customers) diagram, followed by a top-down chart and completed the mapping with a functional deployment map for loan and lease processes respectively. It was shown in Table 2

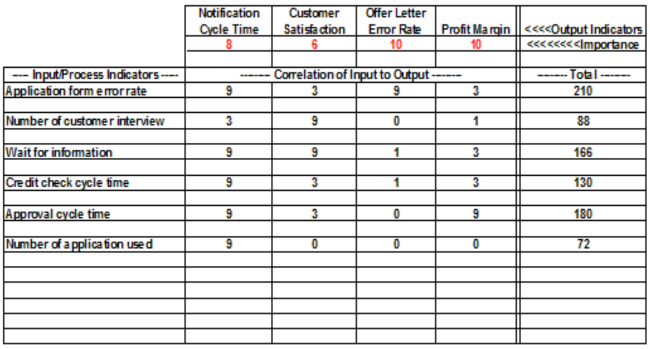
TABLE 2 Top-down chart


Start Boundary: Customer call
Info Gathering
CSR receives call
Gather information
Assign sales reps
Customer Interview
Contact lead
Interview customer
Collect information
Complete Form
Send to lease if not loan
If loan, complete loan application
Submit to Loan Dept Cust. Svc.
Application Review
Verify information on application form
Submit to Credit Dept
Credit Check
Perform credit check
Submit credit check result to Loan Dept Cust. Svc.
Approval
Send rejected application to Customer Service
Send approved loan to Loan Officer
Loan approved by Loan Officer send to Loan Dept Cust. Svc.
Send Notification
Send notification letter to applicants
End Boundary: Received notification letter

With the improved loan process, the team needed to know whether it was feasible to implement the process quickly. Collectively, the team used 5 criteria to determine whether the improved loan process is a "quick win" opportunity. Since the improved loan process is not a "quick win", the team started to work with the "as is" loan process. The next step for the team was to brainstorm and work out the VOC (Voice Of Customer) and VOB (Voice Of Business). This was done to verify ability of current "as is" process in meeting critical customer and business requirements. The team used KANO analysis and C&E matrix to prioritize the CCR (Critical Customer Requirement) and CBR (Critical Business Requirement) obtained from VOC/VOB brainstorming session. The team needed to collect data to determine the baseline cycle time performance of the loan and lease processes.

To be effective, the team wanted to ensure that data for the important input and process indicators was collected simultaneously. They used the SIPOC diagram, fishbone diagram and C&E matrix to establish the relationship between input, process and output indicators.

TABLE 3 Cause & Effect Matrix for Loan Process.



	Notification Cycle Time	Customer Satisfaction	Offer Letter Error Rate	Profit Margin	<<<<Output Indicators
	8	6	10	10	<<<<<<<<Importance
--- Input/Process Indicators ----	------ Correlation of Input to Output ------				------ Total ------
Application form error rate	9	3	9	3	210
Number of customer interview	3	9	0	1	88
Wait for information	9	9	1	3	166
Credit check cycle time	9	3	1	3	130
Approval cycle time	9	3	0	9	180
Number of application used	9	0	0	0	72

SCALE : 0=NONE 1=LOW 3=MODERATE 9=STRONG

After collecting the numerous data about the process time, the team decided that the Pareto chart was selected to start problem identification; the Control chart was used to check for process stability and to identify trends; and Graphical summary was generated to ensure that the team was dealing with not more than one population, and to determine normality of the data distribution.

The team used a scatter plot to view potential correlation between variables. The challenge is how to fix the new process control system, it was created by a team responsible for assuring the quality of a commercial loan process for loans <$1,000,000.

The two customer expectations being tracked were:

Prompt notification of loan approval or non-approval (CCR #1: <24 hrs from submission of application

Timely availability of funds (CCR #2: within 1/2 business day from notification of loan approval)

The duration of various key activities in the loan process are tracked in order to maintain performance at acceptable levels.

Additionally, the team gathers data on incomplete loan applications and rejection reason codes to help in the analysis of loan process cycle time. Outcome Indicator

O1 = loan submission to notify O2 = notification to disbursement

Process Customer Branch & District Commercial Loan Applicants for loans $1,000,000 Critical Customer Requirements:

a) 24 hour (M-S) response time on loan approval

b) Funds dispersed within ½ business day upon approval notification

A process control system organizes relevant information about a process in a meaningful and useable format. The information included in the process control chart may represent the effort of weeks, months and even years of process and customer experience and data collection. Once the right process measures, etc., are validated, a team can gather and organize the information into a process control system in a relatively short amount of time.

The Cycle Time was finally improved to 3.38Ds; Sigma was improved to 3.11 as shown in Figure1

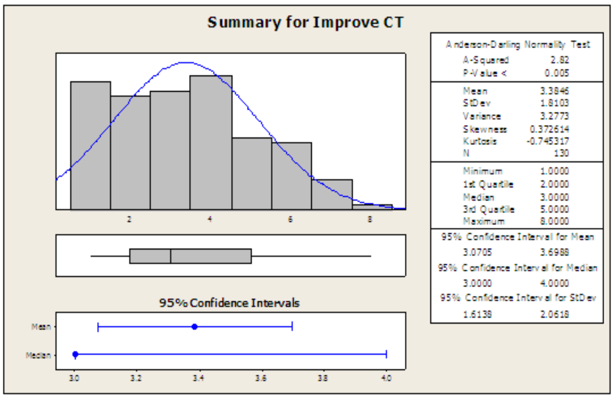


FIGURE 1 Graphical summary post-solution cycle time distribution

And the related whole process control map was shown as FIGURE 2

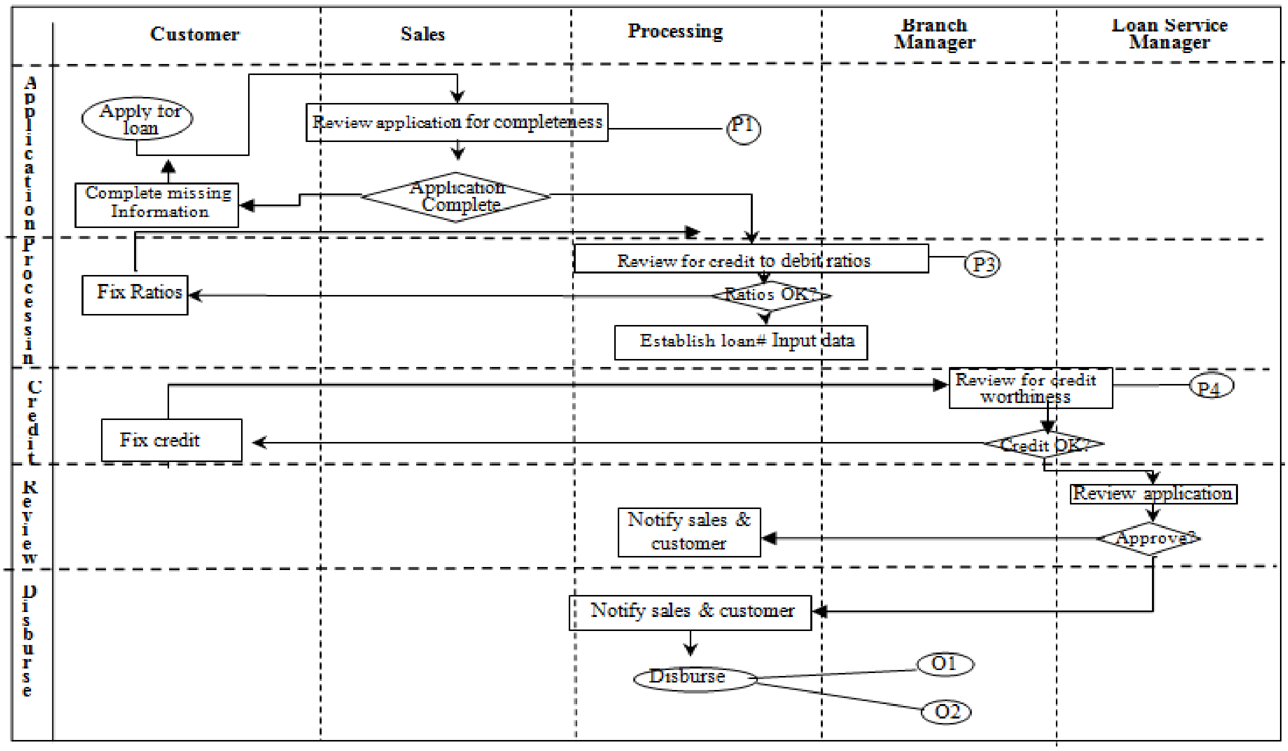


FIGURE 2 Process control map

Using the data collected from 3 sites (Beijing, Shanghai, Guangzhou), the Six Sigma team in Sigma Savings and Loans calculated the DPMO (Defects Per Million Opportunities) and the sigma level using the discrete notification cycle times:

D = 921
N = 2020

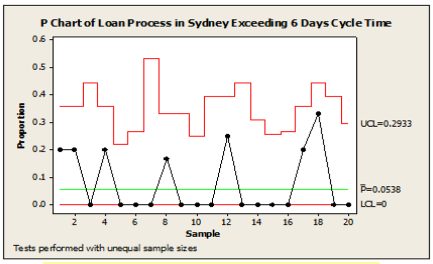
O = 1 (There is only one opportunity for a defect per application. Either the notification is delivered within 8 days or it is a defect.)

DPMO = 921 * (10)^{6} /2020*1=455,941

Sigma Quality Level is approximately 1.61

The team prioritized potential root causes: Wait for information because no guideline was given for information compilation; Long approval time because managers were busy and not confident that subordinates were capable of evaluating application effectively; Location issue due to IT protocol problems, resulting from different technology platform.

An improvement team recommended the use of criteria that significantly reduced the number of approvals by providing new loan representatives with the same tools that management used for loan application reviews, and the process was finally shorten to around 6days as shown in FIGURE 3



Sigma for improved loan process = 3.11

FIGURE3 Control chart post-solution process performance

The effort has been synthesizing Service SCM and six sigma and developing a unique Lean Six Sigma based methodology to improve service process time. The process control check items were shown as Table 4.

TABLE 4 Process control check items

Indicators	Control Limits and/or specs.	Checking Item	Checking Frequency	Responsibility	Actions	Misc. Information
P1 – activity duration, min.	≤ 5 minutes for all loan types	Time stamp, in and out	All loans on receipt	Branch sales representatives	Call customer Complete and validate applications Items 4,5,6,8 & 9	
P2 - # of incomplete loan applications	N = number of defects	All loans record on travel log	All loans on receipt	Branch sales representative	Call branch sales Reponses for all ratios above 0.8	
P3 – activity duration, hrs	≤ 5 minutes for loans ≥ $100k,\;≤20 minutes for loans > $100k	Time stamp, in and out on log		Processing clerk		See branch policy variations on ratios
P4 – type & reason for application rejection		Reason code sheet and log	All loans record on log	Branch Manager		
O1 – loan submission to notify	≤ 5 minutes for loans > $500k	Time stamp, in and out on log	Only loans > $500k	Loan Service Manager		District centre Service manager only reviews loans > $5600k
O2 – notification to disbursement		Reason code sheet and log		Branch sales representative		

The above case demonstrate the use of Lean Six Sigma mechanisms to improve the process of service supply chain which create value-added to customer, A combined management approach for Lean and Six Sigma will both accelerate the improvements achieved and will make a significant difference to the financial performance of the improvement program and the business.(the answer of the RQ2).

5 Lean Six Sigma to improve the performance of Service SCM

The case has set out to investigate Lean Six Sigma practices in the ABC Savings and Loans Bank through DMAIC method and to elaborate on whether it makes sense to apply the Lean Six Sigma concept in a Service SCM context.

Service SCM has its focus on the objectives of both service improvements and cost reduction with the purpose of providing the customer with the best possible service. The theoretical description of Lean Six Sigma has outlined a number of characteristics for being lean, i.e. customer focus, flow production and standardization of processes.

Lean Six Sigma is analytic tools and a disciplined, standardized methodology for their use.it Integrated approach to leading improvement efforts in Service SCM, Six Sigma is also the principles of leadership and Driving results through engaged teams to improve the whole Service Supply Chain process continuously. The Service process may include: Customer requirements, Process alignment, Analytical rigor, Timely execution and etc. Service SCM aims to deliver the required service in the most cost-effective way. Differentiation is an Important determinant in developing different types of relationships with customers, as well as suppliers, through CRM and SRM. Some customers demand special services requiring a special setup to make tailor-made solutions and to fulfil flexibility and lead-time requirements. Other customers may be satisfied with standard services. On the supply side, there may be arm's length relationships based on market prices for commodity products and more strategic partnerships based on trust with some suppliers in which sensitive information about.

Service SCM is always impacted by the market environment, cultural environment, policy-oriented, intra-industry and etc., the relationship among various factors is complex and uncertainties related to these complex systems of non-linear, this increases the complexity of the Services SCM. The "service-efficient" service strategy is deemed appropriate when the services offered are heterogeneous and thus require customized "production" processes. For example, heterogeneous services are characterized by its unpredictability of demand and time consumption to deliver the service and the actual resource spending.

6 Conclusions

Using the Lean Six Sigma processes, this study found several areas for improvement in the process studied. Bottlenecks and variation in the process were identified through in-depth analysis of the statistical information. Complex processes, no matter how finely tuned, have areas of bottlenecks or consistent delays, which can be identified and resolved through the Six Sigma process.

The resolution of indicated delays will invariably result in a decrease in variation, and ultimately, improved service to customers. The improved service will ultimately translate into improved profitability

Who you serve what services you provide and how you are going to achieve competitive advantage is the mission of Service SCM whose vision is to deliver added-value to its customers. Service SCM is always in a dynamic, complex and rapidly changing market environment, Lean Six Sigma is strategic management tool, which be able to fit in the Service SCM and improve the performance from the basis .Only the "zero error" concept is rooted in the quality of Service SCM, can the organization achieve both its development and sustainability goals.

Acknowledgments

This work was supported in part by the China Federation of Logistics and Purchasing (No: 2014CSLKT3-119) and Suzhou Social Sciences Union (No: 2014LX087).

References

[1] Abdi F, Shavarini S K, Hoseini S M S 2006 Glean lean: how to use lean approach in service industries *Journal of Services Research* **6**(special issue) 191-206

[2] Stentoft Arlbjørn J, Vagn Freytag P, de Haas H 2011 Service supply chain management -A survey of lean application in the municipal sector *International Journal of Physical Distribution & Logistics Management* **41**(3) 277-95

[3] Bendel R B, Afifi A A 1987 Comparison of stopping rules in forward stepwise regression *J. Amer. Stat. Assoc.* 46-53

[4] Atkinson P 2004 Creating and implementing lean strategies *Management Services* **48**(2) 18-33

[5] Ellram L M, Tate W L, Billington C 2004 Understanding service supply chain management *The Journal of Supply Chain Management* **40**(4) 17-32

[6] Youngdahl W E, Loomba A P S 2000 Service-driven global supply chains *International Journal of Service Industry Management* **11**(4) 329-47

[7] Becerra-Fernandez I, Zanakis S H, Alczak S 2002 Knowledge Discovery Techniques for Predicting Country Investment Risk *Computer and Industrial Engineering* **43** 787-800

[8] Hammer M 2002 Process management and the future of six sigma *MIT Sloan Management Review* 43(2) 26-32

[9] Faisal M N, Banwet D K, Shankar R 2006 Mapping supply chains on risk and customer sensitivity dimensions *Industrial Management & Data Systems* **106**(6) 878-95

[10] Antony J, Banuelas R 2002 Ingredients for the effective implementation of six sigma program *Measuring Business Excellence* **6**(4) 20-7

Authors

Ouyang Fang, born on April 11, 1973, Jiangxi Province

Current position, grades: Disciplines of Logistics Leaders/Associate Professor
University studies: MBA/The University of Southern Queensland (USQ)
Scientific interest: Supply Chain Management
Publications: (articles: 11; projects: 3 (with enterprise)/5 (with government)): The Construction Research on the Organization to Achieve three - win of the Green Procurement System; Revelation East Asian countries on China's logistics development planning; Asia logistics planning comparison and experience reference
Experience:

1995.7-1998.2	Makita (China) Co.ltd / Planner
1998.2-2003.7	Volex Interconnect Systems (Suzhou) Co.ltd / Master Scheduler
2003.8-2004.2	Fairchild Semiconductor (Suzhou) Co.ltd/ Logistics Supervisor
2004.2-2006.8	The University of Southern Queensland (USQ)-Master degree reading)
2006.8-2008.7	II-VI Optics (Suzhou) Co.ltd / SCM manager
2008.7-present	Suzhou Institute of Industry Technology/ Disciplines of Logistics Leaders

Chih-hung Hsu, born on September 21, 1971, Taiwan

Current position, grades: The Disciplines of Industry Engineering Leaders/Professor
University studies: Taiwan University
Scientific interest: Industry Engineering; Supply Chain Management, It include optimization concepts applied to various aspects of global supply chain management, information systems, technology management, product and process innovation, quality engineering and capital investment justifications.
Publications: (articles: 132; projects: 7 (with enterprise)/27 (with government)) Data Mining QFD for The Dynamic Forecasting of Life Cycle under Green Supply Chain; Integrating Grey Theory into Kano's QFD Based on Data Mining to Enhance Supply Market Survey with Purchasing
Experience:
1993-present Hsiuping University of Science and Technology,
2007-2011 Taiwan University (PhD reading)

The study of urban traffic modal splitting method based on MD model under the low-carbon mode

Shimei Wu*, Yulong Pei, Guozhu Cheng

School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150090, China

Received 1 March 2014, www.tsi.lv

Abstract

Aimed at the problem of overly-simplify in the factors of travel cost in the traffic modal splitting method, built an Urban Traffic Modal Splitting Method Based on MD Model Under the Low-carbon Mode, to predict transportation share rate; Put forward four considerations such as the travel time, cost, safety and low carbon to describe the travel cost on the basis of the application of MD model; Gave the forecasting process of the prediction model and key variables algorithm, applied the model by the examples of DONGGUAN city. The results show that the urban structure of the transportation changed in DONGGUAN with rapid construction, development in traffic and implementation of transport policy, on the one hand, the travel occupies proportion of public transport (including conventional bus and rail transit) will increase significantly in the future, expected to reach more than 25% by 2020; on the other hand, motorcycle travel will gradually fade away.

Keywords: Transportation Planning, Low-carbon Transportation, MD model, traffic modal split

1 Introduction

Along with China's rapid economic development, our motorization process is accelerating. According to National Bureau of Statistics, the holding number of our civilian vehicles has already reached to 120,890,000 by the end of 2012. However, when our motorization process develops rapidly, it also brings a series of problems, such as traffic safety, environmental pollution, traffic congestion and energy shortage. The study on travel mode choice can be traced back to the development of the residents travel survey. With the analysis of the survey data on residents travel, more and more people begin to study residents travel, such as the research on the travel intentions, the influence factors of travel mode choice and travel mode prediction model. The study on the travel mode split makes the study on the urban road planning develop from three-stage to four-stage, which greatly promotes the study on modern urban road planning, and has become an important part of the four-stage method study on the urban road planning. Therefore, the study on the travel mode split has significant meaning for the urban road planning, and it is necessary to strengthen the research on this area.

The splitting models of the transportation means are mainly divided into two categories: one is the Aggregate Model based on the statistics, which includes the transfer curve model, the gravity model's transformation model and regression model. This model is simple and convenient with better application effect, but it needs to investigate a large number of statistical materials and has huge amount of work with difficult data collection [1-4]. Another is the Disaggregate Model based on the "random utility theory", which attempts to conduct modelling to the travellers in the possible choices on the problems, such as whether the individuals as the traffic behaviour decision unit travel, where to go, and what kind of transportation do they use and which path do they choose. Then it sums the sampling results of each person, and thus obtains the total traffic demand. Disaggregate model is the discrete choice model based on the maximum effect theory, among which the models with more application are the MNL model (Multinomial Logit Model), Probit model, NL model (Nested Logit Model) and Mixed Logit model [5-9].

Conventional transportation splitting forecasting model has the following defects:

1) Conventional transportation forecasting model does not consider low carbon travel chain

In the conventional mode choice model, traffic sharing rate usually chooses the Logit probability model, which distinguishes various means of transportation without considering the influence of the transportation that generates in the connection between various means of transportation on the residents travel mode choice in the actual traffic travel.

2) Conventional transportation forecasting model does not consider the extra charges caused by carbon emission

In the travel mode choice model, either aggregate model or disaggregate model generally considers the factors, such as travel time and costs in the travel fees, but under the broad environment, such as the global requirement of reducing the carbon emission and national

* ***Corresponding author*** e-mail: 4770260@qq.com

strict control on the carbon emission in various industries, the additional costs caused by carbon emission are increasingly ignored. Adding carbon factor in the generalized travel costs will well control the carbon emission in the transportation in order to achieve a low-carbon transport requirement.

Aiming at the defects of the conventional mode choice model in the low carbon mode, it improves the travel mode choice model in the low-carbon mode based on that the four-stage prediction method introduces the low carbon factor and safety factor when building the modal split.

2 Basic principle of the MD model based on the low carbon mode

The basic principle of MD model main expresses by using the following concepts and assumptions [10]:

(1) Potential passenger transport demand Q_{ij}

Potential passenger transport demand means the total number of travellers of OD in all the travel possibilities between i to j, including the actual OD and possible generated OD, which does not consider the travellers' payment capacity, nor the ultimate realization of this travel need.

(2) Travel expense quantity S_{mij}

Travel expense quantity can also be called generalized cost, which means the travellers' money, time and energy consuming on the road. In the MD model, it assumes that travellers always choose the means of transportation on the principle of minimum travel expense, and consider the travel time, cost, safety, and low-carbon and other factors all have an impact on travel behaviour. The travel expense quantity indicated by logarithm shows as follows (curve relationship in Figure 1):

$$\ln(S_{mij}) = \ln\left[D_{mij}^{-1}(C_{mij} + vT_{mij} + Ca_{mij})\right]. \tag{1}$$

In this formula, S_{mij}, C_{mij} and T_{mij} respectively show the travel expense quantity, travel costs and travel time of OD to the means of transportation m between *i* to *j*; *v* is the value of travel time, that is the coefficient which converts the time unit into the monetary unit, and different types of travellers have different time value. This type mainly sets the income level and trip purpose as the discriminate criterion (in the MD model, it assumes that the time value conforms to the lognormal distribution, and the distribution parameters change over time); Ca- carbon factors, $Ca_m = v_c A_m$. In this formula, A_m is the m means of transportation's CO2 emissions per capital in per unit of time; v_c means the unit value of carbon; is the safety factor of the m means of transportation between the regions *i* to region *j* [11].

Security is the first factor in the travel choice. If the security of the travel mode cannot be guaranteed, this kind of travel mode will not be selected, so security can be considered to be independent with other factors. The D_{mij} safety factor is to use the numerical value between 0 ~ 1 to qualitatively characterize the safety, in which the smaller the numerical value is, the worse the security is. In travel expense quantity, An Wenjuan and other people use the reciprocal of the safety factor to represent the "security costs" that the residents choose to pay for some travel modes, and point out that safety, the larger the safety factor is, the smaller the travel expense quantity of this travel mode is.

(3) Travel utility u

In the MD model, travel utility does not consider the net effect of travel expense quantity is the benefit that the travellers expect to get regardless of the travel cost, so the travellers with different individual characteristics (mainly the travel purposes), their travel utilities are also different. The same as the value of travel time, MD model assumes that the travel utility also confirms to the lognormal distribution, and its distribution parameters do not change over time; meanwhile, it assumes that the standard deviation of travel utility equals to the standard deviation of travel time value.

(4) Boundary replacement ratio $v_{m,m-1}$

Figure 1 takes 5 means of transportation as example to depict the expense quantity curve of each mode of transportation, and the point that the travel expense quantities equal between transportation mode m and m-1 is the intersection of two curves, which is called boundary replacement ratio or demarcation point $v_{m,m-1}$, and can be derived by equation (1), see equation (2):

$$\ln v_{m,m-1} = \ln\frac{D_{(m-1)ij}(C_{mij} + R_{mij}^{m}) - D_{mij}(C_{(m-1)ij} + R_{(m-1)ij}^{m})}{D_m T_{(m-1)ij} - D_{(m-1)ij} T_{mij}}. \tag{2}$$

In this formula, $v_{m,m-1}$ is the boundary replacement ratio of transportation mode m and its nearest transportation mode m-1; others are as the above.

On the boundary replacement ratio, the expense quantities are equal on adopting the m transportation mode and the nearest m-1 transportation mode, that is:

$$C_{mij} + v_{m,m-1}T_{mij} = C_{(m-1)ij} + v_{m,m-1}T_{(m-1)ij}. \tag{3}$$

According to the equation (2) of the boundary replacement ratio, it can reckon the time value section of different means of transportation, and calculate the selected portion of each mode.

(5) Manifest rate of potential demand R_{mij}

In the MD model, according to its basic assumption, the travel utility and time value of the travellers are uncertain. Travellers choose the transportation mode with minimum travel expense by considering their own time value. Only when travellers' travel utility is greater than the expense quantity of the selected travel mode, will the travel demand be implemented. In the MD model, after the implementation, the potential passenger transport demand quantity is the actual demand quantity, and the ratio between these two is called the manifest rate of potential demand:

$$R_{mij} = \frac{q_{mij}}{Q_{ij}} . \tag{4}$$

In this formula, R_{mij} is OD's manifest rate of potential demand on the transportation mode between *i* to *j*; q_{mij} and Q_{ij} are respectively the actual and potential travel demand quantity of OD to the transportation mode between *i* to *j*.

When the travel expense quantity of the transportation mode m is the minimum expense quantity, it will be chosen by travellers; and only when the travellers' travel utility is greater than the expense quantity of transportation mode m, the travel will be achieved. Therefore, the manifest rate of potential demand can be obtained by calculating the volume of two probability distribution points of the time value and utility, shown as the formula (5).

$$R_{mij} = \int_{\ln S'_{mij}}^{+\infty} f(\ln u)\int_{\ln v_{m,m-1}}^{\ln v_{m,m+1}} f(\ln v)d(\ln v)d(\ln u) . \tag{5}$$

In formula (5), $f\left(\bullet\right)$ is the probability density function of normal distribution.

Travellers have two options of transportation modes on the mini car and regular bus. The intersection of the expense quantity curves of min car and regular bus is v_1, if the travellers' time value is higher than v_1, they will choose the mini car; otherwise, they select the regular bus. Secondly, after choosing travellers' own travel modes, the travel will actually occur only when the travel utility is greater than the expense quantity. By calculating the volume of the curve part that effectiveness of various modes of transportation exceeds the expense quantity, the ratio of the actual demand manifested by the ultimate potential demand of the regular bus and mini car can be obtained, namely, the implementation rate of the potential transport demand. The ratio of the actual demand manifested by potential demand of regular bus is the volume of two combination points of probability distribution abc of the time value and utility, while that of the mini car is the volume of bcd.

3 MD models prediction method on the transportation mode under the low carbon mode

3.1 PREDICTING THOUGHT

Based on the basic principle of MD model and formula (4), it predicts the actual demand of the transportation mode m in the t year can be obtained according to its potential travel demand and the manifest rate of potential demand, that is:

$$q_{mij}^{t} = Q_{ij}^{t}\left(R_{mij}^{'}\right)^{t} . \tag{6}$$

The probability of selecting the transportation mode m, namely the travel demand proportion occupied by transportation mode m, can be shown as:

$$P_{mij}^{t} = \frac{q_{mij}^{t}}{\sum_{m} q_{mij}^{t}} = \frac{Q_{ij}^{t}(R'_{mij})^{t}}{\sum_{m} Q_{ij}^{t}(R'_{mij})^{t}} = \frac{(R'_{mij})^{t}}{\sum_{m}(R'_{mij})^{t}} . \tag{7}$$

Therefore, the key of predicting the sharing rate of transportation mode m is to obtain the manifesting rate of potential demand of transportation mode m.

3.2 MODAL CALCULATION THOUGHT

Currently, researches on the MD model calculation method are seldom published at home and abroad, and there is basically no mature achievement that can be directly applied. The potential passenger transport demand in the MD model itself is a relatively abstract concept, which is an invisible, intangible thing in the social and economic phenomena. Therefore, it cannot be obtained through investigation and statistics, while the manifested passenger capacity can get a more accurate value by observation and investigation. In this paper, the calculation of the potential passenger transport demand between the traffic zones is achieved by twice applying the relativity principle on two levels.

The key for the model solution is the solution of four variables of time value, travel expense quantity, travel utility and potential passenger transport demand.

(1) Estimation of the travel time value

The first thing to use the MD model to predict is to make sure the status quo and future value of the time value on the area of the research object. There are many factors that affect the travel time value, but they are mainly connected with the travel purpose and the level of household income. According to the correlation researches of the World Bank on the traffic assessment, the considerations and computational methods of each trip purpose are shown in Table 1:

TABLE 1 The considerations and computational methods of each trip purpose

Travel type and purpose	Consideration	Computational method
Working trip	Traveler's labour charge	1.29×salary/hour
Business trip	Traveler's labour charge	1.29×salary/hour
Commuting and other working trips	Empirical observation	0.2×family income/hour

Calculation on the time value of per person with various vehicle models shows as formula (8):

$$V_{person}^{k}=\frac{1.29*W*QT_{work}^{k}+0.2*H*QT_{unwork}^{k}}{QT_{work}^{k}+QT_{unwork}^{k}}. \quad (8)$$

In the formula (8), V_{person}^{k} - the time value per person per time of the motor vehicle typed K, Yuan/person/hour;

W - the average hourly earnings level, Yuan/hour;

H - the average hourly family income level, Yuan/hour;

QT_{work}^{k} - the turnover of the residents' working trip, person/minute;

QT_{unwork}^{k} - the turnover of the residents' nonworking trip, person/minute;

On the consideration of allocating the model parameters, the parameter of the time value has more important role on choosing the rail rapid transit system for the passengers. The higher the unit time value is, the more the passengers tend to choose the rapid system to complete travel.

For the time value of future years, due to the rising tendency of the time value, and the lower increase of the time value than that of the GDP (Because people has relatively prior tendency on the improvement of the lifestyle other than the travelling expenses as well as the increase of the expend on education during the process of income incensement), based on the regression analysis of historical data and combine the experience, if assuming the variance and base year of the time value in the future years are the same, its mean value can be calculated by the following formula:

$$v_t=v_j\sqrt{\frac{GDP_t}{GDP_j}}. \quad (9)$$

In this formula: v_t is the mean value of the time value for future years; v_j is the mean value of the time value for base years; GDP_t is the future gross domestic product; GDP_j is the gross domestic product for the base years.

(2) Estimation of the travel expense quantity

According to formula (1), the travel time value, travel time and low carbon factors can be obtained by RP / SP [12] investigation and research results, the calculation of the travel expense C_{mij} in the travel expense quantity in this paper uses the formula (10).

$$C_{mij}=R_{mij}\times\frac{L_{mij}}{n}. \quad (10)$$

In this formula: R_{mij} is the freight rate of transportation mode m, Yuan/ (person • km), the mini car's travel expense is equal to the corresponding road transport cost plus fuel cost; L_{mij} is the running mileage of transportation mode m, km; n is the number of calculation, the passenger car gets 1, while the mini car gets the average real carrying number.

(3) Calculation of the travel utility

Using the computational formula of the manifest rate of potential demand in the MD model, that is $R_{mij}=\frac{q_{mij}}{Q_{ij}}$, it can get:

$$Q_{ij}=\frac{q_{1ij}}{R_{1ij}}=\frac{q_{2ij}}{R_{2ij}}=\ldots=\frac{q_{mij}}{R_{mij}}. \quad (11)$$

Putting the formula (5) into formula (11), the $q_{1ij},q_{1ij},\ldots,q_{mij}$ in the equation can be obtained by actual passenger travel statistics, probability density function form of the time value and utility is known, the mean value and variance of the time value can be estimated by the aforementioned method. If assuming that the variance of the utility is same as the time value, then only the variance $u_{\ln u}$ of the utility is unknown; but it can get C_m^2 equations from the formula (11), and the C_m^2 equations summation can get C_m^2 utility average solution. Since the average value of a number of data and sum of squares of deviations in the individual data are the minimum, it takes the averaging approach to confirm the mean value of utility and finally the utility average adopts the average value in order to be better fitted with the current situation, namely:

$$u_{\ln u}=\frac{1}{C_m^2}\sum_{k=1}^{C_m^2}(u_{\ln u})_k. \quad (12)$$

(4) Calculation of the manifest rate of potential demand

Respectively putting the calculated time value, travel utility and travel expense quantity of the base years and forecasting years into the formula (5), it can get the status quo R'_{mij} and predicted value $(R'_{mij})^t$ of OD for the manifest rate of potential demand between each transportation mode.

4 Analyses of examples

This paper takes Dongguan City as an example to predict the urban means of transportation under the low carbon demand. In 2009, Dongguan City did the research on residents traveling (RP survey) and residents traveling intention (SP survey) directing at the adjustment of the Rail Transit Planning. The research content includes the personal socio-economic characteristics survey (gender, age, occupation, education and income) , family socio-economic characteristics (home address, family population, family income, number of family cars and bikes), activity chain survey data (objectives, start time, trip distance, travel modes, travel costs, whether needs transfer) and residents' travel intention choice under the assumptions of the completion of various modes of rail transit network, site settings and different levels of fares. The number of effective samples of this survey is about 6,000, which distributes in the 32 towns of the whole city. These data gets settling and applying in the Rail Transit Planning Adjustment in Dongguan City.

It is known that, the total area of the administrative region in Dongguan City is 2465 km2. According to the sixth census, in 2010 the city's permanent resident population is 8.22 million, of which the household population is 1.8177 million. This paper will divide Dongguan City into four major traffic areas, 55 traffic zones. Based on the survey, the means of transportation in Dongguan City are mainly in the way of private and individual transportation mode. In 2010, the proportion of each transportation mode is: walking 25.9%, private cars 25.2%, bikes 15.0%, buses 15.2%, motorcycles 14.2%, taxis 1.9%, company cars 0.4%, others 2.3%.

Compared with the survey data in 2001, 2005, 2008 and 2010, the travel time and travel distance status quo remain the same, and have no larger changes, as shown in Table 2.

TABLE 2 The average travel time consumption of different transportation of DongGuan City

Transportation mode / Year	Walking	Bicycle	Bus	Taxi	Motorcycle	Company car	Private car	Total
2001	17.1	18.7	31.2	19.4	20.5	23.4	21.0	19.3
2005	17.1	17.2	28.2	24.1	19.1	24.5	27.7	19.7
2008	15.5	18.4	35.0	24.8	18.0	27.8	24.4	20.0
2010	19.4	20.8	37.1	30.1	18.4	29.1	25.4	24.1

By calculating, the parameters of the unit time value in the main characterized year predicting by the passenger traffic in Dongguan City, as shown in Table 3.

TABLE 3 The time value per people of the characterized year

Year	2020	2022	2024	2027	2029	2031	2042	2044	2046	Prospect
VOT(Yuan/Person Hour)	19.7	21.4	22.3	23.1	23.7	24.1	25.6	25.9	26.1	27

According to the data of residents travel characteristics (travel strength, consumption and trip purpose) and major road traffic volume, it gets the basic data of travel time and travel costs of all modes of transportation in the base year. The MD model which is built by applying the low carbon traffic mode divides the residents travel modes within the scope of Dongguan city. Setting 2020 as the transportation forecast characterized year, the results are shown in Table 4.

From the predicted results, with the rapid traffic construction, development as well as the improvement and implementation of the traffic policies in Dongguan city, its structure of travel mode has changed, especially the further implementation of the ban on motorcycles greatly develops the call for public transport and increase of the demand on low-carbon, which predicts that the share of Dongguan motorcycle travel mode has decreased from 14.2% in 2010 to 3.5% in 2020 in the total travel modes; the share of public transport mode has increased from 15.2% in 2010 to 25.1% in 2020; the change of the share of the company car and other cars is less significant.

TABLE 4 The urban traffic modal splitting results in the domain of DongGuan City

Year / Transport mode	2010	2015	2020
Walking	25.9%	26.7%	28.6%
Bicycle	14.9%	15.3%	15.8%
Motorcycle	14.2%	10.5%	3.5%
Conventional public traffic	15.2%	16.1%	14.8%
Rail transit	0.0%	3.8%	10.3%
Taxi	1.9%	2.1%	2.3%
Mini car	26.2%	24.5%	23.8%
Company car	0.4%	0.3%	0.3%
Others	1.3%	0.7%	0.6%

5 Conclusions

Based on the basic principles of the MD prediction model, this paper considers the low-carbon policies impact on residents travel choice and depicts the travel costs by using the factors of travel time, cost, safety, and low-carbon to propose the MD forecasting model split by urban transport mode under the low-carbon mode. It puts forward the model calculation process and has applied the improved model. The application results show that the prediction result has some rationality. However, the mean value of the travel utility defaulted by the MD model does not change along with time, which still has a certain gap with the actual situation, and needs to be further improved.

References

[1] Train K E 2003 *Discrete Choice Methods with Simulation* Cambridge: Cambridge University Press 45-62

[2] CHEN Zheng 2004 *Research on The Combined Modal Split/Trip Distribution Model and Software Design* Southeast University 12-3

[3] NIU Xue-qin, WANG Wei, YIN Zhi-wei 2004 Research on Method of Urban Passenger Traffic Mode Split Forecast *Journal of Highway and Transportation Research and Development* **19**(1)

[4] WANG Yu-ping 2011 *Study on Urban Rail Transit Passenger Forecast and Analysis* Chang`an University

[5] Chieh-Hua Wen, Koppelman F S 2001 The generalized nested logit model *Transportation Research B* **35**(7) 627-41

[6] Hensher D, Greene W 2002 Specification and Estimation of the Nested Logit Model: Alternative Normalizations *Transportation Research B* **36**(1) 3-15

[7] GUAN Hong-zhi 2004 *Disaggregate model-A Tool of Traffic Behavior Analysis* Beijing: China communication press 27-42

[8] HE Ming, GUO Xiu-cheng, RAN Jiang-yu, ect. 2010 Forecasting Rail Transit Split with Disaggregated MNL Model *Journal of Transportation Systems Engineering and Information Technology* **10**(2) 136-42

[9] GONG Bo-wen 2007 *Study on the Disaggregate model of Traffic Split and Application* JiLin University

[10] SONG Xue-mei, JIANG Yang-sheng, Yun Liang 2010 Study on the calculation method of MD forecast model *Journal of Transportation Engineering and Information* **2**(8) 65-70

[11] AN Wen-juan, CHEN Feng 2012 Regional Traffic Modal Splitting Method Based on Improved MD Model *Journal of Chongqing Jiaotong University (Natural Science)* **8**(4) 824-7

[12] Elisabetta Cherchi, Juan De Dios Ortuzar 2002 Mixed RP/ SP models incorporating interaction effects *Transportation* (29) 41-7

Authors	
	Shimei Wu, born in **Current position, grades:** School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin, P.R. China. Ph.D candidate of transportation science and engineering **Scientific interest:** urban planning, transportation planning, traffic management, traffic demand forecasting, etc **Publications:** 13
	Yulong Pei **Current position, grades**: School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin, P.R. China . Ph. D in Transportation Planning and Management **Scientific interest:** transportation planning, transportation safety, traffic control etc **Publications:** 300 scientific papers and 18 books
	Guozhu Cheng **Current position, grades:** School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin, P.R. China. associate professor **Scientific interest:** transportation safety **Publications:** 70 scientific papers and 6 books

Research on dynamic evolution of innovative virtual prototyping technology diffusion based on cellular automata

Guangbin Wang[1], Honglei Liu[1], Lei Zhang[2*]

[1] *School of Economics and Management, Tongji University, Shanghai 200092, China*

[2] *School of Business, Shandong Jianzhu University, Jinan 250101, China*

Received 1 March 2014, www.tsi.lv

Abstract

The construction industry plays a very important role in the national economy; it is widely criticized because of its slow technical progress and long-term inefficiency all over the world. Building information modelling (BIM) is a transformative virtual prototyping technology for construction industry. VP (Virtual Prototyping Technology) based on BIM as the core technology has been widely regarded as a tool to solve this problem, but was questioned by both academia and industry due to its delayed diffusion. To solve this problem, this paper is based on the characteristics and the evolution rules of cellular automata, built on the CA model of the BIM proliferation process in construction projects, simulating this process, then analysing the impact of important factors such as diffusion willing, decision-making preferences, national and industry support and other factors to the BIM technology diffusion, studying the changes in the proportion of BIM recipients and the importance of the distribution position of the initial to the BIM proliferation process. Finally, it analyses the randomness of BIM technology diffusion.

Keywords: Building Information Model Diffusion, CA Evolution Model, Diffusion Willingness, Decision-Making Preferences, National and Industry Support

1 Introduction

The construction industry is the backbone of social and economic life, and supports national development and prosperity, but the long-term low productivity that results in energy consumption and serious environmental pollution makes it gain widespread attention. VP (Virtual Prototyping Technology) based on BIM as the core technology has been widely regarded as a tool to solve this problem in the whole world, and some scholars think it will bring an unprecedented revolution in construction industry because of its enormous potential impact for enhancing the construction project performance and industry production efficiency [1]. It is widely regarded as the most important technical concepts produced in the AEC industry (Architecture, Engineering and Construction Industry) in the last decade [2-4], and it can be effectively used in the whole construction project lifecycle such as site planning, collaborative design, clash detection, energy consumption analysis, construction schedule simulation, cost control and other aspects through digital and intelligent building facilities. Although the VP-related ideas and prototypes have already been considered to have broad application prospects as early as the 1970s, the breadth and depth of its application are not satisfactory now [5, 6]. Construction enterprises are the players in the construction market, and construction projects are the service targets and entities of BIM applications, so only the effective diffusion and application of BIM can affect project performance, enterprise performance and even industry efficiency. Nowadays, how to promote the diffusion of VP (Virtual Prototyping Technology) such as BIM in the construction industry has become the new proposition confronted with both academia and businessmen.

To solve this problem, many scholars have carried out relevant studies. In view of the progress and upgrading of hardware and software technology, the focus of related research has been shifted from technical issues to problems of implementation. Dulaimi investigated the motivation of technological innovation diffusion and innovation-related interaction between organizations of the construction projects, and identified the impact of the target incentive, participants' commitment and other factors on motivation of innovative applications [7]. Davis et al. pointed out that the main reason for the boycott to BIM lies in the failure to support organizational change of BIM application [8]. Jernigan also pointed out that the significant change that BIM brought in the construction industry caused organizational inertia and resistance to change [9]. The major barriers to BIM application include poor interoperability among different software, unclear business value and return on investment of BIM, and reluctance among construction companies to share information [10-12]. The above studies provide an important basis for understanding the impact of BIM

[*] ***Corresponding author*** email: zhlei06@gmail.com

proliferation factors of construction projects. However, these studies, mostly based on case studies or surveys on various factors affecting the static identifier, also emphasize more on the technical aspects of the study, and take fewer dynamic evolution of the BIM technology diffusion process into account. This paper is based on the characteristics and the evolution rules of cellular automata, builds the CA model of BIM proliferation process in construction projects, simulates this process, then analyses the impact of important factors such as diffusion will, making preferences, national and industry support and other factors to the BIM technology diffusion, focused on the randomness of BIM proliferation, in order to provide a theoretical basis for the diffusion and application of VP whose core technology is BIM in the construction industry.

2 Dynamic evolution of the BIM proliferation model

Cellular automata, which is in accordance with certain local rules in a discrete, finite cellular space composited of cellular, is dynamical systems in the evolution of discrete time dimension. CA models, constructing individual (cell) on the micro-level and obtaining the macro results by summing the micro individuals from the perspective of complex systems, is a bottom-up (bottom-up) research method. Essentially, the CA model is a dynamic evolution system based on the interactions of a large number of cellular.

2.1 RESEARCH ASSUMPTIONS AND THE DYNAMIC EVOLUTION OF THE CA MODEL BUILDING

(1) Hypothesis 1: The BIM proliferation process of Construction Projects is a dynamic system according to certain rules of non-linear interaction of actors. In the process of BIM technology diffusion, the owner, designer and construction side, government departments and agencies and other project participants showed a nonlinear and varied relationship on the overall level; from the perspective of the individual level, the individual participants are the core elements in the process of BIM technology diffusion. This process manifested in the interaction relationships of project participants, and in a certain context, can achieve the dissemination and diffusion of BIM technology.

(2) Hypothesis 2: In the BIM technology diffusion process, the project involved individual implemented dissemination and diffusion by communicating with its nearest individuals. The occurrence of local interactions between individuals makes the project happen. Individuals involved in the project send or receive innovative BIM technology are limited only by the project participants or the closest; at the same time, there is no priority when many diffusion processes of BIM technology are at the same observation point.

(3) Hypothesis 3: The project involved individuals adjusting the BIM technology diffusion process based on their relationship. They assess their own scenarios, and then adjust the BIM technology diffusion rate according to the above evaluation and expected results to be adjusted.

(4) Hypothesis 4: Participating individuals have completed information about the individual states of the diffusion process. Meanwhile, once innovative technology is received, it will never be changed.

(5) Hypothesis 5: Every involved individual will not be affected by their property differences in each project. Individuals act independently without disturbing each other.

2.2 MODEL

Based on the above assumptions, the diffusion process for the BIM evolution model is established as follows:

(1) Set the rows or columns of the square-type grid of the cellular automaton model. The number of cells of cellular automata $n \times n$ in the grid represents the number of individuals $N = \sum i$ involved in the process of BIM proliferation. This model uses a 50 × 50 grid to construct cellular space, the grid intersection points of rows and columns represent recipients and proliferators of BIM technology.

(2) There are two like bodies $\{0,1\}$ representing whether the individual accepts the diffusion of BIM technology: 1 represents that BIM innovation has been accepted by the individual, and 0 represents the opposite.

(3) Neighbours form. In the CA model, there are many neighbours' forms. This model uses mole (Moore) type; that is, each user has eight neighbours, and the eight neighbours at the t-1 moment will affect the central users who are at the t moment.

(4) The evolution function and state transition rules of cellular.

This article assumes that all individuals involved in the evolution of BIM technology diffusion are affected by diffusion willingness (γ) and decision-making preferences (δ), then select support on BIM from country and industry bodies (ε) and the choice of other ones (i) around the involved individual as external factors. Its influence function is:

$$f(t) = \gamma(t) \times \delta(t) + \varepsilon(t) + p\sum_{i=1}^{8} s_i(t-1) \tag{1}$$

In the formula (1): (γ) Diffusion willingness: the possibility of BIM technology diffusion $\gamma(0 \le \gamma \le 1)$, $\gamma = 0$ means unwilling, $0 < \gamma < 1$ means the degree of willingness, $\gamma = 1$ expressed complete readiness, the innovators of BIM who want to obtain and maintain a reputation for higher qualifications expect to solve the practical obstacles and has a higher degree of willingness, but considering the competition, the experience will be reserved to a certain extent. Therefore, the range of value is $(0.5 \le \gamma \le 0.8)$; (2) Decision Preference (δ): This article use emphasize degree on income from BIM innovation accepted corporate representing decision-making preferences $\delta(0 \le \delta \le 1)$, $\delta = 0$ means no attention, $0 < \delta < 1$ means the degree of attention, $\delta = 1$ means absolute attention. There are some differences in decision-making preferences of trying BIM technology in different individuals. (3) National and industry organizations supporting (ε). The value of support from country and industry bodies to BIM is $0 \le \varepsilon \le 1$, which indicates the support is 0, the diffusion of BIM mainly relies on self-development of the industry, the value of ε is more which means the support is greater.

This article will use the simulation tool - MATLAB to program the evolution model of the BIM technology diffusion based on cellular automata, and analyse the simulation results based on the above.

3 Simulation of evolution model of BIM technology diffusion

The diffusion of the BIM Construction Project is the process of partial interactive process between individuals, and its complexity is suitable for the "bottom-up" complex systems - cellular automaton model. Through the establishment of the diffusion process evolution model of BIM technology based on cellular automata, this paper simulated the diffusion process, analysed the individual wishes, internal factors of the participants in decision-making preferences and the influence of relationships between individual and neighbours, and also analysed the variation of acceptable BIM technology innovation under certain parameters.

3.1 PARAMETER INITIALIZATION

In this paper, the initial value of willingness, decision-making preferences, the relationships with neighbours of BIM involved individual, support from country and industry bodies on BIM and the position of the initial proliferators is defined, in which diffusion willingness and decision-making preferences are shown in Table 1.

TABLE 1 The value of Diffusion willingness and decision-making preferences

Decision-making preferences δ **Diffusion willingness γ**	$\delta = 0.3$	$\delta = 0.6$
$\gamma = 0.5$	$\gamma \times \delta = 0.15$	$\gamma \times \delta = 0.30$
$\gamma = 0.6$	$\gamma \times \delta = 0.15$	$\gamma \times \delta = 0.36$

In BIM proliferation, the simulation is divided into two cases, no support from state and industry bodies and the presence of support. In this paper, cellular space is represented by a square grid $n = 50$ that takes Moore type. Assuming that the 1/10 (diffusion random position) of the involved individuals accept the innovation in the initial state, while the 9/10 did not receive the innovation. The frequency of simulation is 50 times.

3.2 ANALYSIS OF SIMULATION RESULTS

(1) The change of proportion of BIM recipients in different iterations.

Setting of parameters: the value of diffusion will and decision-making preferences are 0.5 and 0.3. Without the support from national and industry bodies, BIM proliferation depends on the innovative diffusion of core corporate, which will be affected by the external environment. The black areas in the figure consist of the cells obtaining the innovation of BIM, while the cells that are not transmitted are components the white areas.

As we can see from Figure 1, BIM technology transmits to each direction at almost equal speed, and also the transmission is random. BIM technology will soon transmit around the diffused sources (several companies who initially mastered the technology) under the bringing in of a number of aggressive decision-makers and the initiating of the conservative ones. Finally, apart from a few enterprises, almost all companies will introduce the technology.

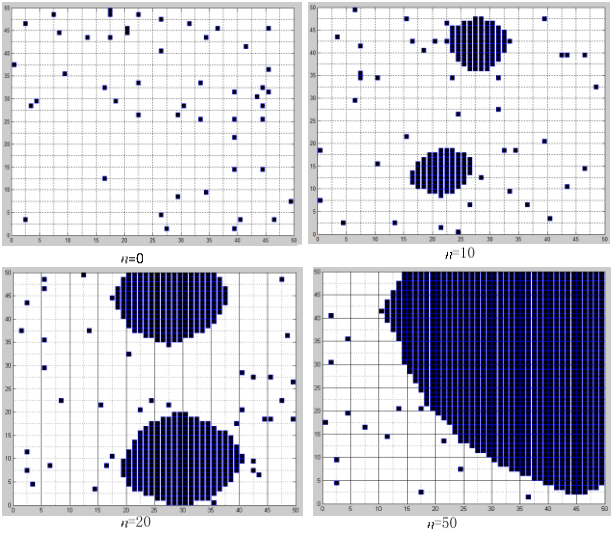
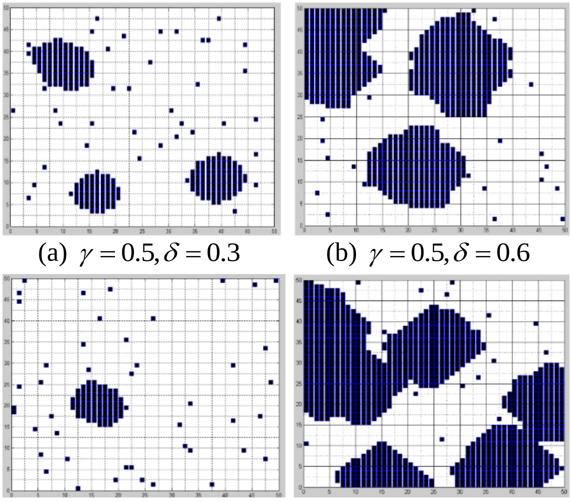


FIGURE 1 Influence of Different Iterations of Simulation on the Diffusion Process

(2) Impact of different values of γ and δ on BIM proliferation

Iterations are analysed from the number $n = 10$ that the involved individual has been diffused by BIM technology. In Figures 2 (a) and (c), when the value of the diffusion willingness of participants and decision-making preferences of recipients are both small, BIM knowledge spreads in fewer individuals in the diffusion process; in Fig 2 (b), BIM knowledge spreads in a large number of individuals in the diffusion process; in Fig 2 (d), when the value of the diffusion willingness of participants and decision-making preferences of recipients are both large, BIM knowledge spreads to almost half of the involved individuals in the diffusion process.



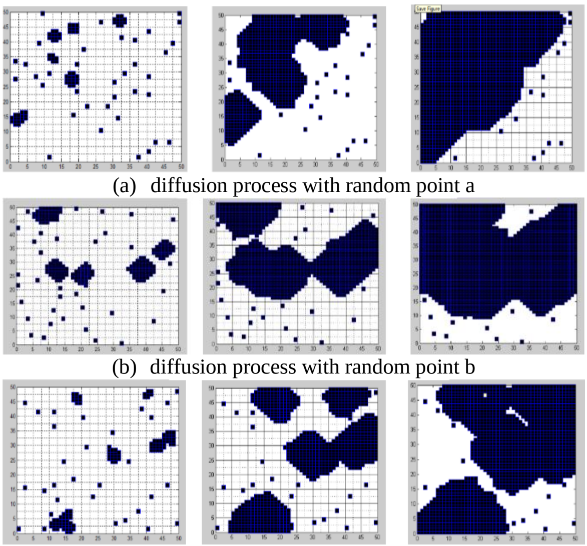
(a) $\gamma = 0.5, \delta = 0.3$ (b) $\gamma = 0.5, \delta = 0.6$

(c) $\gamma = 0.6, \delta = 0.3$ (d) $\gamma = 0.6, \delta = 0.6$

FIGURE 2 Simulation Diagram about the Influence of Diffusion wishes γ and Decision-making Preference δ on BIM Proliferation Process

Meanwhile, it can be seen from Figure 2, as the value of the diffusion willingness and decision-making preferences of the involved individual increase, the average speed of BIM proliferation is accelerates. Decision-making preferences of the involved individual have great influence on the average speed of BIM proliferation, while diffusion willingness has great impact on the diffusion region of BIM proliferation.

(3) Impact of the location of initial knowledge proliferators on the BIM proliferation process



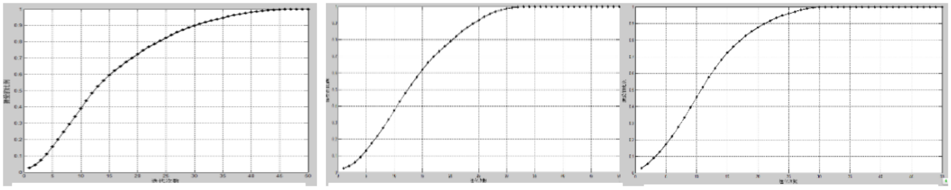
(a) diffusion process with random point a

(b) diffusion process with random point b

(c) diffusion process with random point c

FIGURE 3 Simulation Diagram about the impact of the Location of Initial Knowledge Proliferators on the BIM Proliferation Process

Keep the value of the diffusion willingness of participants and decision-making preferences of recipients unchanged, take the influence of BIM proliferators Moore type, and compare three different and random positions with the same number of iterations, as shown in Figure 3. From the perspective of the number of the involved individual who has been proliferated by BIM innovators, the value of the diffusion willingness of participants and decision-making preferences of recipients are. Selecting no support from national and industry organizations, after 50 simulation clockings, the state of the iterations at different and random locations are shown in (a) in Figure 4. Compared to (b) and (c), the proliferated cells appeared in a small black area, and only occupies the top left of the grid spatial location, and the diffusion rate is relatively slow, indicating that the BIM is spread to a small number of individuals involved when the initial positions of the proliferators are far away from the organized centre in the BIM proliferation process. Compared to (a), the proliferated cells are relatively in the centre of Figures (c) and (b), and diffusion speed is fast, which shows that BIM spreads to a large number of individuals involved when the initial positions of the proliferators are close to the organized centre in BIM proliferation process.



(a) random location a, n=45 (b) random location b, n=32 (c) random location c, n=30

FIGURE 4 Simulation Diagram about the impact of Proliferators' Initial Position on BIM Proliferation

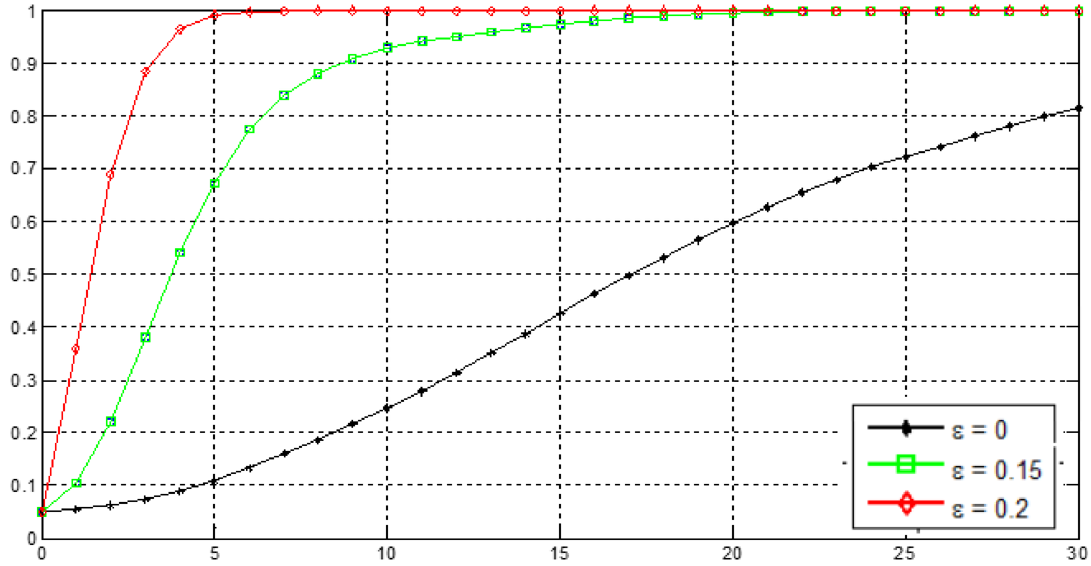
From Figure 4 we can see that the iterations of BIM proliferation spreading to the whole are at locations b and c, while to achieve the overall diffusion at location a, the three curves all need to be in line with the classic innovation diffusion model of Bass "S" curve. Thus, the initial location of BIM proliferation in the organization has a greater impact on the BIM proliferation rate. If the initial BIM proliferators are the core participating member, they can ensure a faster diffusion rate of BIM; conversely, if the initial BIM proliferators are at the edge, it is not conducive to BIM proliferation. This reveals that the core participants plays an important role in the construction project.

(4) Randomness of BIM technology diffusion

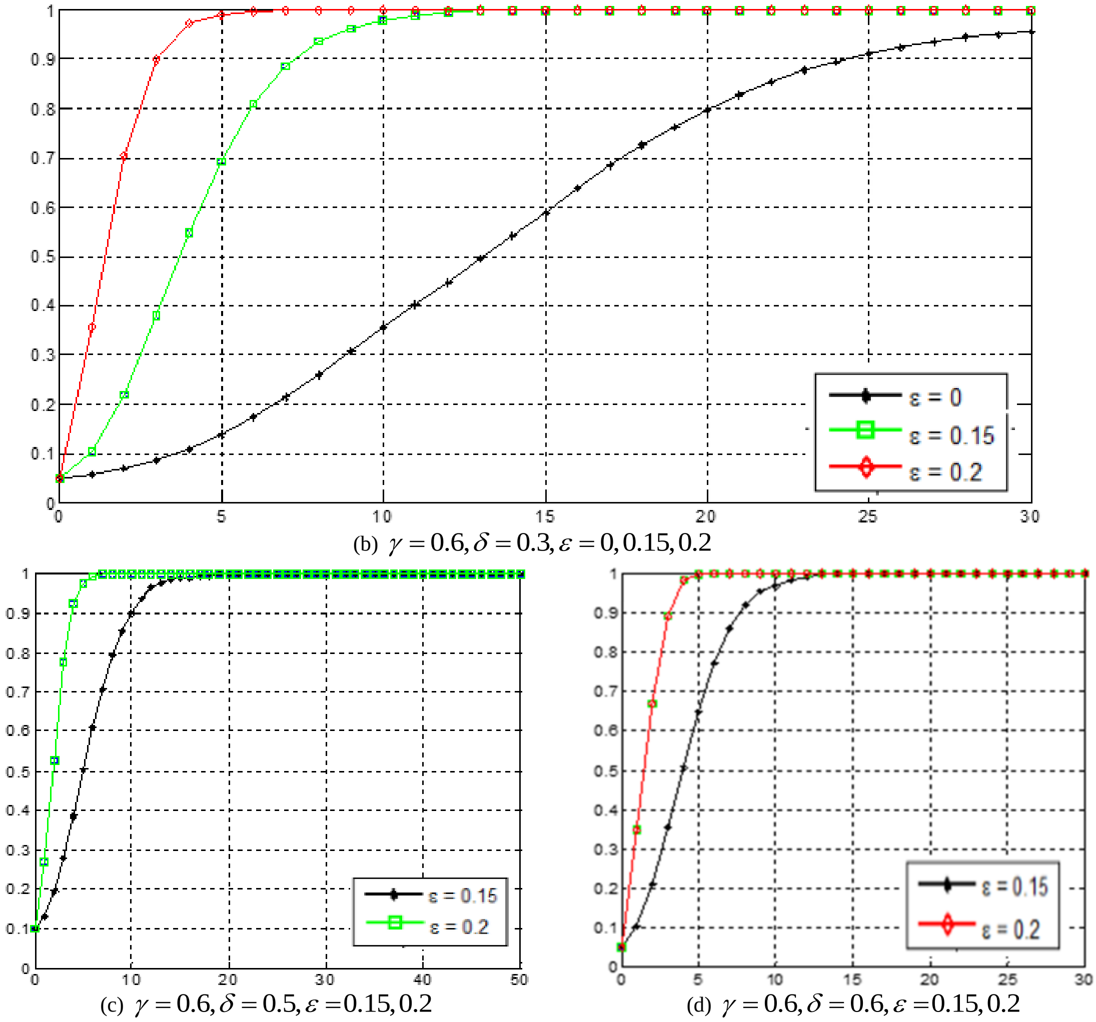
From the above analysis about the simulation results, we can conclude that establishing good relations between involved individuals and encouraging the spreading wishes of BIM proliferators can effectively improve the speed and efficiency in BIM proliferation. As mentioned above, the arbitrariness is largely due to the different positions of the distribution of the initial BIM proliferators. The grid cells at different locations means that the initial BIM proliferators face different environments, which will influence the direction and speed of diffusion greatly. In the practice of BIM proliferation, this arbitrariness is mainly broken down by the software vendor's marketing promotional efforts and a market leader in the design side. Project participants, including the owners, have passively accepted the form that regarding projects as a link to BIM implementation at present. Therefore, BIM technology diffusion will show certain regional stability. On the other hand, the BIM theory and practical experience in foreign countries show that the influence of external factors on BIM technology (such as the support from government and industry organizations) is very important. With the popularity and recognition of BIM technology, BIM software developers have completed the mission of marketing and have begun to focus on technical services and consulting. Leading designers have changed direction to the diffusion of inter-enterprise technology, adjustment of organizational structure, and the change of workflow designing, the internal training of technical staff, and pilot projects have become the main task for initial BIM proliferators.

(5) Impact of external factors ε on the BIM diffusion process

Keeping the value of the diffusion willingness of initial participants γ and decision-making preferences of recipients δ unchanged, defining the national and industry agencies’ supporting ε as influence of BIM projects proliferators, executable program for Moore-type, and comparing the change of influence. In order to conveniently observe the ratio of the initial proliferators, we take is 5%. As we can see from Figure 5, the effect of the national and industry agencies’ supporting ε on BIM proliferation is obvious, and greater supporting strength will be more conducive to the innovation diffusion. Taking (A) as an example, when the value of ε is 0, BIM techniques spread from the initial proliferators to the entire grid, and the number of iterations is up to 37 times (Figure 30 shows only 30 iterations), the whole diffusion process is relatively long, while the value of ε are 0.15 and 2, the number of iterations which will significantly decrease are respectively 12 and 7. Therefore, national and industry agencies’ support has obvious promoting effect on BIM proliferation. Through comparing (a) and (b) in Figure 5, ε has high rate of sensitivity to BIM diffusion. Each small increase proportion will influence BIM diffusion greatly. The value of ε is 0.2, the expansion curves are almost the same, which can compensate the lack of willingness and decision-making preferences to a large extent.



(a) $\gamma = 0.5, \delta = 0.3, \varepsilon = 0, 0.15, 0.2$



(b) $\gamma = 0.6, \delta = 0.3, \varepsilon = 0, 0.15, 0.2$

(c) $\gamma = 0.6, \delta = 0.5, \varepsilon = 0.15, 0.2$

(d) $\gamma = 0.6, \delta = 0.6, \varepsilon = 0.15, 0.2$

FIGURE 5 Simulation Diagram of Impact of National and Industry Organizations Supporting on Different Parameters on BIM Diffusion Process (X: number of iterations; Y: acceptance ratio)

4 Conclusions

This article is based on characteristics and evolution rules of the cellular automata, selected the CA as the BIM proliferation process model, and established the evolution function and state transition rules of the cellular simulate the diffusion process, and analysed the relationships between individual wishes of project proliferators and internal factors of participation in decision-making preferences, participated individuals and neighbours. Studies have shown that the design unit as a leader in the application of BIM selected random diffusion in a certain internal factors and external factors, showing obvious regional (block) distribution with the diffusion process. Moreover, the decision-making preferences of Individual recipients of project participants showed greater impact on the average speed of BIM proliferation, but wishes of BIM proliferators influenced greatly on BIM proliferation regions. In the different process of the simulation graph, we can find that the position of the initial BIM proliferation in construction project organization has a greater impact on BIM proliferation velocity. Due to the large randomness between the diffusion and the recipients, the distribution area and diffusion wishes of the diffusions are different, which lead to a greater difference in the proliferation of BIM. Taking the convergence of research into account, this paper was unsuccessful in quantifying the various parameters in the model in order to further explore the influence degree of various factors on the dynamic evolution of BIM proliferation. This paper is just a beginning on the applications of BIM proliferation in the construction industry, which need a more in-depth follow-up study in the future.

Acknowledgments

This paper is supported by the Ministry of Science and Technology of China and South Korea Joint Research Projects (No.OS2012GR0050, Research on ICT-based Construction Management using Open BIM), as well as Natural Science Foundation of China (NSFC) (NO.71272046).

References

[1] Young N W, Jones S A, Bernstein H M 2008 *Building Information Modeling: Transforming Design and Construction to Achieve Greater Industry Productivity* New York: McGraw - Hill Construction

[2] Eastman C, Teicholz P, Sacks R, Liston K 2011 *BIM handbook: A guide to building information modelling for owners, managers, designers, engineers, and contractors* (2nd Edition) New Jersey: John Wiley & Sons, Inc.

[3] Guo H L, Li H, Skitmore M 2010 Life-Cycle Management of Construction Projects Based on Virtual Prototyping Technology *Journal of Management in Engineering* **26**(1) 41-7

[4] Jernigan F 2007 *Big BIM little BIM - The practical approach to Building Information Modelling-Integrated practice done the right way* USA: Booksurge Llc

[5] Taylor J E, Bernstein P 2009 Paradigm Trajectories of Building Information Modeling Practice in Project Networks *Journal of Management in Engineering* **25**(2) 69 - 76

[6] Hartmann T, Gao J, Fischer M 2008 Areas of Application for 3D and 4D Models on Construction Projects *Journal of Construction Engineering and Management* **134**(10) 776 - 85

[7] Dulaimi M F, Ling F Y Y, Bajracharya A 2003 Organizational Motivation and Inter - organizational Interaction in Construction Innovation in Singapore *Construction Management and Economics* **21**(3) 307 - 18

[8] Davis K A, Songer A D 2008 Resistance to IT change in the AEC industry: An individual assessment tool *Electronic Journal of Information Technology in Construction (ITcon)* **13**(14) 56-68

[9] Jernigan F 2007 *BIG BIM little bim - The practical approach to Building Information Modelling - Integrated practice done the right way! (1st ed.)* Salisbury: 4Site Press

[10] http://images.autodesk.com/adsk/files/final_2009_bim_smartmarket_report.pdf (last access: 3/11/2013)

[11] He G P 2012 *The Theory of BIM and Nonsense Arguments: The Theory is not Nonsense, and Vice Visa* ChinaBIM, Source: http://www.chinabim.com/school/heguanpei/2012-12-13/4715.html (last access:3/11/2013)

[12] Pan J, Zhao Y 2012 Research on Barriers of BIM Application in China's Building Industry *Journal of Engineering Management* **26**(1) 6-11

Authors

Guangbin Wang, born in April, 1967, Heze, Shandong province

Professor, doctoral advisor and PhD. Professor Wang got PhD in Tongji University in 1998; engaged in research work in Stanford University in 2007-2008; several times went to the United States, Germany and other countries for short-term academic exchanges. Major research fields include: project management, project finance, construction management information. Professor Wang also undertakes international cooperation research subject of the Ministry of science and scientific research subject in Shanghai, as well as a number of enterprise research subject. Main courses: project management, project finance. Published materials: construction project management, Investment construction project organization, over 80 papers, and so on. Participate in project research and practice: Shanghai Yindu Building project management and Guangzhou Airport overall management, Beijing Olympic Games project management information platform and Sino-German Friendship Hospital.

Honglei Liu, born in December, 1982, Puyang, Henan province

PhD student. Research Field: Theory and Practice of BIM-Based Project Management, Information Integration;
Academic Experience:2013/1~Present Research on the Internal Mechanism and Structure Optimization for Contractual Governance of Construction Project with BIM Application Project supported by the National Natural Science Foundation of China (71272046); 2012/1~Present Research on ICT-based Construction Management using Open BIM, International cooperation in science and technology special project (OS2012GR0050)
Working Experience:2012/7~2012/10 Consultant of strategic planning of BIM implementation for Shanghai Jianke Engineering Consulting Co., Ltd; 2009/10~2011/8 Project Management of Shanghai Electronic Industries Procurement Centre

Lei Zhang, born in November, 1973, RiZhao, Shandong province

Associate Professor and PhD. Zhang worked in Shandong Jianzhu University since 1998, and got PhD in Tongji University in 2014. Major research fields include: project management, construction management information, theory and practice of BIM. Main courses: project management, Real estate market research. Academic Experience: 2013/1~Present, Research on the Internal Mechanism and Structure Optimization for Contractual Governance of Construction Project with BIM Application Project supported by the National Natural Science Foundation of China (71272046); 2010/9~2012/7, key technology research on integrated management of Shanghai West Railway Station Hub project based on BIM supported by Science and Technology Program Foundation of Shanghai (10DZ1202603). 2008/1~2009/4, Status and Trends of Shandong Province residential real estate analysis supported by Soft Science Project in Shandong Province. Published materials: Market research and forecasting, over 20 papers, and so on.

Research on well-formed business process modelling mechanism

Chen Kai[1], Xie Yi[2]*

[1] *College of technology, Lishui University, Lishui, China*

[2] *College of computer science & information engineering, Zhejiang Gongshang University, Hangzhou, China*

Received 1 March 2014, www.tsi.lv

Abstract

It is very important to ensure that the logic structure of business process model is correct before the model is implemented. Because traditional graphical process modelling methods lack efficiency mechanisms or rules to ensure correctness of the logical structure during business process modelling, they need additional methods to verify its correctness of the logic structure after the business process model is established. Therefore, the well-formed business process modelling mechanism is researched. The business process logic structure model is built firstly. Then the semantic and syntactic rules are presents for the correctness of business process logic structure model, and the algorithm is proposed to detect whether the model meets the rules. The modelling mechanism has been applied in our business process scheduling optimization system with integration of modelling and simulation, which shows its feasibility and effectiveness.

Keywords: business process modelling, modelling mechanism, well-formed model, model verification

1 Introduction

Process structure is the most important and primary aspect of a process model [1]. After the business process or workflow is deployed, detecting and repairing errors of process structure at run time is very expensive and time consuming. Therefore, a critical challenge in business process modelling lies in the design-time verification of business process models [2].

In order to adaptive to the varying needs of the organizations, business processes or workflows are often modelled as graphs [2, 3], in which individual activities within the process and their temporal constraints are shown as a series of nodes and edge (e.g. rectangles and arrows). Process modelling methods based on graphs (e.g., BPMN, activity diagrams of UML) are easy to use and master, the built process models using these methods are simple and intuitively understandable at a glance. However, work of general process modelling based on graphs includes arbitrariness and lacks strictness [4]. Moreover, process modelling methods based on graphs themselves do not provide some effective methods, rules or mechanism against errors in logical structure of process model. Thus, before utilizing the process model, it must be verified by following methods or techniques: Petri-net reduction techniques [5-8], graph reduction techniques [9-10], integer programming [2], simulation techniques [11], process logic [12], π calculus [13], semantic deduction [14], matrix calculus [15].

Petri-net reduction techniques, integer programming, simulation techniques, process logic, π calculus, semantic deduction, matrix calculus are indirect verification methods, and model transformations are required for verifying process model. After model transformations, process models lose their natural structures and are not easy to be understood at a glance [16]. Moreover, the performance of these methods drops off precipitously when the process model becomes more and more complicated with the node increasing, thus it is difficult to complete the verification of the large complicated business process models in distributed real-time interactive environment. Graph reduction techniques can directly detect the structural conflicts of graphical process model through a reduction process based on reduction rules. However, they can detect a limited set of process anomalies because the set of the developed reduction rules is not complete [12, 17]. They give no details about the causes of these conflicts, and, therefore, provide no help for further improvement [16].

In fact, there are two approaches to ensuring the correctness of process model. Besides to check it completely based on these approaches above, another is to build it correctly, which relies on strict grammatical rules that govern the composition of the various elements in the process model [18]. In reference [19] the concept of well-formed business process are proposed based on the idea of structured programming. In reference [19-22] the performance of well-formed business process is analysed and optimized. However, modelling mechanism of well-formed business process is not studied well yet, and there are lacks of formal, systematic description and supporting algorithm.

****Corresponding author** e-mail: xieyi@mail.zjgsu.edu.cn

To fill the gap, this study aims to formalize modelling mechanism of well-formed business process and design the effective algorithms of rule verifications, which can satisfy the requirements to verify the correctness of the large complicated business process in the real-time distributed interactive modelling environment. The proposed approaches have been applied to our developed Business Process Scheduling & Optimization System.

2 Business process logical structure model

2.1 FORMALIZING THE BUSINESS PROCESS LOGICAL STRUCTURE MODEL

The business process logical structure model (BPLSM) represents the control dependency or temporal constraint relationship between activities in business process, and can be formally defined below.

Definition 1 (Business process logical structure model): A BPLSM is defined by a 3-tuple $BPLSM=(A,C,L)$, which is characterized as follows:

(1) A is the set of activities.

(2) C is the set of connectors. $N=A\bigcup C$ is the set of nodes in the business process.

(3) $L\subseteq N\times N$ is the set of links. $l\in L$, $l=\langle n_1,n_2\rangle$ represents the link from n_1 to n_2.

Apparently, a BPLSM is a directed graph $G_L=(N,L)$ which is composed of a set of nodes $N=A\bigcup C=\{n_i\}$ and a set of links $L=\{l_k\}\subseteq N\times N$.

Definition 2 (τ , η). $\tau:C\rightarrow\{And,Or\}$ is a function, which maps each connector onto this connector's logical type. $\eta:L\rightarrow(0,1]$ is a function, which maps each link onto the execution probability of this link.

Definition 4 (Directed path, elementary path). A directed path p from a node n_1 to a node n_k is a sequence $\langle n_1,n_2,\cdots,n_k\rangle$, so that $\langle n_i,n_{i+1}\rangle\in L$ for $1\le i\le k-1$. p is elementary path iff for any two nodes n_i and n_j on p, $i\ne j\Rightarrow n_i\ne n_j$.

Definition 5 ($^{\bullet}$,$|\ |$). $|S|$ is the number of elements in the set S. For $n\in N$, $^{\bullet}n=\{m|(m,n)\in L\}$ is the set of input nodes, and $|^{\bullet}n|$ is the number of input nodes. $n^{\bullet}=\{m|(n,m)\in L\}$ is the set of output nodes, and $|n^{\bullet}|$ is the number of output nodes.

2.2 CONTROL DEPENDENCY RELATIONSHIPS BETWEEN ACTIVITIES IN BPLSM

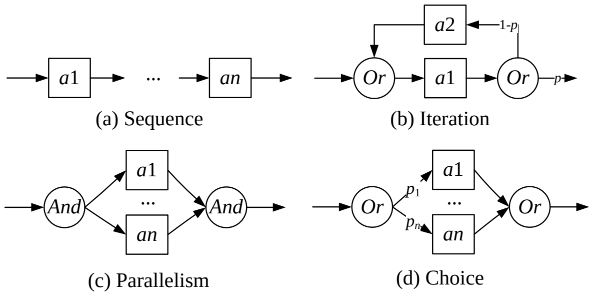
A common relationship between activities in business process model is logistic order that denotes a kind of partial order, called control dependency. The control dependencies such as sequence, And-Split, And-Join, Or-Split, Or-Join, and iteration, compose the model structures of business process. Four kinds of basic model structures shown as Figure 1 are sequence (SEQ), iteration (LOOP), parallelism (AND), and choice (OR). p is the probability of exit loop in Figure 1(b). p_i is the probability of selection branch in Figure 1(d), where $\sum_{i=1}^{n}p_i=1$. The LOOP, AND, and OR basic model structures are called the non-sequential basic model structure. The semantics of these basic model structures excerpted from WfMC (1999) are listed below.

(1) Sequence (SEQ): Activities are executed in order under a single thread of execution, which means that the succeeding activity cannot start until the preceding activity is completed.

(2) Iteration (LOOP): A business process cycle involves the repetitive execution of one (or more) business process activities until a certain condition is satisfied.

(3) Parallelism (AND): A single thread of control splits into two or more threads that are executed in parallel within the business process, allowing multiple activities to be executed simultaneously. Once these parallel executing threads are all completed, they converge into a single common thread of control.

(4) Choice (OR): A single thread of control makes a decision upon which branch to take when encountered with multiple alternative business process branches. No synchronization is required because of no parallel activity execution.



(a) Sequence

(b) Iteration

(c) Parallelism

(d) Choice

FIGURE 1 Four kinds of basic model structures

3 Modelling rules and its verification algorithm for well-defined BPLSM

The correctness of BPLSM includes two aspects: syntax and semantics.

3.1 SYNTAX RULES

A correct BPLSM must satisfy the following syntax rules:

(1) There is only one node $n_s\in N$, $|^{\bullet}n_s|=0$, called as starting node. And there is only one node $n_e\in N$, $|n_e^{\bullet}|=0$, called as ending node.

(2) $\forall n\in N$, $\exists p=\langle n_s,\cdots,n,\cdots n_e\rangle$. The node n must locate in a directed path from the starting node n_s to the

ending node n_e, namely, isolated nodes are not allowed.

(3) $\forall a \in A$, $\left|{}^\bullet a\right| \le 1 \wedge \left|a^\bullet\right| \le 1$; $\forall c \in C$, $(\left|{}^\bullet c\right| > 1 \wedge \left|c^\bullet\right| \le 1) \vee (\left|{}^\bullet c\right| \le 1 \wedge \left|c^\bullet\right| > 1)$. For every activity, the number of its preceding and succeeding node are respectively less than 2, namely, every activity cannot have multiple input links and multiple output links. Whereas, a connector is either a join node $(\left|{}^\bullet c\right| > 1 \wedge \left|c^\bullet\right| \le 1)$ or a split node $(\left|{}^\bullet c\right| \le 1 \wedge \left|c^\bullet\right| > 1)$.

(4) $\forall l \in L$, if $l \in \left\{\langle c, n\rangle \middle| \tau(c) = OR \wedge \left|c^\bullet\right| > 1\right\}$ then $\sum_{i=1}^{\left|c^\bullet\right|} \eta(l_i) = 1$, else $\eta(l) = 1$. Namely, for every split connector whose logical type is "*OR*", the sum of execution probabilities of its output links is equal to 1.

3.2 SEMANTIC RULES

A BPLSM, which satisfies the syntax rules may still have some semantic errors which lead to appear abnormality in execution of business process. To identify the semantic errors in BPLSM, the instance sub-graph and correctness criterion for BPLSM are defined firstly as follows:

Definition 6 (Instance sub-graph).An instance sub-graph represents a subset of activities that may be executed for a particular instance of a business process.

Definition 7 (Correctness criterion for BPLSM). A BPLSM is correct if all instance sub-graph can be executed without exception from starting node to ending node, namely for any one execution of a business process the ending node can be executed only once.

According to the correctness criterion for BPLSM, there are two semantic errors in BPLSM: deadlock and lack of synchronization. A deadlock conflict will be introduced if exclusive choice paths are joined with a synchronizer (a join node whose logical type is "*AND*"). A deadlock at synchronizer blocks the continuation of a business process path since one or more of the preceding transitions of the synchronizer are not triggered. A lack of synchronization will be introduced if concurrent paths are joined with a merge node (a join node whose logical type is "*OR*"). A lack of synchronization at a merge node results into unintentional multiple activation of nodes that follow the merge node.

It is very important to ensure that a business process model has correct logical structure before it is deployed. Because process modelling methods based on graphs themselves do not provide some effective methods, rules or mechanism against errors in logical structure of built BPM, based on the Jin Hyun Son's research [19], modelling rules are introduced and formalized for well-defined business process as follows:

Rule 1: An AND-Split control dependency should have its matching AND-Join control dependency, which forms a correct *AND* model structure.

The rule 1 is formalized as follows: for any $c_i \in C$, $\tau(c_i) = And \wedge \left|c_i^\bullet\right| > 1$, there is a $c_j \in C$, $\tau(c_j) = And \wedge \left|{}^\bullet c_j\right| = \left|c_i^\bullet\right|$. The connectors c_i and c_j satisfy with ① $\forall p = \langle c_i, \cdots, n_e\rangle$, $c_j \in p$; $\forall p = \langle n_s, \cdots, c_j\rangle$, $c_i \in p$; ② $\neg n \in N$, $\forall p = \langle c_i, \cdots, c_j\rangle$, $n \in p$. All nodes and links located in $p = \langle c_i, \cdots, c_j\rangle$ form a parallel model structure $PMS\left\{c_i, c_j\right\}$.

Rule 2: An OR-Split control dependency should have its matching OR-Join control dependency, which forms a correct *OR* or *LOOP* model structure.

The rule 2 is formalized as follows: for any $c_i \in C$, $\tau(c_i) = Or \wedge \left|c_i^\bullet\right| > 1$, there is a $c_j \in C$, $\tau(c_j) = Or \wedge \left|{}^\bullet c_j\right| = \left|c_i^\bullet\right|$. The connectors c_i and c_j must satisfy one of the following two cases:

Case 1: ① $\forall p = \langle c_i, \cdots, n_e\rangle$, $c_j \in p$; $\forall p = \langle n_s, \cdots, c_j\rangle$, $c_i \in p$; ② $\neg n \in N$, $\forall p = \langle c_i, \cdots, c_j\rangle$, $n \in p$. In this case, all nodes and links located in $p = \langle c_i, \cdots, c_j\rangle$ form a *OR* model structure $SMS\left\{c_i, c_j\right\}$.

Case 2: ① $\left|{}^\bullet c_j\right| = \left|c_i^\bullet\right| = 2$, $\exists p = \langle c_i, \cdots, c_j\rangle \wedge \exists p' = \langle c_j, \cdots, c_i\rangle$; ② $\forall n \in p$, $n \ne ci \wedge n \ne cj$, $\overleftarrow{n} \in p$, $\vec{n} \in p$ Where $\overleftarrow{n} \in {}^\bullet n$, $\vec{n} \in n^\bullet$, $p = \langle c_i, \cdots, c_j\rangle$ or $p = \langle c_j, \cdots, c_i\rangle$. In this case, all nodes and links located in $p = \langle c_i, \cdots, c_j\rangle$ and $p = \langle c_j, \cdots, c_i\rangle$ form a *LOOP* model structure $LMS\left\{c_i, c_j\right\}$.

Rule 3: an activity in the model structure can be replaced by a basic model structure to form a new model structure.

Rules 1 and 2 are match rules. Rule 3 is replacing or nesting rule.

Definition 8 (nested model structure)**.** A nested model structure is a model structure that contains non-sequential model structures.

Definition 9 (basic model structure)**.** A basic model structure is a model structure that does not contain any non-sequential model structures.

The simplest BPLSM which has only one activity node is shown as Figure 2. Based on the simplest model, a complex BPLSM can be usually formed by nesting four kinds of basic model structures (*SEQ*, *LOOP*, *AND*), and *OR*) according to the rules 1-3. Figure 3 shows an example to form a complex BPLSM.

Definition 10 (well-defined BPLSM)**.** A well-defined BPLSM is a BPLSM that is formed by nesting four kinds of basic model structures (*SEQ*, *LOOP*, *AND*), and *OR*) based on a single activity according to the rules 1-3.

Theorem 1. A well-defined BPLSM must be correct in logical structure.

Proof: firstly, there are no errors in syntax. Secondly, because the process of forming well-defined BPLSM is reversible, any well-defined BPLSM can be simplified to a single active by the inverse process. This single active is both stating node and ending node, thus when business process can be executed once the ending node will be executed only once. According to definition 7 this BPLSM is correct.

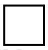


FIGURE 2 The simplest BPLSM, which has only one activity node

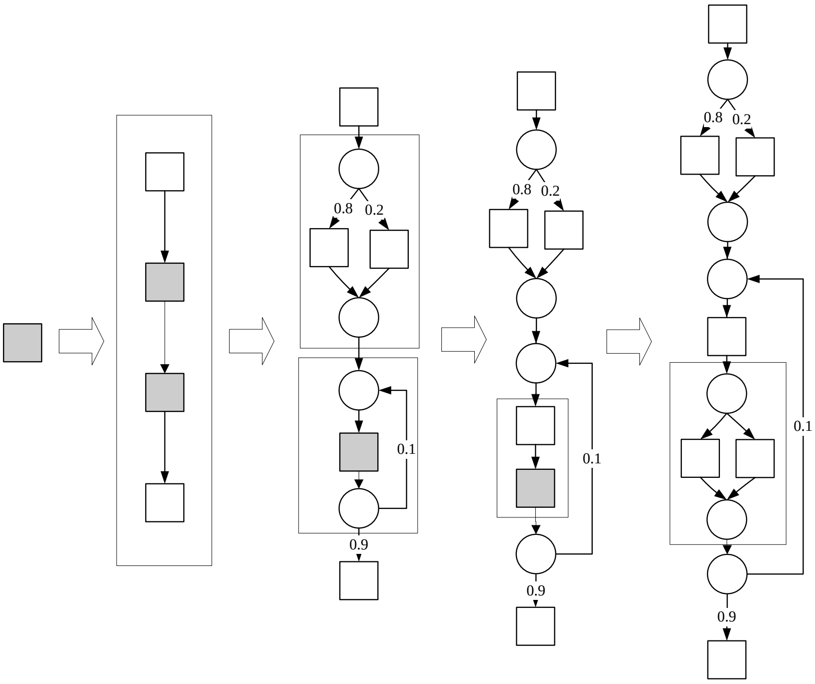


FIGURE 3 A process of forming a complex BPLSM from the simplest BPLSM

3.3 CHECKING ALGORITHM

It is not necessary to design the special algorithm for checking whether the BPLSM meet these simple syntax rules. However, the semantic rules are more complex than the syntax rules, the algorithm for checking whether the BPLSM meet the match and nesting rules are developed as follows:

Algorithm 1:

Step1: If the BPLSM is a basic sequence model structure then proceed to Step5, else proceed to Step2.

Step2: If $|JN| = |SN| > 0$, where JN is the set of join node and SN is the set of split node, then proceed to Step3, else proceed to Step6.

Step3: if the set of join node JN is not null then find out a join node jn_i and proceed to Step4, else proceed to Step5.

Step4: According to rules 1-3 find the split node sn_j from SN which matches with the join node jn_i. if the split node sn_j exist then jn_i, sn_j, and their matching type are recorded, and they are deleted from JN and SN respectively, proceed to Step3, else proceed to Step6.

Step5: The BPLSM meet the match and nesting rules, and output matching connectors. The algorithm terminates.

Step6: The BPLSM does not meet the match and nesting rules. The algorithm terminates.

To estimate the efficiency of this algorithm, "Big-O" is used to estimate its time complexity. In particular, the time complexity of the proposed algorithm will be $O(\frac{\lfloor |C|/2 \rfloor^2}{2})$ in the worst case. Thus, the developed algorithm can complete verification of well-defined BPLSM in very short time, and can be applied to verifying the large complicated business process models in distributed real-time interactive environment.

4 Conclusion

Because traditional graphical process modelling methods lack efficiency mechanisms or rules to ensure correctness of the logical structure during business process modelling, they need additional methods to verify its correctness of the logic structure after the business process model is established. Therefore, the well-formed business process modelling mechanism is researched. The semantic and syntactic rules are presents for the correctness of business process logic structure model, and the algorithm for checking if the model meets these rules is proposed. The modelling mechanism has been applied in our business process scheduling optimization system with integration of modelling and simulation. Compared with conventional verification methods, for example, Petri-net reduction techniques, integer programming, process logic, π calculus, they have the following characteristics:

- The matching and nesting rules are simple and effective, and they are easy to use and master. And the correctness of the logical structure can be ensured by employing these rules.
- Model transformations is not required for checking if business process models satisfy the matching and nesting rules to verify their correctness.
- The algorithm for check rules is simple and its time complexity is small. It can avoid the sharp decline in the performance of algorithm with the growing process model and increasing node. Thus, it can be applied to verify the correctness of the large complicated business process in distributed real-time interactive environment.

Future studies can be done to enhance flexibility of modelling and presentation power of model, by incorporating some unstructured model structure into the well-formed business process modelling mechanism.

Acknowledgments

The authors would like to Ministry of Education of China, Zhejiang Provincial Natural Science Foundation of China, and Zhejiang Provincial Science Technology Department of china for financially supporting this research under the Grants Nos.11YJA630161, Y1111039 and 2014C31031.

References

[1] Touré F, Baïna K, Benali K 2008 An efficient algorithm for workflow graph structural verification. Proc. Of the Int. Conf. on Cooperative Information Systems *Lecture Notes in Computer Science* **5331** 392-408

[2] Eshuis Rik, Kumar Akhil 2010 An integer programming based approach for verification and diagnosis of workflows *Data & Knowledge Engineering* **69**(2010) 816-35

[3] Giouris A, Wallace M 2007 Graph Based Workflow Validation *Artificial Intelligence and Innovations* **247** 45-53

[4] Morimoto Shoichi 2008 A Survey of Formal Verification for Business Process Modeling. ICCS2008 *Lecture Notes in Computer Science* **5102** 514-22

[5] Van der Aalst W M P, Ter Hofstedev A H M 2000 Verification Of Workflow Task Structures: A Petri-net-based Approach *Information Systems* **25**(1) 43-69

[6] Naijing Hu, Liang Zhao, Hu Jinhua 2007 Verification of Evolution Rules on Workflow Net Based on Petri-Net *Journal of Chinese Computer Systems* **28**(6) 1076-9

[7] Jiantao Zhou, Meilin Shi, Xinming Ye 2005 A Method for Semantic Verification of Workflow Processes Based on Petri Net Reduction Technique *Journal of Software* **16**(7) 1242-51

[8] Liang Zhang, Shuzhen Yao 2007 Research on Correctness of Workflow Model Based on Petri Nets Reduction Techniques *Computer Engineering* **3**(9) 60-1

[9] Zhengjun Dang, Zhongjun Du 2011 Improved Algorithm combining graph-reduction and graph-search for workflow verification *Computer Engineering and Applications* **47**(4) 226-8

[10] Sadiq W, Orlwska M E 2000 Analysing process models using graph reduction techniques *Information Systems* **25**(2) 117-34

[11] Ke Ning Qing Li, Yuliu Chen 2005 Verification of IDEF3 process models *J Tsinghua Univ (Sci&Tech)* **45**(4) 540-4

[12] Bi H H, Zhao J L 2004 Process logic for verifying the correctness of business process models *Proc. of the 25th International Conference on Information Systems (2004ICIS)*.Washington, D.C., USA 91-100

[13] Abouzaid Faisal, Mullins John 2008 A Calculus for Generation, Verification and Refinement of BPEL Specifications *Electronic Notes in Theoretical Computer Science* **200**(2008) 43-65

[14] Hong Ling, Jiangbo Zhou, Zhengchuan Xu 2006 Semantic deduction-based workflow structure verification method *Computer Integrated Manufacturing Systems* **12**(6) 893-8

[15] Choi Y, Zhao J 2002 Matrix-based abstraction and verification for E-business processes *Proc. of the 1st Workshop on e-Business* Barcelona, Spain 154-65

[16] Liu R, Kumar A 2005 An analysis and taxonomy of unstructured workflows *Proc. of the 3rd Conference on Business Process Management (BPM 2005)* Lecture Notes in Computer Science **3649** 268-84

[17] Van der Aalst W M P, Hirnschall A, Verbeek H M W 2002 An Alternative Way to Analyse Workflow Graphs *Proc. of the 14th Int. Conf. on Advanced Information Systems Engineering (CAiSE'02)* Lecture Notes in Computer Science **2348** 535-52

[18] Sivaraman E, Kamath M 2002 On the use of Petri nets for business process modelling *11th Annual Industrial Engineering Research Conference*, Orlando, Florida

[19] Hyun Son Jin, Sun Kim Jung, Ho Kim Myoung 2005 Extracting the workflow critical path form the extended well-formed workflow schema *Journal of Computer and System Sciences* **70**(2005) 86-106

[20] Sheng Liu, Yushun Fan, Huiping Lin 2009 Dwelling time probability density distribution of instances in a workflow model. *Computer & Industrial Engineering* **57**(2009) 874-9

[21] Zuoxian Nie, Xinhua Jiang, Jiancheng Liu, Haiyan Yang 2009 Performance analysis for instances of generalized well-formed workflow *Computer Integrated Manufacturing Systems* **15**(12) 2424-31

[22] Liqin Tian, Chuang Lin 2003 Method for computing performance of a kind of workflow models nested by basic model *Acta Elect ronica Sinaca* **31**(12A) 2167-70

Authors

Kai Chen, born on April 26, 1968, Zhejiang China

Current position, grades: senior engineer at Lishui University, Lishui, China
University studies: M.S. degrees in Technology of Computer Application from Hangzhou Dianzi University in 2007
Scientific interest: system modelling and optimization, mechanical and electrical engineering and related control techniques

Yi Xie, born on July 23, 1975, Zhejiang China

Current position, grades: professor of Zhejiang Gongshang University, Hangzhou, China
University studies: M.S. and Ph.D. degrees in Mechanical Engineering from Zhejiang University in 2003, and 2011, respectively
Scientific interest: workflow, business process management, scheduling and optimization

Research on wisdom urban public security management system integrated into the situation of urban safety

Luya Wang[1], Liang Xiao[2,3]*

[1] *School of science, Zhejiang Sci-Tech University, Private bag 310012, Hangzhou, China*

[2] *Centre for Studies of Modern Business Zhejiang Gong Shang University, Private bag 310015, Hangzhou, China*

[3] *School of Business Administration, Zhejiang Gong Shang University, Private bag 310015, Hangzhou, China*

Received 1 March 2014, www.tsi.lv

Abstract

With the expanding of the sizes of the cities, the urban population and property space distribution becomes more concentrated, urban public safety incidents into the increasingly frequent stage. How to intelligent and efficient manage the urban public safety is imminently. On the basis of defining the urban security situation management model systematic, this article will establish the urban safety stratified hierarchical data acquisition of internet of things which is based on urban monomer-group region, study the tracking-summarized-warning-optimization handling mechanism which support the city security complex event, construct the wisdom urban public security management system which is integrated into urban security situation and provide an effective means to realize the wisdom management of the city public security.

Keywords: public security, safety situation, wisdom city, internet of things

1 Introduction

With the acceleration of the process of Chinas urbanization and the expanding of the sizes of the cities the space distribution of urban population and property become more concentrated which bring a few challenges to urban public security management. Statistical material shows that Chinas annual economic losses due to public security issues are about 650 billion Yuan, accounting for 6% of the total GDP. In order to build an efficient urban public security management system, many scholars and enterprises conduct a large number of explorations respectively from theoretical and practical aspects. Study from the theoretical point of view, like Chengyu Zhan set Beijing for example, raised the urban emergency response system which is to prevent and control the uncertainty of public risk [1]. Rongzhi think that to manage urban risk, which is often complex and likely to go out of control, it is highly necessary to establish integrated and highly-effective risk control system and public security management system so that emergency response and post-accident management would give way to the proactive public risk management system [2]. Through comparative study of urban public safety management system between Japan and the United States, Chen Hua put forward the four stages of urban disaster emergency management system which including disaster prevention, disaster early warning before the response, post-disaster recovery and after monitoring assessment [3]. From the practical application perspective, after announce The advice of further carry out the construction of Ping An by the central politics and law committee, the central social security comprehensive management committee in December 2005, the construction of urban public safety management system has made outstanding progress [4-6]. By layout all kinds of safety monitoring terminal equipment in the major hazards and social public facilities, the government has initially established an internet of things system which can real-time security monitoring and early warning the urban public facilities and source of danger [7]. And on this basis, the government has initially establish the urban public safety management system to different industries, and to a certain extent, realize the tracking, early warning, analysis and rescue to urban public security.

However, the current construction of the urban public security management system still has the following problems: first of all, the existing system is mostly set up by industry, which leads to relative isolate between application system, interactive coordination is not enough, and the source of data between different application systems has a typical distributed heterogeneous data characteristic. The system also faced with some problems, such as scattered, isolated, single and the effective integration of heterogeneous data resources [8].Second, most of the existing system is a kind of extensive static management and affairs management, but as the rapidly application of the internet of things technology, it is possible to realize the real-time tracking and accurate data collection of urban public security object. Third, what the existing system mostly considered is unidirectional management, especially in

* ***Corresponding author*** e-mail: hz-sigma@vip.sina.com.cn

the process of urban public safety emergency rescue, the system generally acquisition the scene of the accident data simply and unidirectional. But with the development of the internet of things, how to feedback all kinds of safety information to the rescue site by the internet of things in the process of emergency according to the requirements and characteristics of the new period of urban public safety management, this paper has put forward the architecture of the wisdom urban public security management system which is integrated into the situation of urban safety, and mainly studies the stratified hierarchical safety data collection network, the urban security situation management model, the urban security incident two-way processing mechanism, the urban safety data exchange framework and some other key technologies [9].

2 System framework

The wisdom urban public security management system framework this paper has put forward is shown in Fig.1. This system includes facilities layer, network layer, data layer, function layer and application layer, its corn is to layout different types of the internet of things facilities near the urban safety management object and bring different kinds of urban safety management object into the range of the wisdom urban public safety management framework. Moreover, with the support of the urban public safety management system to realize whole journal, dynamic, intelligent and fine management of the urban security management object according to different industry application requirements [10]. Among the layers, the facilities layer refers to different kinds of equipment and facilities, which can conduct real-time tracking and data collection to different sites. The network layer is refers to network technology which support the safety data highly speed acquisition and reliable transmission. The data layer introduced into the safety situation ontology model and then put forward the complex urban security environment-safety management object-safety resources mapping mechanism, established the heterogeneous distribution of urban public safety data integration view, which integrated the urban safety situation, and urban public safety data exchange framework. The function layer includes safety early warning management, event location management, safety plan management, field rescue management, resource scheduling management and statistical analysis management. The application layer combined with application requirements from different industries, relying on all kinds of application systems developed by urban public safety management platform.

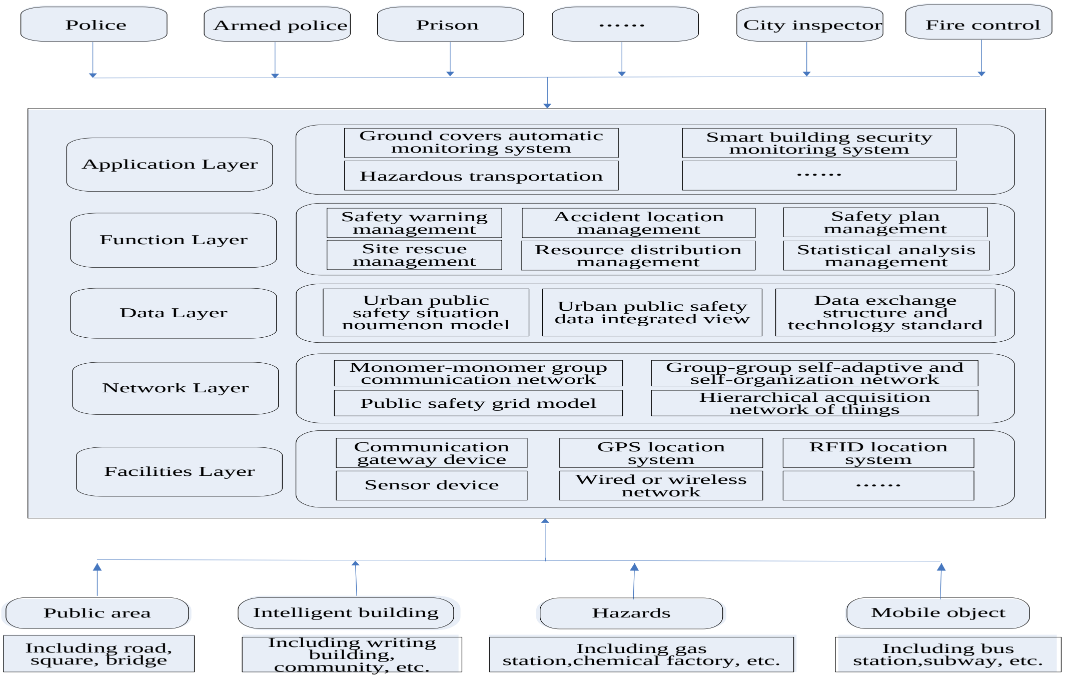

Police
Armed police
Prison
……
City inspector
Fire control
Application Layer
Ground covers automatic monitoring system
Smart building security monitoring system
Hazardous transportation
……
Function Layer
Safety warning management
Accident location management
Safety plan management
Site rescue management
Resource distribution management
Statistical analysis management
Data Layer
Urban public safety situation noumenon model
Urban public safety data integrated view
Data exchange structure and technology standard
Network Layer
Monomer-monomer group communication network
Group-group self-adaptive and self-organization network
Public safety grid model
Hierarchical acquisition network of things
Facilities Layer
Communication gateway device
GPS location system
RFID location system
Sensor device
Wired or wireless network
……
Public area
Intelligent building
Hazards
Mobile object
Including road, square, bridge
Including writing building, community, etc.
Including gas station,chemical factory, etc.
Including bus station,subway, etc.

FIGURE 1 Intelligent urban public security management system structure

3 The key technology and method study

3.1 THE URBAN PUBLIC SAFETY DATA HIERARCHICAL COLLECTION NETWORK CONSTRUCTION TECHNOLOGY.

On the one hand, the urban public safety management environment is getting more and more complicated, such as the typhoon and other natural disasters, the harmful gas leakage other safety accidents, both have all kinds of buildings distributed in the centre of the city where traffic flow is concentrated, and have bridges, and other public infrastructures distributed in the complex geological environment [11]. On the other hand, the complexity of urban safety management object may make different urban safety node data acquisition rate inconsistent.

Therefore, using conventional uniform method to deploy the sensor node may lead to inefficient of energy consumption of the sensor node in the region even premature failure and cause the urban public safety things united network disconnected or produce hollow. In addition, we found from practice that the data obtained from the monitoring of some adjacent sensors to the same environment factor from the same urban safety node. Therefore, in order to reduce energy consumption and improve the quality of the data communication, it is necessary to deal with the data comprehensively, which is monitored from multiple adjacent sensor nodes. According to the characteristics of urban public safety data information collection and transmission, this system using grid optimization method to division the urban public safety area, establish the urban public safety group, which support the communication between safety node within the scope of particular area. The urban public safety data collection network, which is based on the urban security node monomer-security groups and security zone, has the feature of self-adaptive self-organizing on data collection and transmission, and it supports the dynamic configuration of urban public safety data monitoring physical quantities and monitoring point. Among them, the safety data collection network sensor node cannot only realize the function of safe environment automatically data acquisition but also realize the function of forwarding and self-checking the safe data, support to send out the acquisition safety situation data. The safety data collection network structure based on grid division method, on one hand, puts forward a fascicles topology system which can minimize information transmission volume, reduces and evenly distribute node power consumption, suits for wireless sensor network, supports dynamic monitoring of multi-situation factors in urban public security environment, on the other hand, develops an urban public safety things networking grid partition algorithm which shows a special 3d layout and adapts to kinds of obstacles, satisfies regional connection and coverage of.

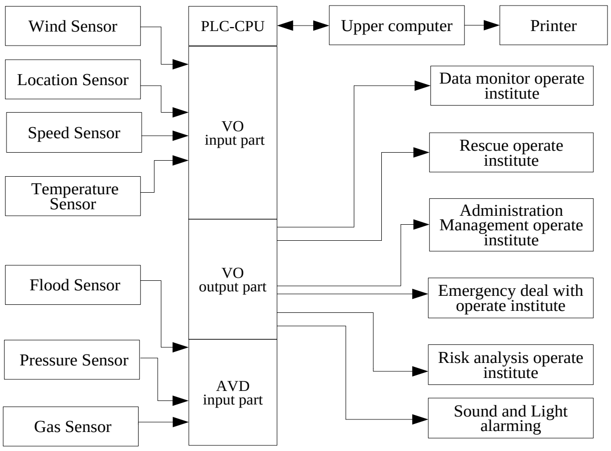


FIGURE 2 Urban public safety situation data acquisition network (1)

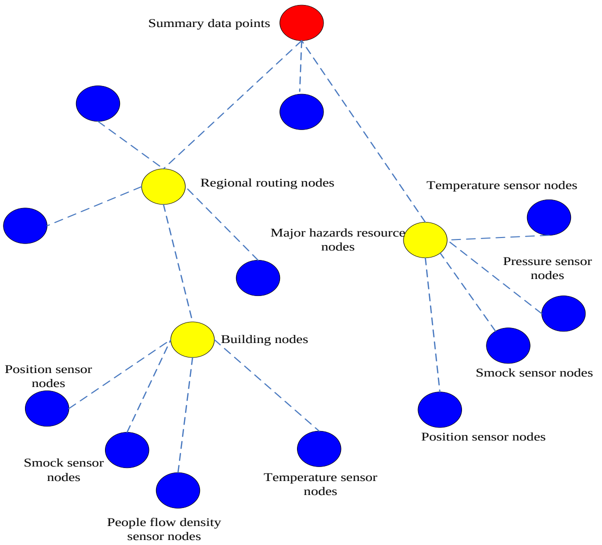


FIGURE 3 Urban public safety situation data acquisition network (2)

3.2 URBAN PUBLIC SAFETY SITUATION MODEL AND MANAGEMENT METHOD.

The situation is often used to describe a variety of internal and external environment information faced in the process of a physical activity. Urban public safety incident belongs to the typical unconventional emergencies; the situation information related to it the urban public safety incident occurred faced. The urban public security ontology model is an effective means to accurately describe or portray all kinds of complicated environment information the urban public safety management object faced. By constructing urban public safety situation ontology model the government can build urban public safety environment-public safety management object-public security resources mapping mechanism and accurate description all kinds of social attributes and real-time status information of the urban public safety management objects.

According to the content of the urban public security management object and the features of the urban public security incident decision-making, this paper will abstract summarize the urban public safety situation for rescue resources situation, the scene of the accident situation, surrounding risk situation three categories of situation factors. Among the three situations, the rescue resources situation refers to the resources situation, which can service urban public safety management object and is available for dispatch at the time of accident. The accident site situation is point to the indicator, which can describe the comprehensive state of urban public safety accident site, in order to make sure they are effective protected when accident occurred; the site staff situation, which reflects the space distribution of the victims and potential victims of the accident site. The surrounding risk situation refers to spatial distribution state and related information of the property, the staff and the dangerous source in 1km area around the scene of the accident [12].

According to the definition of urban public safety situation, the situation ontology model and the sample are showed as below. Among them, the relationship Subclass-of reflect the father and son relationship or inheritance relationship between different subclass in situation domain ontology, subclass situation can inherit his father situation properties and extension appropriately; the relationship Attribute-of reflect the affiliation between situation subclass concept and subclass attribute, the relationship Instance-of reflect the assignment type of a particular attribute of the subclass situation. When all of the attribute of a situation subclass have been assigned, you can get a group of assignment, which reflect the characteristics of this situation, called situation instance, the function create the situation subclass instance is called situation function. Obviously in order to realize the structure description of the situation, we need to create the situation function to structure deal with the distribution task situation. As follows:

3.3 THE TWO-WAY FEEDBACK MECHANISM AND KEY METHOD FOR SUPPORTING THE URBAN PUBLIC SAFETY INCIDENT EMERGENCY MANAGEMENT [11]

The two-way feedback system of urban public security management refers to in the process of dealing with the urban public safety incident, on the one hand, feedback all kinds of instruction information to scene of urban public safety accident by the internet of things system, like the emergency rescue instruction lamps in the high-rise buildings, to help the accident staff to conduct self-rescue; [13] On the other hand, tracking and monitoring the process status of the accident, collecting all kinds of process status data of the public safety accident site by remote start high performance wireless sensor equipment set up in the accident site, and then feedback to the urban public safety management system to guide the rescue teams to adjust and optimize the emergency rescue plan.

The system including four key feedback mechanisms and methods, as followed: the self-adaptive mechanism of safety intelligent management system, which is based on the two-way feedback of urban public safety data. Focus on researching the process data change model of all of the monitor nodes of urban public security internet of things, set the corresponding threshold, and start the emergency rescue terminal equipment. When the internet of things terminal equipment of original site suffered damage, it can research the urban public security internet of things emergency data collection system by applying the sequence cut method. The self-leaning mechanism of safety intelligent management system, which based on the two-way feedback of urban public safety data. Its kernel is to design self-learning mechanism through CAS theory, so that the urban public security intelligent system can proceed self-leaning and processing and form a new leaning experience which means the network structure and weight, according to the historical data of the safety of things data centre and the processing state data of current safety incidents and apply the urban public security model and Bayes network. The self-adjustment mechanism of safety intelligent management system, which based on the two-way feedback of urban public safety data. Its corn is to generate the simulation matrix, apply the best projection direction and other method to evaluate the degree of the accident, and by use the orderly composite strategy of the maximal frequent item sets mining method over data stream, an improved clustering algorithm for dynamic data based on principal component analysis and density and other method for analysis and mining the state information of the accident site according to the five evolution processes and characteristics of the urban public accidents. The self-optimization mechanism of safety intelligent management system, which based on the two-way feedback of urban public safety data. Its corn is to realize automatic optimization of the urban security resource rescue plan based on SVM and finally find out the feasible technical route which can dissolve the urban accident resource rescue to support the self-optimize of the urban public security intelligence management plan, designing the Agent-DEVS model group which is the emergency rescue plan that can dissolve the urban accident and build the emergency resource rescue collaborative environment which is based on the HLA.

4 The positive research based on a certain company product

The design thought of this system has been preliminary reflected on the relevant prototype products in a certain information technology co., LTD. At present, related application cases are including Xinglin-Bay Business Parkis intelligent systems engineering, Zhejiang south lake prison security system, Sichuan environmental monitoring centre construction project monitoring and emergency command engineering project. Although the application fields have some differences, they both have innovations below. Support the management of monitoring and pre-warning urban safety hazards. Improve the level of alarming, monitoring, pre-warning and supervision through establishing the system of hazards data collecting and monitoring. Support the collection and track the safety data network. Combine next-generation internet and wireless sensor network together, and connect urban management departments at various levels, each unit grid and each city parts. Support intelligent dispatch of security rescue. The results of this project can realize remote emergency command dispatching within GIS, integrated positioning system, signal monitor system and GWSN [14].

Take the development of a city construction safety management platform by demonstration application units as an example. This platform takes the grid management method, marks the important building, which may occur safety accident within the scope of management area as different types of urban public security node, builds the city public safety date acquisition network by modern information technology. Combined with the actual demand of urban public security management and monitoring, the urban safety management grid is divided into business grid and geographic grid in the implementation process of demonstration project. Imaginary arcs represent the logical correspondence between business grid and geographic grid.

On this basis, Project has proposed and implemented urban public security management platform architecture, basing grid management. This platform adopts the hybrid network design of three video networking structure and analogy digital video, and makes full use of existing mature wired/ wireless technology, achieves real-time acquisition and analysis of urban public safety data, realizes urban public security management tasks of urban public safety data exchange, intelligence warning and analysis of security incidents, security incidents intelligent rescue and disposal [15]. As follows

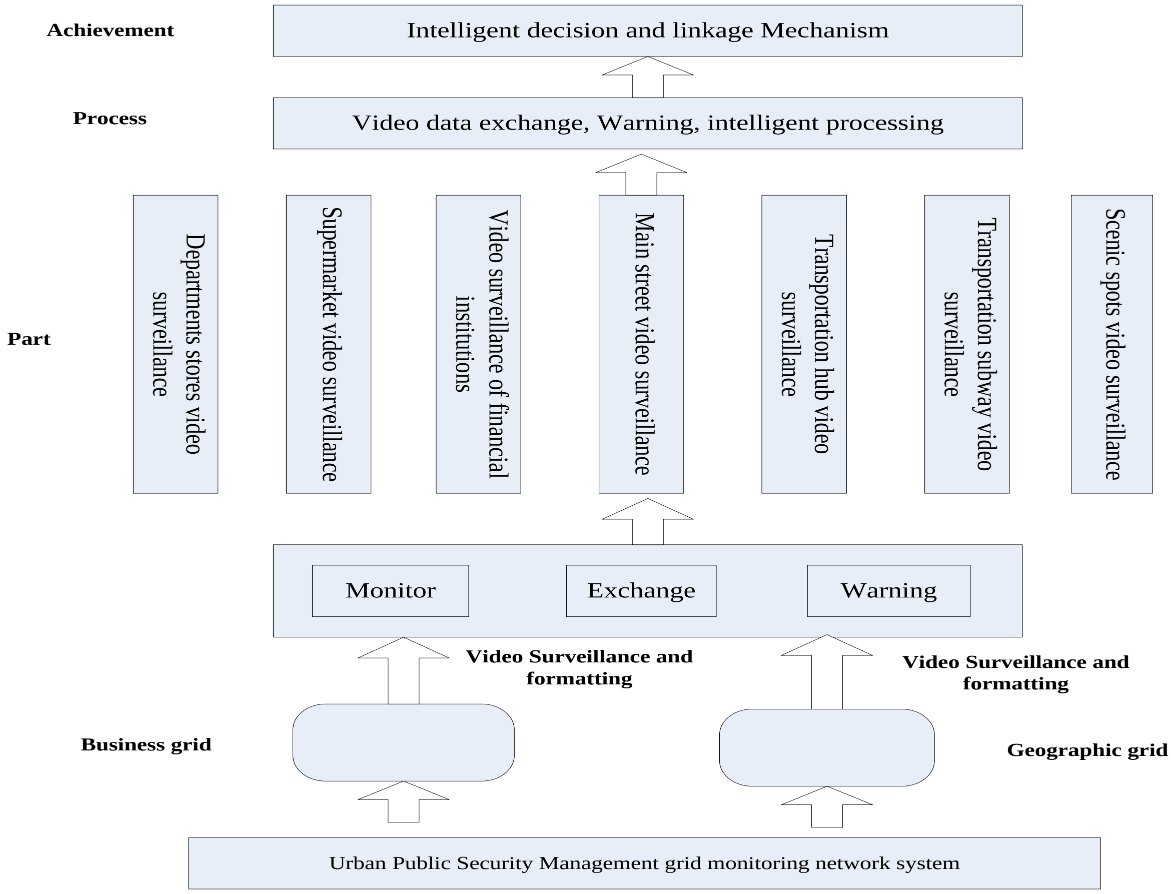

Achievement
Intelligent decision and linkage Mechanism
Process
Video data exchange, Warning, intelligent processing
Part
Departments stores video surveillance
Supermarket video surveillance
Video surveillance of financial institutions
Main street video surveillance
Transportation hub video surveillance
Transportation subway video surveillance
Scenic spots video surveillance
Monitor
Exchange
Warning
Video Surveillance and formatting
Video Surveillance and formatting
Business grid
Geographic grid
Urban Public Security Management grid monitoring network system

FIGURE 4 Urban public security management platform architecture

Among this, urban public security event early warning divides two parts, first is establishing early warning process, which includes finding urban security alert, seeking security event source, analysing security situation, confirming security event importance and starting alarming. Second is developing a comprehensive index to measure the total alert degree through the establishment of monitoring index system. Meanwhile, divide the alert interval into five areas, safety area, light warning area, moderate warning area, serious warning area and severe warning area, arranging the corresponding early warning plan aimed at every area, the details see Figure 6. Urban public security event linkage disposal and rescue command management. Adopting the model of receive unified, dispose classified, large alarm system, establishing an unified command scheduling management system which is oriented by first class monitoring command centre, supported by second and third class monitoring centre [16]. This system includes receiving and disposing command scheduling module based on computer network, wired and wireless communications and other systems, rescue information repository supported by inputting, maintaining, updating, sharing and dispatching the urban public security event disposal method, realizing functional feedback between safety incident site and command management centre by all kinds of Internet of things technology, and also proceeding rescue command management system of urban public security event intelligent disposal, the details see Figure 7.

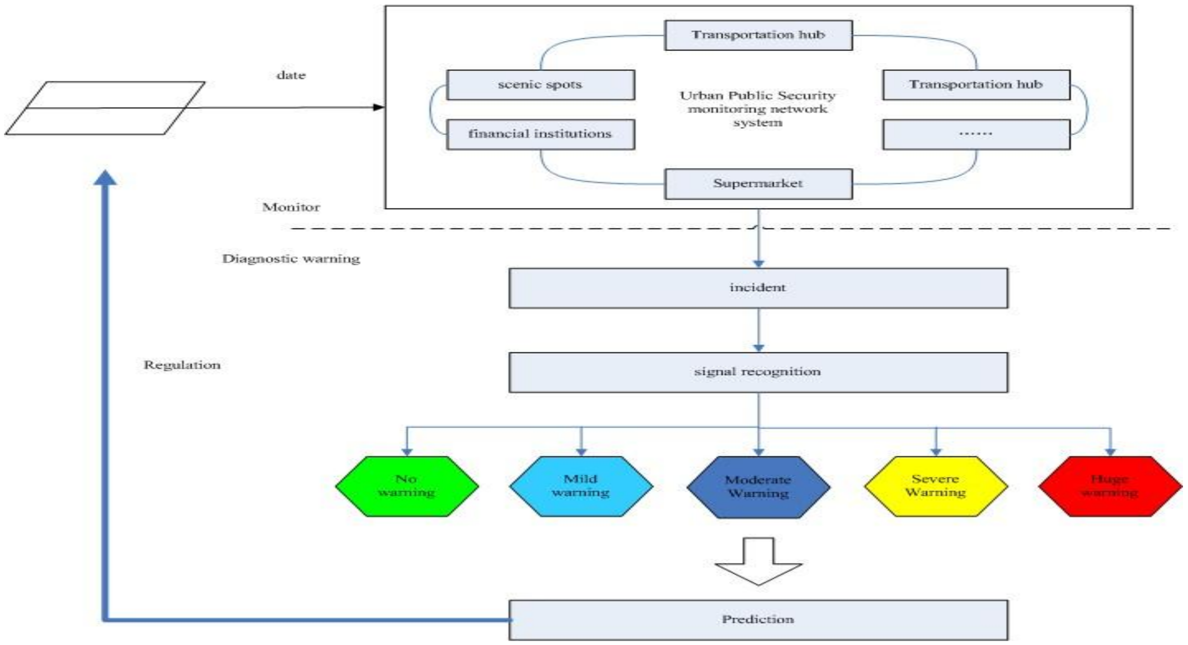


FIGURE 5 Urban public security event early warning model

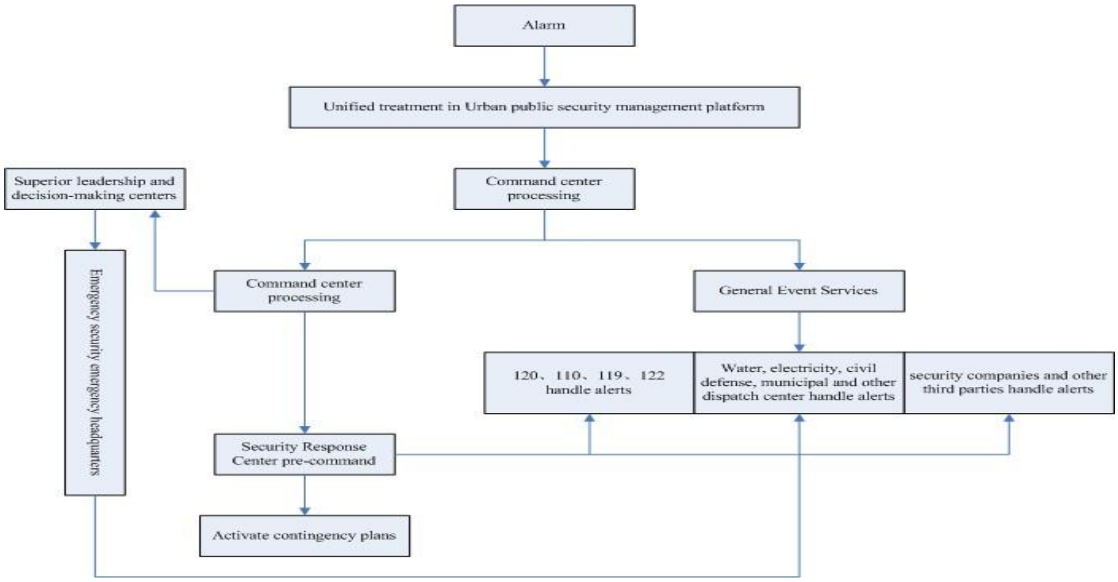

Alarm
Unified treatment in Urban public security management platform
Superior leadership and decision-making centers
Command center processing
Emergency security emergency headquarters
Command center processing
General Event Services
120、110、119、122 handle alerts
Water, electricity, civil defense, municipal and other dispatch center handle alerts
security companies and other third parties handle alerts
Security Response Center pre-command
Activate contingency plans

FIGURE 6 Urban public security event linkage disposal and rescue command management model

5 Conclusions

According to the demand and characteristic of urban public safety management, this paper put forward smart urban public safety management system, which is fit in urban security situation. The target is to establish an urban public safety intelligent management system, which is including data collecting of Internet of things security, safety situation ontology, security data exchange, security emergency management and security data analysis. Then we mainly put forward and studied urban safety data stratified hierarchical collection of Internet of things which is based on monomer- group-region grid structure, safety situation ontology model which support normalized description of urban security resources and management object in complex environment, handling mechanism of self-adaptive-study-adjust-optimize which support urban public safety complex event emergency management. This system can not only help prevent urban public security incident from occurring, but also understand the security incident spot situation. Besides, it can also help improving public security incident.

Acknowledgment

This paper was supported by the National Natural Science Foundation of China (Project Number: 71001088), Ministry of Education Humanities and social science research base project(13JJD630017),Natural Science Foundation of Zhejiang Province (Project

Number Z13G010002, LZ13G010001), and Public Technology Research and Social Development of Zhejiang Province (Project Number: 2012C23125), and Zhijiang Youth Social Science Scholars Action Plan (G178,13ZJQN059YB). The author is grateful to the anonymous referees for their valuable comments and suggestions to improve the presentation of this paper.

References

[1] Cheng Yu Zhan, Cheng Wei Li 2010 Process and Trends: The Development Research of Emergency Response System in Beijing *Social Science of Beijing* (5) 20-6

[2] RONG Zhi 2012 Theorizing on Urban Safety and Risk Control System: with Insight from Shanghai Expo *Shanghai University (Social science Edition)* **29**(3) 116-29

[3] Chen hua 2010 The International Comparative Study of The Modern Urban Disaster Management Based on The Efficiency *Modernization of Management* (3) 59-61

[4] Lu Yong 2011 Public Security Management in the Context of the Internet of Things *Urban Problems* (2) 80-4

[5] Tong Yang, Benching Shia, Jinrui Wei, Kuangnan Fang 2012 Mass Data Analysis and Forecasting Based on Cloud Computing *Journal of Software* **7**(10) 2189-95

[6] Baomin Xu, Ning Wang, Chunyan Li 2011 A Cloud Computer Infrastructure on Heterogeneous Computing Resources *Journal of Computer* **6**(8) 1789-96

[7] Xue Li Wang 2012 Problems of Urban Security System and Countermeasure to them *Urban Problems* (7) 79-83

[8] Komninos N 2009 Intelligent cities: Towards interactive and global innovation environments *International Journal of Innovation and Regional Development* **1**(3) 337-55

[9] CHEN You-liang 2011 Research on risk investigation system on urban public supervision object *Journal of Safety Science and Technology* **7**(7) 79-82

[10] Hu Zhilaing, Gao xaingduo 2012 The Construction and Planning Application of City public Security Infrastructure under the Comprehensive Disaster Prevention Concept *Area Research and Development* **31**(2) 49-53

[11] Ingrid Nielsen, Russell Smyth 2009 Perception of Public Security in Post-reform Urban China: A Routine Activity Analysis *Asian Journal of Criminology* **4**(2) 145-63

[12] Qihai Zhou, Xuejun Zhang 2012 Special Issue on Current Research in Computer Science and Information Technology *Journal of Computer* **7**(7) 1543-4

[13] Zoltan Csefalvay 2011 Gated Communities for Security or Prestige? A Public Choice Approach and the Case of Budapest *International Journal of Urban and Regional Research* **35**(4) 735-52

[14] Guoping Zhang, Wentao Gong 2011 The Research of Access Control Based on UCON in the Internet of Things *Journal of Software* **6**(4) 724-31

[15] Weipeng Liu, Jun Hu, Xing Zhang 2009 A Novel Secure Terminal System Based on Trusted Hardware: U-Key *Journal of Internet* **4**(3) 222-9

[16] Yu Zhang, Shenghui Wang, Ke Xiong, Zhengding Qiu, Dongmei Sun 2010 DPCM Computer for Real-Time Logging While Drilling Data *Journal of Software* **5**(3) 280-7

Authors

Luya Wang, born in 1979, Hangzhou, China

Current position, grades: A lecturer in Zhejiang SCI-Tech University, Hangzhou, China
University studies: Zhejiang SCI-Tech University, Hangzhou, China.
Scientific interest: Knowledge Management, Educational Informationization and Information Management System.
Experience: LuYa Wang, female, master degree. Now, she is a lecturer in Zhejiang SCI-Tech University, Hangzhou, China. Her research interests include Knowledge Management, Educational Informationization and Information Management System.

Liang Xiao, born in 1976, Hangzhou, China

Current position, grades: A professor at Industry and Commerce Administration in Zhejiang Gong Shang University
University studies: received his Ph.D. degree in management science and engineering from Zhejiang University, Hangzhou, China.
Scientific interest: logistics & supply chain management, information management system and marketing channel.
Experience: Xiao Liang male, received his Ph.D. degree in management science and engineering from Zhejiang University, Hangzhou, China. Now, he is a professor at Industry and Commerce Administration in Zhejiang Gong Shang University, Hangzhou, China. His research interests include logistics & supply chain management, information management system and marketing channel.

Study on the difference of urban heat island defined by brightness temperature and land surface temperature retrieved by RS technology

Wenxia Qiu[1, 2], Huixi Xu[1, 2], Zhengwei He[1]*

[1] *Key Laboratory of Geo-special Information Technology, Ministry of Land and Resources, Chengdu University of Technology, Chengdu 610059, China*

[2] *Institute of Engineering Surveying, Sichuan College of Architectural Technology, Deyang 618000, China*

Received 1 March 2014, www.tsi.lv

Abstract

At present, the Remote Sensing is the most advantage method of studying on the Urban Heat Island (UHI) from the space. In general, the method uses remote sensing images to inverse the brightness temperature or land surface temperature to define the UHI. But have any differences of UHI defined by the two kinds of temperature? And what are the differences? This problem is rarely being studied now. Based on this, the brightness temperature (BT) and the land surface temperature (LST) of the Chengdu City were retrieved using Landsat ETM+ image obtained on July 30, 2006. And then, the differences of UHI defined by the BT and the LST were studied from three aspects, which were temperature value, temperature classification and heat island intensity respectively. Research result are the following: (1) There were some differences between BT and LST, and the variation level of LST was higher than BT. (2)There was a slight difference only on the area covered by the low temperature and the secondary low temperature, and the area covered by the others was equal. Therefore, there was no difference on the area of UHI defined by BT and LST. (3) The UHI intensity defined by LST was slightly higher than that was defined by BT, and the intensity value was determined by the method used.

Keywords: Urban Heat Island (UHI), Brightness Temperature (BT), Land Surface Temperature (LST), Remote Sensing Technology (RS)

1 Introduction

Since Lake Howard discovered the temperature of urban centre in London was higher than that in suburban and proposed the concept of “Urban Heat Island (UHI)” in 1833 [1.2], the research in this area has been intense. The research methods [3] included meteorological data analysis, sit observation, numerical simulation and remote sensing. Above methods, the remote sensing, for large observation range, many times, high-speed, good dynamic performance, low cost, strong spatial response ability, etc, in recent years has become the mainstream method of UHI research [4, 5]. It mainly uses thermal infrared images to inverse the brightness temperature or land surface temperature to study the UHI [6, 7]. Then, have any differences of UHI defined by brightness temperature and land surface temperature? And what are the differences? This problem is rarely being studied now.

Based on this, the brightness temperature and the land surface temperature of study area of the Chengdu city were retrieved, and the difference of UHI defined by them was studied from the temperature value, the temperature grade and the urban heat island intensity, respectively.

2 Research Region and Data Source

The Chengdu City is the capital of the Sichuan Province of China, which is the centre of policy, economy and culture and the hub of communications and transportation in the southwest of China. The longitude of the border is from 102°54′ E to 104°53′ E, the latitude is from 30°05′ N to 31°26′ N, and the altitude is from 387 m to 5364 m. In the area, the plain, the hilly area and the mountain area accounts for 40.1%, 27.6% and 32.3%, respectively. The Chengdu city is located in the east of the Chengdu Plain known as “the land of abundance”, and the average elevation is 500 m.

The research region is the area within the loop expressway in Chengdu city, its area is 540.83 km^2. In the region, there are three important ring roads. The First Ring Road was completed in 1987, the Second Ring Road was well versed in 1993, and the Third Ring Road was finished in the end of 2002. The data was Landsat-7 ETM+ image obtained on July 30, 2006, the projection of which was UTM 48N, and the spheroid of which was WGS84. The spatial resolution of the multispectral bands, the thermal infrared band and the panchromatic wave band was 30 m, 60 m and 15 m respectively.

* ***Corresponding author*** e-mail: hzw@cdut.edu.cn

3 Data Processing

3.1 RADIATION CALIBRATION

Radiation calibration is a process of converting the digital value (DN) of remote sensing data into spectral radiance value of sensor. For Landsat-7 ETM+ data, the model is formula (1) [8, 9].

$$L_\lambda = \frac{L_{\max_\lambda} - L_{\min_\lambda}}{Q_{\max} - Q_{\max}} (Q_\lambda - Q_{\min}) + L_{\min_\lambda}, \qquad (1)$$

where L_λ is the spectral radiance by the sensor ($W\cdot m^{-2}\cdot sr^{-1}\cdot \mu m^{-1}$), Q_λ is the digital number of analyzed pixel, Q_{max} is the maximum digital number (255), Q_{min} is the minimum digital number, $L_{\max_\lambda}$ and $L_{\min_\lambda}$ are the maximum and minimum spectral radiance.

3.2 ESTIMATION OF LAND SURFACE EMISSIVITY

The NDVI threshold method [10] proposed by Beck et al. (1990) was adopted to estimate the land surface emissivity, and only the natural surface was considered. If NDVI<0.2, the land surface is considered to be the bare soil. If NDVI>0.5, the land surface is thought to be completely covered by vegetation. And if 0.1≤NDVI≤0.5, the land surface is deemed to be covered by vegetation and bare soil mixing. The emissivity of the bare soil and the total vegetation-covered area take the empirical value, which are 0.973 and 0.986 respectively. The bare soil and vegetation mixed coverage area is calculated by formula (2) [11].

$$\varepsilon = \varepsilon_v P_v R_v + \varepsilon_s (1 - P_v) R_s + d_\varepsilon,$$

$$\begin{aligned} R_v &= 0.92762 + 0.07033 P_v \\ R_s &= 0.99782 + 0.08362 P_v \end{aligned}, \qquad (2)$$

where, ε_v is the vegetation emissivity (0.986), ε_s is the bare soil emissivity (0.973). P_v is the vegetation proportion in pixel, and is estimated by formula (3). d_ε is the terrain factors, but it is negligible for the study area located in Chengdu Plain. $\rho$is the surface reflectance acquired in the red and near-infrared band.

$$P_v = \left[\frac{NDVI - NDVI_{\min}}{NDVI_{\max} - NDVI_{\min}} \right]^2$$

$$NDVI = \frac{\rho_{band4} - \rho_{band3}}{\rho_{band4} + \rho_{band3}}. \qquad (3)$$

Considering the NDVI threshold method only for natural surface, the pseudo-colour image of study area, spatial resolution of which was 15 m, was obtained by the fusion of the image, which was layered by red, green and near-infrared band selected, and the panchromatic wave band. Then, the land surface of study area was divided into three classes, which were the natural surface, the construction surface and the water by the supervised classification method. Thirdly, the NDVI image of study area was masked by the vector data of construction and water surface, the NDVI data of the natural surface was obtained, and so the emissivity of which was obtained by the NDVI threshold method. The emissivity of the construction and water surface was given the empirical value, 0.970 and 0.995 respectively. Finally, the land surface emissivity data of study area (see, Figure1) was generated by overlying the three kinds of images in space.

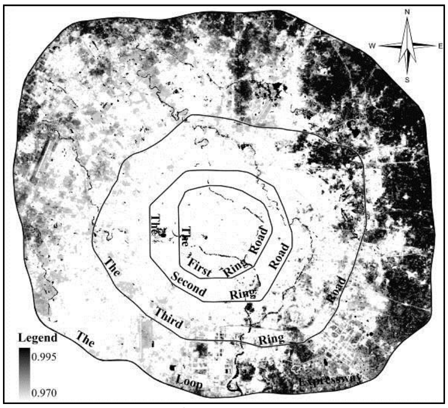


FIGURE 1 Land surface emissivity research region

4 Temperature Inversion

4.1. BRIGHTNESS TEMPERATURE INVERSION

Based on spectral radiation value of pixels on sensor, brightness temperature can be directly calculated by Planck's radiation function or an approximation formula (4) [9, 11, 13] The result is displayed in Figure.2.

$$T_{rad}=K_2/\ln(1+K_1/L_\lambda), \qquad (4)$$

where T_{rad} is brightness temperature of pixels and its unit is K, K_1 and K_2 are pre-launch calibration constants, K_1 is $666.093 W.m^{-2}.ster^{-1}.\mu m^{-1}$, and K_2 is 1282.708K [14].

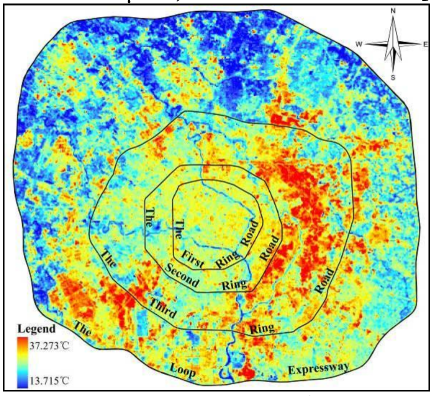


FIGURE 2. Brightness temperature of research region

4.2 LAND SURFACE TEMPERATURE INVERSION

The land surface temperature can be calculated according to formula (5) [9, 15]. And the result is displayed in Figure 3.

$$T_s = \frac{T_{rad}}{1+(\lambda \cdot T_{rad} / \rho)\ln\varepsilon} - 273.15\,, \tag{5}$$

where, T_s is LST and its unit is ºC, T_{rad} is the brightness temperature and its unit is K, λ is the center wavelength (11.4μm), $\rho = h \cdot c / \sigma$, where h is Planck constant $(6.626\times10^{-34}\, J\cdot s)$, c is the velocity of light $(2.998\times10^{8}\, m/s)$, σis Boltzmann constant $(1.38\times10^{-23}\, J/K)$, ε is the land surface emissivity.

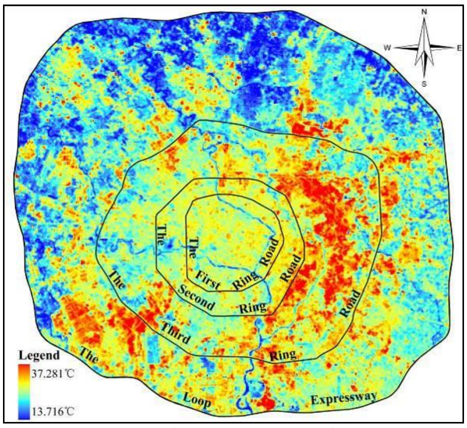


FIGURE 3 Land surface temperature of research region

5 Difference Analysis of UHI

5.1 LAND SURFACE TEMPERATURE INVERSION

According to the statistics from the BT image, the maximum value is 37.273ºC, the minimum value is 13.715ºC, the average value is 26.125ºC, the standard deviation is 1.600. And according to the statistics from the LST, the maximum value is 37.281ºC, the minimum value is 13.716ºC, the average value is 26.131ºC, the standard deviation is 1.601.

By comparison, the characteristic values of the LST are greater than the BT, and the value is from 0.001ºC to 0.008ºC, the average is 0.006ºC.

5.2 DIFFERENCE ANALYSIS OF TEMPERATURE CLASSIFICATION

Mean-standard deviation method [16] was used to divide temperatures of the research region into five grades, which are high temperature area, secondary high temperature area, medium temperature area, secondary low temperature area and low temperature area. High temperature area and secondary high temperature area are defined as the heat island area; others are called non-heat island area.

From the statistics of the BT classification image, the area of the five grades from high to low temperature respectively was 80.444 km^2, 85.997 km^2, 26.158 km^2, 9.472 km^2 and 78.759 km^2, and the UHI area was 166.441km^2, accounting for 30.78% of research region. And the statistics from LST classification image showed, the area of the five grades from high to low temperature respectively was 80.444 km^2, 85.997 km^2, 226.158 km^2, 68.124 km^2 and 80.107 km^2, and the UHI area was 166.441 km^2, accounting for 30.78% of research region.
From the comparison between the BT classification image and the LST classification image, there was a slight difference only on the area covered by the low temperature and the secondary low temperature, and the area covered by the others was equal. Therefore, there was no difference on the area of UHI defined by BT and LST.

5.3 DIFFERENCE ANALYSIS OF UHI INTENSITY

1) Comparison of the Difference of UHI and non-UHI Average Temperature

The method defines the UHI intensity is the difference between UHI average temperature and non-UHI average temperature. The math model is as the following:

$$P = T_{R,avg} - T_{V,avg}\,, \tag{6}$$

where P represents heat island intensity, TR, avg is average temperature in UHI area, TV, avg is average temperature in non-UHI area.

By calculation from the BT image, the average temperature of UHI and non-UHI images were 27.9202ºC and 25.3271ºC respectively, and the UHI intensity was 2.5931ºC. And statistics from the LST image showed that the average temperature of UHI and non-UHI images were 27.9273ºC and 25.3329ºC respectively, and the UHI intensity was 2.5944ºC. Apparently, the UHI intensity defined by LST was 0.0013ºC higher than what defined by BT.

2) Heat island area index method

This method used average temperature of non-UHI as basis, define the UHI intensity is the sum of products of the difference between the average temperature of every grade in UHI area and the basis, and its weight, which is the percentage of every grade in UHI area. The math model is formula (7).

$$P = (T_H, avg - T_V,_{avg}) \times A_H + (T_{SH,avg} - T_V,_{avg}) \times A_{SH}\,, \tag{7}$$

where P represents heat island intensity, TH, avg is the average temperature of high temperature area, TSH, avg is the average temperature of secondary high temperature area, TV, avg is the average temperature of non-UHI area, AH is the weight of high temperature area, ASH is the weight of secondary high temperature area.

The statistics from UHI image defined by BT showed that, the average temperature of high temperature area was 28.6376ºC, accounting for 14.87% in UHI area, the average temperature of secondary high temperature area was 27.2492ºC, accounting for 15.90%, the average temperature of non-UHI was 25.3271ºC, and the UHI intensity calculated by formula (7) was 0.7980ºC. And in UHI image defined by LST, he average temperature of high temperature area was 28.6447ºC, accounting for 14.87% in UHI area, he average temperature of secondary high temperature area was 27.2561ºC, accounting for 15.90%, the average temperature of non-UHI was 25.3329ºC, and the UHI intensity calculated was 0.7984ºC.

By comparison, the UHI intensity defined by LST was 0.0004ºC higher than what defined by BT.

6 Conclusions

Taking the Chengdu City as the research object, the brightness temperature (BT) and the land surface temperature (LST) were retrieved using the Remote Sensing technology, and the difference of the UHI of research region defined by them was researched. The conclusions are as follows:

(1) There were some differences between BT and LST, and the variation level of LST was higher than BT.

(2) There was a slight difference only on the area covered by the low temperature and the secondary low temperature, and the area covered by the others was equal. Therefore, there was no difference on the area of UHI defined by BT and LST.

(3) The UHI intensity defined by LST was slightly higher than that was defined by BT, and the intensity value was determined by the method used.

The above conclusions were based on the remote sensing image of specific phase and specific landscape. If the remote sensing image was different and the study area landform types was different, the results may be different, which need further study.

Acknowledgements

This work was supported in part was supported by the open foundation of the Laboratory of Geo-special Information Technology Ministry of Land and Resources, Chengdu University of Technology China (NO.KLGSIT2013-10), and by the key project of Sichuan Environmental Protection Bureau (NO.2010HBY003), and by the research project of Sichuan College of Architecture Technology (2011) and by Key Project of Education Department of Sichuan Province under Grant 12ZA255 and by The key science and technology project of the Deyang City (NO. 2013ZZ074-05).

References

[1] Howard L 1833 *Harvey and Darton* **1**(3) 1-24

[2] Matori A, Basith A, Harahap I S H 2012 *Arablan Journal of Geosciences* **5** 1069–84

[3] Liu C, Shi B, Shao Y X, Tang C S 2013 *Bulletin of Engineering Geology and the Environment* **72** 303-10

[4] Cheng F L, Di S, Jiang S D, Jing Y Y, Jun J Z, Dan X 2013 *Arabian Journal of Geoscience* **6**(8) 2829-42

[5] Marialuce S, Marco S 2012 *Computational Science and Its Applications* Part II LNCS **7334** 599–608

[6] Gantuya G, Ji Y H, Young H R, Young Y B 2013 *Journal of Atomosphere Science* **49**(4) 535-41

[7] Petr Dobrovolný 2013 *Theoretical and Applied Climatology* **112** 89-98

[8] Xu H Q 2007 *Geomatics and Information Science of Wuhan University* **32**(1) 62-7

[9] Zhang Y L, Bai Z K, Liu W B 2013 *Geo-Informatics in Resource Management & Sustainable Ecosystem* Part I CCIS398 264-73

[10] Becker F, Li Z-L 1990 Temperature independent spectral indices in thermal infrared bands *Remote Sensing Environment* **32** 17-23

[11] Qin Z H, Li W, Xu B, Chen Z, Liu J 2004 *Remote Sensing for Land and Resources* **3** 28–42

[12] Mustard J F, Camey M A, Sen A 1999 *Estuarine Coastal and Shelf Science* **49** 509-24

[13] Li C F, Yin J Y, Zhao J J. 2010. *International Journal of Environment Science Development* **3** 234-7

[14] Landsat Project Science Office 2010 *Landsat 7 science data user's handbook*

[15] Artis D A, Carnahan W H 1982 *Remote Sensing of Environment* **12**(4) 313-29

[16] Chen S L, Wang T X 2009 *Journal of Geo-Information Science* **11**(2) 145-7

Authors	
	Wenxia QIU, born on April 18, 1982, Linfen, Shanxi, China **Current position, grades:** Lecturer **University studies:** Application of Remote Sensing and GIS **Scientific interest:** Application of Remote Sensing and GIS **Publications:** Several articles (EI, Chinese core) **Experience:** Engaging in teaching and research work at Sichuan College of Architectural Technology, China
	Huixi XU, born in 1979, Deyang, Sichuan, China **Current position, grades:** Doctor (postdoctor), Associate Professor **University studies:** Application of Remote Sensing and GIS; Surveying and Mapping **Scientific interest:** Application of Remote Sensing and GIS; Surveying and Mapping **Publications:** Several articles (SCI, EI, Chinese core) **Experience:** Engaging in teaching and research work at Sichuan College of Architectural Technology, China
	Zhengwei HE, born in 1966, South County, Sichuan, China **Current position, grades:** Doctor(postdoctor), Professor, Doctoral tutor **University studies:** Remote sensing geology, Ecological Geographic Information System, Ecological environmental geology **Scientific interest:** Remote sensing geology, Ecological Geographic Information System, Ecological environmental geology. **Publications:** Several articles (SCI, EI, Chinese core), Several monographs (Chinese, English) **Experience:** Engaging in teaching and research work in Chengdu University of Technology, China

Multi-level dosing and preact self-adaption correcting automatic batch control model

Xuechao Liao[1, 2], **Zhenxing Liu**[3*]

[1] *College of Computer Science, Wuhan University of Science and Technology, Wuhan 430081, P. R. China*

[2] *Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430081, P. R. China*

[3] *College of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430081, P. R. China*

Received 1 May 2014, www.tsi.lv

Abstract

The process flow and system structure of automatic batch weighing system are presented. In order to increase production speed and dosing accuracy, the multi-level dosing control model (high/low speed dosing + inching dosing) is designed. Besides, the inching dosing mode is adopted to accurately compensate the weight deviation. In order to solve the problem that the fall of materials in-air cannot be easily controlled and out of tolerance. The multi-level dosing control model and preact will correct after each dosing dynamically with iteration method, moreover, the target value is predicted with second-order estimator, so as to increase the dosing speed with high weighing accuracy. The successful application proves that the control model can realize the rapid and accurate control of batch weighing process and has quite favourable control and reliability.

Keywords: Automatic batch, Multi-levels dosing, Fall of dosing, Self-adaption correcting, preact

1 Introduction

Automatic batch weighing system [1] is a very important procedure for the meticulous factory production technology. Velocity and accuracy of the batch weighing [2] is vital for the efficiency of entire production line and the product quality [3]. Automatic batch control process is a multiple inputs & outputs system. Various batch-conveying lines will be coordinated and controlled as per the formulating ratio set in advance [4]. Control system shall realize timely and accurate monitoring and regulation of material level and flow. This system will adopt two-level dosing control modes, respectively high/low speed and inching dosing. The two-level dosing control model and the preact will be corrected after each dosing dynamically with iteration method, moreover, the target value is predicted with second-order prediction, so as to realize the dynamic correction [5] of every dosing process and improve the dosing speed with high weighing accuracy.

2 Process Flow and Control Principles of the System

This system is mainly comprised of 8 large silos, 10 small silos, 8 liquid silos, 1 artificial dosing silo, 4 sets of weighing hoppers and weighing instruments (W1-W4), three sets of dosing vibrating screens (B1-B3), three sets of vibrating screen drivers (M1-M3), mixer (M5) and conveyor (M6) (as shown in Figure 1).

The system will firstly select the type of auxiliary materials to be dosed and calculate the material weight required for each silo according to the requirements for ingredients of feeds produced at present, and then come into operation after being confirmed by the operator. Large & small silos, artificial dosing silo and liquid silo shall set value as per the formulation of each silo. Start up the electromagnetic vibrating feeders (vibration velocity is in direct proportion to the dosing speed) as per the sequence set for dosing and feed solid materials into the hopper for weighing. When materials in the hopper reaching the set value, vibrating feeder will stop and the material feeding and weighing in the next silo will be done in sequence. After completing the weighing in all silos, solid materials in each hopper will be placed into the mixer through vibrating screens (M1, M2 and M4), for primary mixing. During that process, vibrating screen of the liquid silo (M3) will then feed liquid materials into the mixer for secondary mixing. Finally, the conveyor will move out the uniformly mixed feeds and the automatic batching process [6] completes.

* ***Corresponding author*** e-mail: Zhenxingliu@wust.edu.cn

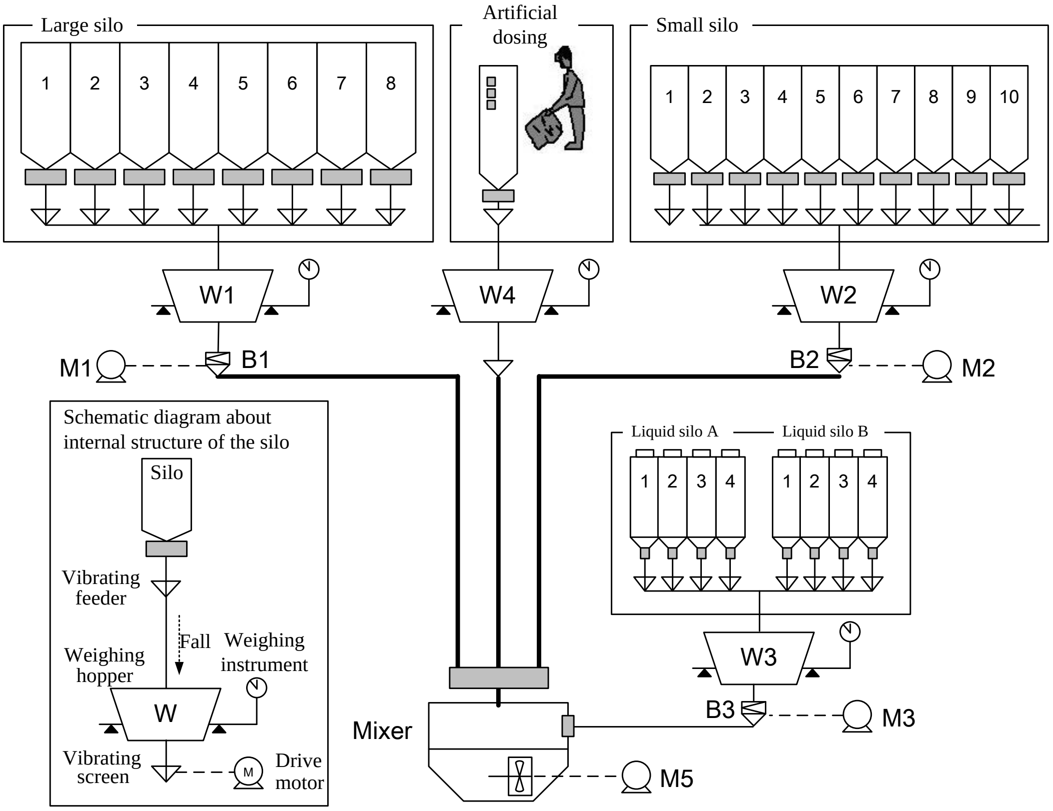


FIGURE 1 Control system process chart

Vibrating feeder is installed at the silo bottom. Vibrating feeder should be started at the beginning of the weighing process and materials will drop from silo to the weighing hopper. Since a certain distance exists between the silo bottom and the weighing hopper [7], materials will fall into the weighing hopper after a certain delay in the air after the vibrating feeder activates and only then weight of materials in the hopper will change. When vibrating feeder stops, all residual materials in the air will only fall into the hopper after a certain delay and then weight of materials in the hopper will be stable. Such out-of-tolerance caused by lagged materials is named as "fall of dosing" [8]. In addition, material level in the silo will fluctuate due to irregular shape of materials, relatively great difference in grain size and unscheduled dosing into silo during the production. Therefore, flow velocity of material changes randomly at every moment, to make the deviation of material weighing caused by "fall of dosing" change [9] every time. This weighing deviation caused by changes in material level and fall of dosing shall be provided with special batch weighing correction model. When weighing is going to be completed, estimate and calculate the preact as per the "fall of dosing" [10]; stop dosing when it reaches the preact; make use of the inertia before stop to realize:

"Current weighing value + Estimate preact = Target weighing value", and to control the error of dosing within the target ± threshold value.

3 Hardware Structure and Control Principles of the System

A three-level computer control network is comprised of the HMI, PLC, weighing instrument and frequency converter (as shown in Figure 2). PLC connects to the frequency converter through field bus and to the weighing instrument through RS232 serial. During the automatic batch production process, mix the main materials in the large silo, auxiliary materials in the small silo, artificial and liquid materials as per a certain proportion, and complete the measurement of materials by the weighing hopper. PLC mainly realize the real-time control of conveyor, weighing and mixing processes; complete the system fault detection, display and alarm; and meanwhile send signal to the frequency converter to regulate the speed of belt conveyor.

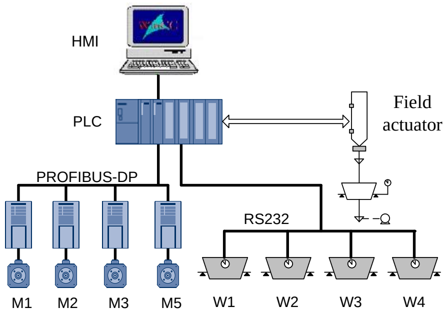


FIGURE 2 Control system hardware structure

To realize the rapid and high-efficiency weighing for the batching process, PLC will send a high-speed signal to the frequency converter when starts the weighing, to drive the vibrating feeder to run at high speed; when the weighing value is close to the target value, PLC will send a low-speed signal to the frequency converter, to drive the vibrating feeder to run at low speed, and; when the weighing value is equivalent to the preact value, PLC will send a stop signal to the frequency converter and inertia before stop will be used to make the weighing value be close to the target value. System control block diagram is shown in Figure 3.

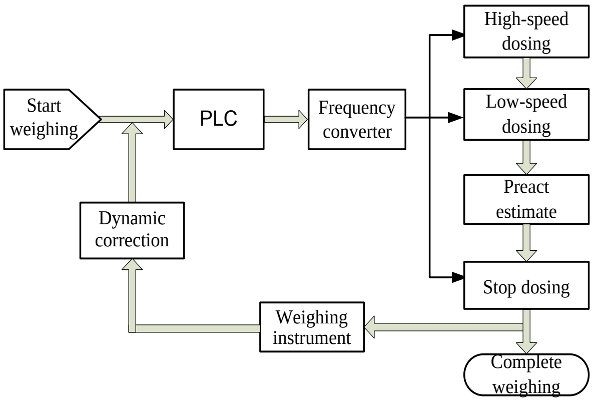


FIGURE 3 System control theory diagram

4 Multi-Level Automatic Batching Control Mode

Time of batch weighing and accuracy of weighing is a contradiction pair [11]. Higher weighing accuracy requires lower speed of dosing, which may lead to the extension of dosing duration. Therefore, that it is to guarantee the accuracy by taking more time, with low production efficiency. However, if the dosing speed is increased, fall of materials in the air can be difficultly controlled within a relatively short duration of weighing and that will easily lead to out-of-tolerance and further affect the weighing accuracy [12].

To realize the rapid and accurate control of the batch weighing process, this system adopts a two-level speed control mode for dosing. High speed (V2) will be adopted for rapid feeding at the beginning of the weighing process [13], and low speed (V1) will be adopted when the materials are close to the target value. Finally, make use of the control mode to calculate the fall in air and the pre-closing time, so as to guarantee the weighing accuracy. The following two factors will affect the performance of weighing for the abovementioned control modes, i.e. changeover between high and low speed and pre-estimate of every fall of dosing [14]. Since material level and shape etc. in the silos are different every time, predicted value of the preact shall be corrected according to the conditions of every dosing and the changeover point for high/ low speed dosing shall be modified as per the conditions such as dosing weight and preact etc.; otherwise cumulative error [15] will appear after several times of weighing. Therefore, the following two dosing control models are designed for control.

4.1 HIGH/ LOW SPEED DOSING CONTROL MODELS

High/ low speed dosing control models are shown in Figure 4 as below. Target of model control is to make sure that the actual weight is between r1 and r2 threshold values for dosing after high/ low speed dosing operation.

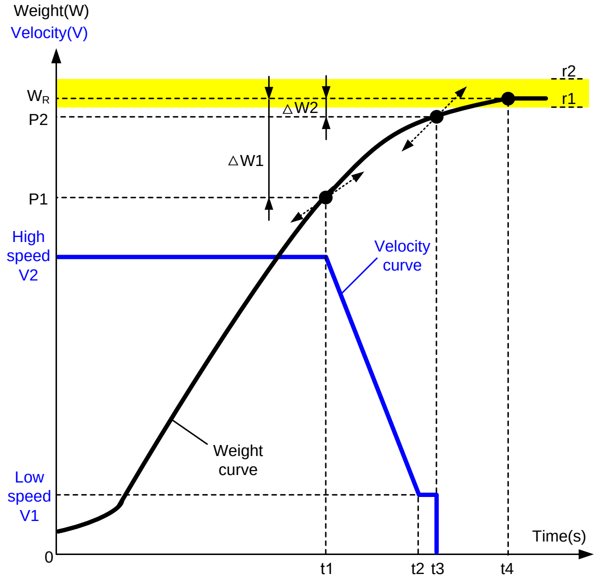


FIGURE 4 High/slow dosing control model

TABLE 1 The four stage parameter of high/low speed dosing control model

Time range	Phase	Weighing value at end point
0-t1	High speed dosing	P_1: weight at high/ low speed changeover point
t1-t2	High/ low speed changeover	
t2-t3	Low speed dosing	P_2: weight at stop point for dosing
t3-t4	Stopping & stabilizing	W_R: stabilized weighing weight

During the dosing process, firstly carry out high-speed vibrating dosing (V_2) before weighing value reaching P_1 and decrease the velocity from V_2 to V_1 by making use of linear method, of which the linear slope will be defined according to time (t2-t1). Then, carry out low-speed vibrating dosing (V_1) and stop dosing when the weighing value reaches P_2. Thereafter, weighing value on the hopper will be stabilized after a certain time due to the fall of dosing. The system will record the dosing stabilization time (t4) and the actual dosing value (W_R), to correct the preact. In addition, to eliminate the cumulative error caused by preact, the system will realize the real-time calculation of the weight difference for high/low speed changeover (preact 1) ΔW_1 and the weight difference between stop dosing – stabilizing (preact 2) ΔW_2 after completing dosing in each silo, including:

Then calculate the next high/ low speed changeover point (P_1) and dosing stop point (P_2) accordingly.

4.2 INCHING DOSING CONTROL MODEL

If high/low speed dosing mode is adopted but the actual weighing value of the silo still cannot reach the target threshold value, inching doing control model will be adopted (as shown in Figure 5).

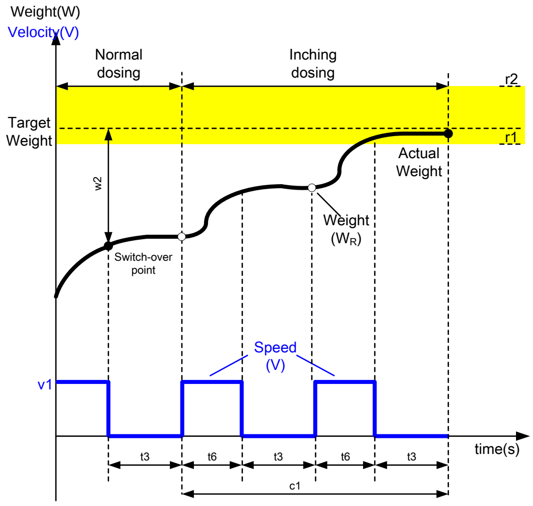


FIGURE 5 Inching dosing control model

Parameters in the figure:

- t6: inching impulse duration;
- t3: inching adjustment duration;
- r1, r2: threshold value for weight target (-)(+);
- w2: weight difference between stop dosing and stabilizing;
- v1: low dosing speed;
- c1: max value of inching dosing number (generally c1=3)

Inching dosing control process includes the following three steps:

1) Start low-speed dosing (v1), time duration t6;
2) Stop low-speed dosing, time duration t3;
3) Calculate the weighing value (W_R); repeat the above inching dosing process if W_R<r1 and the number of inching repeated <= c1.

5 Preact self-adaption correction model

5.1 SELF-ADAPTION CORRECTION OF MODEL

To realize the accurate control of the dosing process and eliminate the cumulative error caused by preact, the system will realize the real-time analysis on the change in deviation during every weighing process and carry out dynamic correction for P1 and P2 after every dosing.

At the time of Nth dosing:

Weighing error:

$$\Delta E(N) = W_R(N) - W_T(N) . \quad (1)$$

Target value:

$$W_T(N) = W_P - \Delta E(N-1) . \quad (2)$$

Error rate:

$$\Delta U(N-1) = \frac{\Delta W_2(N-1)}{\Delta W_T(N-1)} . \quad (3)$$

Preact 1:

$$\Delta W_1(N) = \Delta W_1(N-1) - \Delta U(N-1) . \quad (4)$$

Preact 2:

$$\Delta W_2(N) = \Delta W_2(N-1) - \Delta U(N-1) . \quad (5)$$

Formula (3) is the definition of this weighing error; Formula (4) is the definition of this target value; Formulas (5) & (6) are definition of this preact.

If the last error rate (ΔU(N-1)) is much greater, preact set for this time shall be decreased. $P_1(N)$ and $P_2(N)$ shall be calculated as per Formulas (5) and (6).

$$P_1(N) = W_R(N) + \Delta W_1(N) , \quad (6)$$

$$P_2(N) = W_R(N) + \Delta W_2(N) . \quad (7)$$

The abovementioned method is realized through iteration in PLC and has relatively favourable control effect when the target value is quite great (W_P>30Kg).

If W_P is quite small, ΔE(N-1) in Formula (1) will be relatively low, error rate ΔU(N-1) is unobvious, and actual value $W_R(N)$ is quite close to target value $W_T(N)$. However, cumulative error after several times of dosing will have great fluctuation. To make $W_R(N)$= $W_T(N)$, the following iteration method will be used to carry out dynamic correction for Δ$W_2(N)$ (correction method of Δ$W_1(N)$ is similar to this method):

0th time:

$$\Delta W_2(N) = 0 . \quad (8)$$

1st time:

$$\begin{aligned} &P_2(1) = W_P - \Delta W_2(0) \\ &\Delta W_2(1) = [\Delta W_2(0) + (W_R(1) - P_2(1))]/2 \end{aligned} . \quad (9)$$

Nth time:

$$\begin{aligned} &P_2(N) = W_P - \Delta W_2(N-1) \\ &\Delta W_2(N) = \frac{[\Delta W_2(N-1) + (W_R(N) - P_2(N))]}{2} \end{aligned} , \quad (10)$$

In compliance with the abovementioned iteration method, make use of this preact Δ$W_2(N)$ to recursively calculate the preact (Δ$W_2(N+1)$) of next batch after completing every dosing. Accurate estimate of weighing preact can be realized after several batches of dosing weighing, so as to improve the weighing accuracy.

5.2 DESIGN OF SECOND-ORDER ESTIMATOR FOR TARGET VALUE

Controlled object of this dosing control model is meticulous feed batching system, characterized by low silo weighing value, fast dosing and short duration for stop stabilization. To improve the adaptation of model calculation to the meticulous control objects, the system design adopts the following second-order estimator model to predict the actual weighing value (*W_R*(*N*)):

$$W_R(N+1) = a_1W_R(N)+a_2W_R(N-1) \\ +b_0W_T(N)+b_1W_T(N-1)+b_2W_T(N-2) , \quad (11)$$

Assign: $\omega(N)=[a_1,a_2,b_1,b_2]$, $X^T(N)=[W_R(N),W_R(N-1),W_T(N-1),W_T(N-2)]$.

Then,

$$W_R(N+1)=b_0W_T(N)+\omega(N)\cdot X^T . \quad (12)$$

In Formula (12), "$W_R(N+1)$" refers to the (N+1)th actual weighing value; "$W_T(N)$" the set value for the Nth weighing; "$\omega(N)$" the parameter vector; and "X^T" the transposition of data vector. Second-order estimator of this system is designed for the purpose of selecting appropriate $W_T(N)$, to minimize the index for the next weighing error.

$$J=[W_R(N+1)-W_P]^2 . \quad (13)$$

The following control equation is obtained according to the self-correcting principle:

$$W_T(N)=\frac{W_P-\omega(N)\cdot X^T(N)}{b_0} . \quad (14)$$

Vector $\omega(N)$ shall be recursively estimated by making use of the generalized least squares; and parameter b_0 can be selected at the time of initial calculation. Recursive estimation equation for parameter vector $\omega(N)$ is shown as below:

$$\hat{\omega}(N+1)=\hat{\omega}(N)+K(N+1)\cdot\delta(N) , \quad (15)$$

where "$K(N+1)$" refers to the gain matrix and "$\delta(N)$" is the correction term:

$$\delta(N)=[W_R(n+N+1)-X^T(N+1)\hat{\omega}(N)] . \quad (16)$$

Second-order estimation for system design, n=2

Formula (15) defines that: this estimate value of parameter ($\hat{\omega}(N+1)$) is to add a correction term ($\delta(N)$) based on the last estimate value ($\hat{\omega}(N)$) . The abovementioned second-order estimator has the following advantages: this estimate value is not only related to the last parameter but also the historical data of the previous $(n+N+1)^{th}$ estimate value and the actual weighing value. Therefore, it reflects the influence of the entire fluctuation process of near-term weighing values on the preact of estimate value. In addition, initial value for recursion should be roughly selected during the recursive process of the generalized least squares, then a brand new group of estimate values will be generated through iterative calculation for (2n+1) times, and continue the recursion based on that.

6 Conclusion

This Paper applies the high/ low speed dosing control model to realize the accurate control of batch weighing, i.e. make the material weight be close to the target value as practical as possible during the high-speed dosing process and realize the accurate control of batch weighing during the low-speed dosing process; in addition, adopt the inching dosing method to compensate the weighing accuracy of low-speed dosing. Upon completion of each dosing, carry out dynamic correction for the next preact ($\Delta W_1(N)$ and $\Delta W_2(N)$), and make use of the second-order estimator to predict the target value ($W_T(N)$). This control calculation method has been used in a certain meticulous batch weighing system. Table 2 shows the process data for real-time control (set value for technology 40Kg) of a silo. Before adopting new calculation method, it will take about 5 minutes and 20 seconds for completing one weighing and the accuracy error will reach ±0.12Kg after several continuous weighing. With the increase in weighing times, cumulative error will be greater and greater. After applying the new calculation method, completion of one weighing of three materials will be controlled within 4 minutes and 30 seconds.

TABLE 2 The process data of continuous 20 batches

No	Actual value $W_R(N)$	Dosing stop point $P_2(N)$	Preact $\Delta W_2(N)$	Error $\Delta E(N)$
0	40.0000		0.0000	0.0000
1	38.2520	40.0000	-0.8740	-1.7480
3	40.6780	39.2625	0.3390	-2.5450
5	40.1530	40.6410	0.0765	-1.1100
7	40.5580	40.0420	0.2790	-0.4680
9	39.9250	40.3875	-0.0375	0.2320
11	39.8250	40.0290	-0.0875	0.1150
13	39.8060	39.9980	-0.0970	-0.0830
15	40.0110	40.0020	0.0055	-0.0680
17	40.0650	39.9990	0.0325	-0.0050
19	39.9990	40.0085	-0.0005	0.0110
20	40.0020	39.9995	0.0010	0.0130

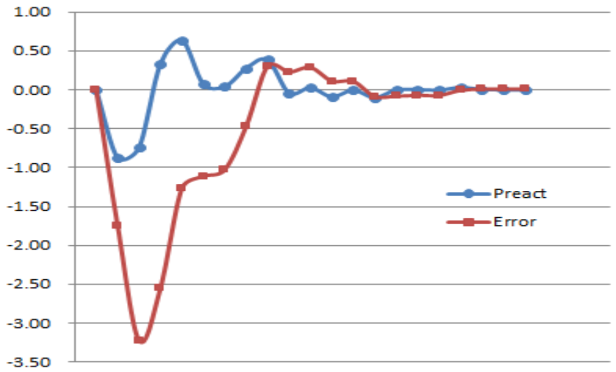


FIGURE 6 the Trend graph of Error and preact of continuous 20 batches

According to the variation trend of error and preact as shown in Figure 6, the weighing accuracy after 20 times of continuous weighing can be controlled approximately ±0.01Kg and the increase in weighing times will not affect the cumulative error and preact. Therefore, the new calculation method can effectively mitigate the contradiction between weighing velocity and accuracy and furthermore fully satisfies the production need of the entire technology. In addition, this control model can, characterized by reasonable design and simple operation, realize the rapid and accurate control of batch weighing process and has quite favourable control and reliability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61174107 and 61100055), Natural Science Foundation of Hubei Province (Grant No. 2011CDB233)

References

[1] Song Yuepeng 2008 Research on Automatic Batching System Based on PLC with Self-Correcting Fuzzy Control *Electrical Drive* **38**(8) 72-4

[2] Song Hui, Fang Zongda 2003 Measures about Weighing of Automatic Batch Control System *Automation and Instrument & Meters* **2** 23-4

[3] Status and Ralph 2002 Improved Temperature Control in Batch Production Systems *ISA TECH/EXPO Technology Update Conference Proceedings* **422** 109-15

[4] Hogenson and David C 1990 Four Approaches to Batch Control *Advances in Instrumentation Proceedings* **45**(1) 133-6

[5] Labs and Wayne 1993 Control Software Features Batch Management and PLC-Based PID *Instruments & Control Systems* **66**(4) 102-11

[6] Jones and Wardin 1991 Design of a PLC Based Control System for a Batch Reactor *Proceedings of the IASTED International Symposium oil Circuits and Systems* 902-11

[7] Skontos and Sam 1991 PLCs Challenge DCSs in Batch Control *Process and Control Engineering* **44**(4) 48-50

[8] Bonnina and Anthony T 1990 Workstations and PLCs for Batch Control *Instruments & Control Systems* **63**(11) 53

[9] Liu Chao, Bai Ling, Liu Feng 2010 Error Compensating Model of Smoothness Self-Adaptation of Feedstuff Mixing and Application *Feed Industry* **31**(21) 1-3

[10] Jeffery R 1998 Integrating Scales and Weight Information and PLCs *Australian Journal of Instrumentation and Control* **13**(3) 10-16

[11] Troys and Dougalas 1996 Development Environment for Batch Process Control *Computers in Industry* **31**(1) 6l-84

[12] Zhang Qingbin Bi Lihong, Wang Zhu 2005 Analysis on Accuracy of Industrial Automatic Batching System *Automation Technology and Application* **24**(5) 79-81

[13] Wishowski and Kirk E 1994 Automated Control of Batch Mix Tank is Wed to Manual Procedures and Operator Intervention *Instrumentation & Control Systems* **67**(7) 59-65

[14] South and Crai 1995 Continual Batch Processing for Liquid Wastes. *Process and Control Engineering* **48**(6) 26-8

[15] Sumi and Takao 1990 Supervisory System for Batch Process *Advances in Instrumentation Proceedings* **45**(3) 1115-22

Authors

Liao Xuechao, born in 1979, Hubei Province, China

Current position, grades: lecturer
University studies: Wuhan University of Science and Technology
Scientific interest: industry process control, electric motor and motor fault diagnosis
Experience: He received the Master degree from Wuhan University of Science and Technology in 2006. Currently, he is an lecturer of the College of Computer Science and Technology at Wuhan University of Science and Technology.

Liu Zhenxing, born 1965, HuNan Province, China

Current position, grades: Professor, Doctoral tutor
University studies: Wuhan University of Science and Technology
Scientific interest: industry process control, electric motor and motor fault diagnosis
Experience: He received the PhD degree from Huazhong University of Science and Technology in 2004. Currently, he is an Profess of the College of of Information Science and Engineering at Wuhan University of Science and Technology.

Experimental research on transmission efficiency of metal belt continuously variable transmission

Wu Zhang[1]*, Wei Guo[1], Chuanwei Zhang[1], Yizhi Yang[2], Yu Zhang[3]

[1] *School of Mechanical Engineering, Xi'an University of Science and Technology, Yanta Str. 58, 710054, Xi'an, China*

[2] *College of Humanities and Foreign Languages, Xi'an University of Science and Technology, Yanta Str. 58, 710054, Xi'an, China*

[3] *The 41st Research Institute, The 6th Academy of China Aerospace Science And Industry Corporation, 1055 mailbox, 010010, Hohhot, China*

Received 5 May 2014, www.tsi.lv

Abstract

Transmission efficiency is one of the main limiting factors on metal belt CVT large-scale assembly car. Metal belt CVT transmission efficiency has been invested in this paper, and, test-bed has been established by L13A3 engine, MB-CVT, brake, input sensor, output sensor, coupling and half shaft. Efficiency test results show that, with the decrease of transmission ratio, CVT efficiency first increases and then decreases. The range of efficiency is nearly 45%-89% in increases part (i>1), the range of efficiency is nearly 85%-89% in decrease part (i<1), the efficiency reaches the highest when transmission ratio is 1. The conclusions are in consistent with others conclusion, whereby demonstrating that the established transmission efficiency test-bed is rational and that the experiment results are reliable.

Keywords: metal belt, CVT, pulley, strain

1 Introduction

The transmission efficiency and cost of metal belt continuously variable transmission (MB-CVT) restraint its wide use. Figures 1 and 2 show the basic structure of metal belt CVT. Both driver and driven pulley contain a moving pulley and a fixed pulley; a metal block contact with driver and driven pulleys, thus torque is transmitted by friction between them. An infinite number of gear ratios have been achieved by pressure regulating device of moving pulley.

Yang YL established MB-CVT efficiency test-bed by inverter Motor, and the efficiency of the CVT was related to transmission ratio, oil pressure and input rotary speed [1]. Three control schemes have been proposed by Deng Tao, the fuel economy can be raised by around 2.9% to 3.5% respectively while the power performance intact has been kept [2]. The transmission efficiency has been improved by means of optimized control strategy by Xue DL [3]. Micklem J D [4] proposed a friction model based on elastohydrodynamic theory. Kim P [5] established an independent pressure-control-type on reduction in pressure fluctuations. Carbone G [6] concerned with the shifting behaviour of a MB-CVT, and the pulley elastic deformations has been described in this paper. Nilabh Srivastava [7-8] focused on a detail transient dynamic model to understand the transient behaviour of MB-CVT and evaluate the system performance under the influence of pulley flexibility and varying friction characteristics of the belt-pulley contact zone, his other paper [9] extensively discussed the concepts, mathematical and computational model in MB-CVT and Chain Belt CVT. Kong LY [10] studied the interaction between the individual bands and between the innermost band and pulley surfaces. Guebeii M [11] shows that, the CVT efficiency reach maximum when transmission ratio is 1. In a previous work [12] Sun Dezhi analysed the efficiency loss of MB-CVT in different torque ratio. This paper showed the friction power loss between metal belt and pulley is the main power loss. Liao Jian [13] reported the influence of power loss by lubricant viscosity and the calculation equation of efficiency has been obtained. Zhang wu [14] and Narita K [15] analysed the friction characteristics between metal belt assembly and pulley. Akehurst et al. [16-18] analysed the loss mechanisms between metal belt and pulley by pulley deflection. Kobayashi et al. [19] analysed the slip behaviour between the metal blocks, but they did not analyse the transmission efficiency of metal belt CVT under a realistic running condition. Zhang wu [20] focused on the pulley deformation by elastic theory, described the relationship between pulley radius, transmission ratio and pulley deformation.

The research reported in this paper focused on the MB-CVT transmission efficiency. The goal is to understand the transmission efficiency under the influence of transmission ratio. The MB-CVT efficiency test-bed by the engine of HONDA L13A3 has been established. Transmission efficiency variation law of MB-CVT on drive and reserve has been tested.

* ***Corresponding author*** e-mail: zhangwu0828@163.com

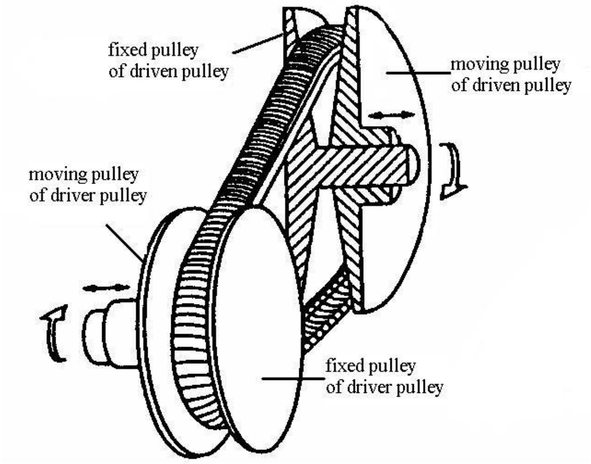

fixed pulley of driven pulley
moving pulley of driven pulley
moving pulley of driver pulley
fixed pulley of driver pulley

FIGURE 1 Basic structure of MB-CVT

FIGURE 2 Metal belt

2 Calculation principle of efficiency

Zhang Wu [14] shows that the CVT power loss was determined as four parts: the radial and tangential friction power loss between segment side and pulley surface, friction power loss between segment shoulder and innermost metal ring, relative slipping power loss between metal rings. Although the several power loss has been obtained alone in theory, but it is difficulty in experiment. Therefore, the test of MB-CVT efficiency has been analysed mostly from an overall perspective.

There is power loss of coupling, bearing and universal joint in experiment, thus the disturber of them should be eliminated. The MB-CVT efficiency can be determined by the following equation

$$\eta_z = \frac{T_2 n_2}{T_1 n_1 \cdot \eta}, \quad (1)$$

where T_1 and n_1 are input torque and speed of torque speed sensor No.1, respectively; T_2 and n_2 are output torque and speed of torque speed sensor No.2, respectively.

Footnote: “z” is total, “1” is input, “2” is output.

Drive,

$$\begin{aligned}\eta &= (2\eta_{zc}) \cdot \eta_{lzq} \cdot \eta_{wxj} \cdot \eta_{cl} \\ &= (2\eta_{zc}) \cdot \eta_{lzq} \cdot \eta_{wxj} \cdot \eta_{cl} \\ &= (2 \times 0.99) \times 0.97 \times 0.98 \times 0.99 \\ &\approx 0.922\end{aligned}$$

Reverse,

$$\begin{aligned}\eta &= (2\eta_{zc}) \cdot \eta_{lzq} \cdot \eta_{wxj} \cdot \eta_{cl} \cdot \eta_{xxjg} \\ &= (2\eta_{zc}) \cdot \eta_{lzq} \cdot \eta_{wxj} \cdot \eta_{cl} \cdot \eta_{xxjg} \\ &= (2 \times 0.99) \times 0.97 \times 0.98 \times 0.99 \times 0.95 \\ &\approx 0.876\end{aligned}$$

Where, η_{zc}=0.99, η_{lzq}=0.97, η_{wxj}=0.98, η_{cl}=0.99, η_{xxjg}=0.95. These data was obtained by mechanical design handbook.

Footnote: “zc” is bearing, “lzq” is coupling, “wxj” is universal joint, “cl” is gear, “xxjg” is planetary gear selector mechanism.

3 MB-CVT efficiency experiment

3.1 EXPERIMENT SYSTEM AND SCHEME

The experiment system involved:

1. Engine: HONDA L13A3-2517413, compression ratio: 10.4, the maximum power: 60/5700(kW/rpm), the maximum torque: 116/2800(Nm/rpm);
2. CVT: BOSCH SERA-PWRS5J15-0257, the range of theoretical transmission ratio: *i*=2.367-0.407.
3. Two torque speed sensor: torque capacity: 200(Nm) and 500(Nm), range of speed: 0-5000r/min;
4. Magnetic: torque capacity: 630(Nm);
5. Other affiliated equipment’s: efficiency test system, electronic control model, power control module, brake control part and CVT lubricant, et al. MB-CVT efficiency test-bed schematic diagram as shown in Figure 3.

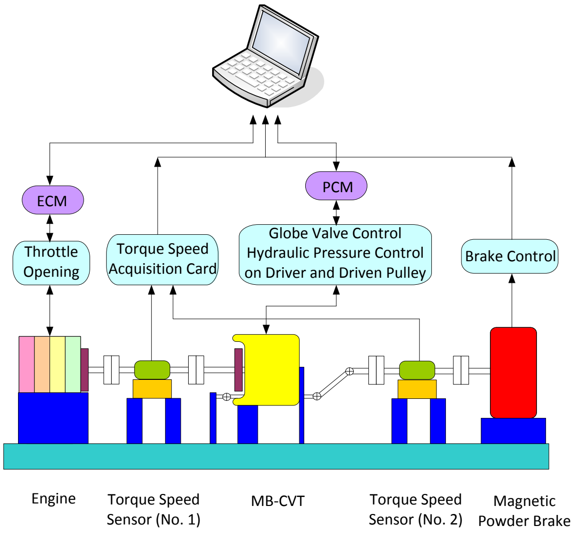

ECM
PCM
Throttle Opening
Torque Speed Acquisition Card
Globe Valve Control Hydraulic Pressure Control on Driver and Driven Pulley
Brake Control
Engine
Torque Speed Sensor (No. 1)
MB-CVT
Torque Speed Sensor (No. 2)
Magnetic Powder Brake

FIGURE 3 MB-CVT efficiency test-bed schematic diagram

The experiment system includes the final drive and the differential. There are two half axles in the experiment, one has been fixed on the test-bed and the other as an output shaft. Final gear transmission ratio is i_e=6.02, the rate (transmission ratio) of two half axes as an output shaft and one as an output shaft is i_b=0.5; the range of transmission ratio between the driver and driven pulley in theory is i=2.367-0.407. The drive transmission ratio is different from the reverse. The torque has been transmitted to the driver pulley by forward clutch in drive, so the transmission ratio is 1. Planet carrier has been fixed by friction plate brake in reverse. The torque has been transmitted to the driver pulley by sun gear, planetary gears and ring gear, the rotational is in the opposite direction to the sun gear, and transmission ratio is i_d=1.7857. The range of the total transmission ratio is i_D=7.125-1.225 in drive and i_R=12.723-2.188 in reserve. MB-CVT transmission ratio general diagram is shown in Figure 4. MB-CVT transmission efficiency test-bed photo is shown in Figure 5. HONDA interface module is shown in Figure 6. Software screen capture is shown in Figure 7.

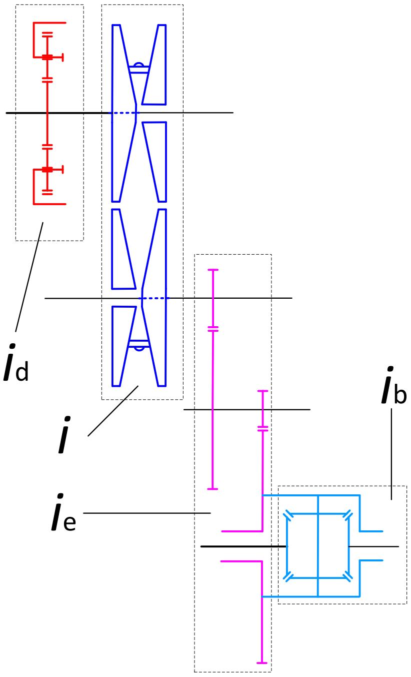


FIGURE 4 MB-CVT transmission ratio general diagram

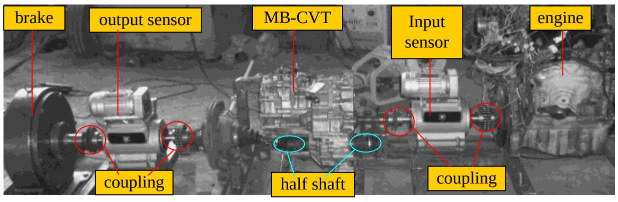


FIGURE 5 MB-CVT transmission efficiency test-bed

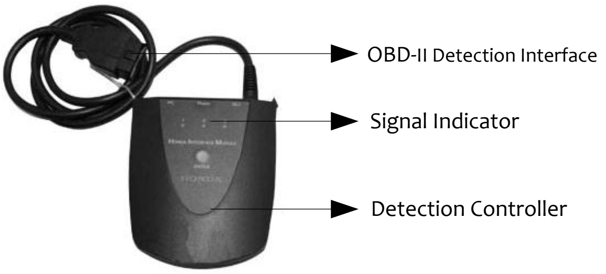


FIGURE 6 HONDA diagnostic system (HDS)

Name: HONDA interface module

Manufacturer: HONDA

Specification: Honda interface module is the newest detecting instrument of Honda. It detect the systems includes: powertrain, body, chassis, ABS, SRS, CVT, etc, and able to reprogram the vehicle control module.

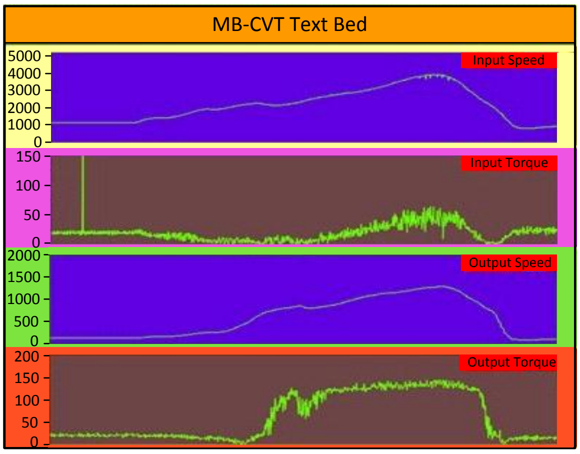


FIGURE 7 Software screen capture

3.2 EXPERIMENT PROCESS

Figure 8 shows the experiment schematic and steps.

1. Check whether the CVT is shifted to neutral. observe whether the pilot lamp of magnetic powder brake is block out, and start the engine;
2. Shift the CVT gear to drive (reserve) when the water temperature of engine reached to about 80°C
3. Connect the power supply of magnetic powder brake, adjust voltage, observes the output torque of software interface; Stop voltage's adjustment and measure the CVT lubricant temperature when the output torque is presetting.
4. Adjust throttle opening slowly from small to big and stop when the speed of engine is higher, and the sound and vibration is bigger, and then ease off the throttle opening quickly. Observe the change law of T_1, n_1, T_2 and n_2 in this process. The whole data have been registered by computer into the text form. The frequency of collection data is 100 Hz (the frequency decided by software and can be changed).
5. Measure the CVT lubricant temperature.
6. Adjust the load torque to the next data, repeat the above steps from the step 3.

7. Shift the CVT gear to Reverse, instead of step 2, and repeat the above steps.
8. Cut off the power supply, and close the engine.

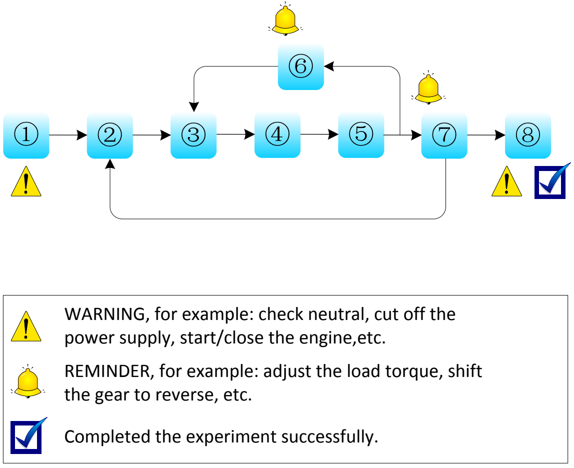

①
②
③
④
⑤
⑥
⑦
⑧
WARNING, for example: check neutral, cut off the power supply, start/close the engine,etc.
REMINDER, for example: adjust the load torque, shift the gear to reverse, etc.
Completed the experiment successfully.

FIGURE 8 Experiment process schematic diagram

3.3 EXPERIMENT DATA

The change of magnetic powder brake's torque has been controlled within a range in experiment, meanwhile adjust throttle's opening range to observe input torque, input speed, output torque and output speed.

The slipping occurred between the pulley and metal belt, the slipping rate of 10% [13] should be considered during the transmission ratio calculation between the two pulleys. The transmission ratio under realistic running condition is *i*′.

$$i' = (1-10\%)n_1 / (n_2 i_e i_b i_d) \,, \qquad (2)$$

where i_d=1(Drive); i_d=1.7857(Reverse).

Tables 1 and 2 show the experiment data on drive and reverse respectively.

TABLE 1 MB-CVT efficiency experiment data (drive)

Serial number	Input speed n_1(r/min)	Input torque t_1(n·m)	Output speed n_2(r/min)	Output torque t_2(n·m)	Realistic transmission ratio i'	Efficiency η_z
1	911	38	116	123	2.348	0.447
2	1103	38	146	124	2.259	0.468
3	1339	38	184	121	2.176	0.475
4	1613	37	230	121	2.097	0.506
5	1619	38	256	118	1.891	0.533
6	1635	37	290	118	1.686	0.614
7	1762	40	396	123	1.330	0.750
8	1936	42	502	120	1.153	0.804
9	2213	42	631	118	1.049	0.869
10	2567	43	764	118	1.005	0.886
11	2916	45	901	119	0.968	0.886
12	3359	46	1086	115	0.925	0.877
13	3651	45	1237	107	0.883	0.874
14	3729	48	1359	105	0.820	0.865
15	3896	48	1430	104	0.815	0.863

CVT lubricant: HONDA ATF-Z1

TABLE 2 MB-CVT efficiency experiment data (reverse)

Serial number	Input speed n_1(r/min)	Input torque t_1(n·m)	Output speed n_2(r/min)	Output torque t_2(n·m)	Realistic transmission ratio i'	Efficiency H_z
1	956	35	69	200	2.320	0.471
2	1023	35	81	205	2.115	0.529
3	1249	36	107	203	1.955	0.551
4	1325	36	121	197	1.834	0.570
5	1355	35	130	188	1.745	0.588
6	1367	37	138	193	1.659	0.601
7	1408	36	149	200	1.582	0.671
8	1474	35	166	190	1.487	0.698
9	1503	36	176	189	1.430	0.702
10	1638	37	216	196	1.270	0.797
11	1831	38	273	195	1.123	0.873
12	2101	43	369	185	0.953	0.863
13	2213	45	411	182	0.902	0.857
14	2453	46	464	181	0.885	0.850
15	2632	47	508	181	0.868	0.849

CVT lubricant: HONDA ATF- Z1

3.4 ANALYSIS AND DISCUSSION OF EXPERIMENT RESULTS

The engine working condition is too easy to be influenced by intake air temperature, intake air pressure and injection quantity, thus the engine output characteristic is worse than motor. The sound and vibration is bigger when the speed of engine is higher because of the installation error and test-bed torsion stiffness, thus the maximum engine speed is about 4000r/min in experiment. The experiment has been done in two modes (drive and reserve).

The efficiency is lower when the load torque is smaller, because the energy loss of mechanism and heat. MB-CVT transmission efficiency in different load and mode is shown in Figure 9.

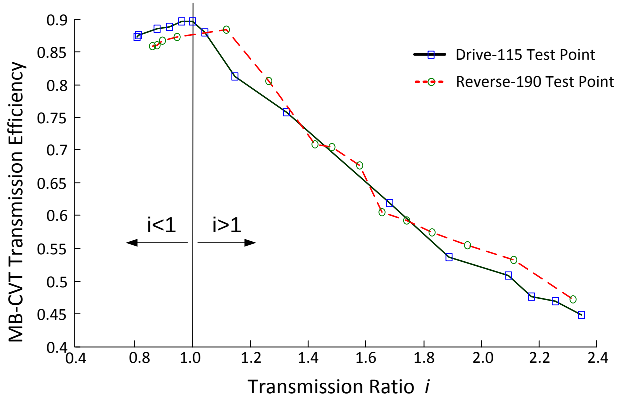


FIGURE 9 MB-CVT transmission efficiency

As indicated in Figure 9:

Transmission ratio: the range of transmission ratio is 0.8-2.367 both on drive and on reverse. The range of transmission ratio *i*=0.407-0.8 is too difficult to obtain for engine, because the speed of engine is limited. Increasing the throttle opening can overcome load when the load torque is small, so the CVT is easier to achieve a growth rate of movement, namely the transmission ratio *i*<1.

Efficiency: experiment data collected while the engine is running at a transient operation. The experiment results show that, with the decrease of transmission ratio, the CVT efficiency increases first, and then decreases. The range of efficiency is nearly 45%-89% in increases part(i>1), the range of efficiency is nearly 85%-89% in decrease part(i<1), the efficiency is the maximum when transmission ratio is 1. It can be described that the efficiency is the minimum in start-up phase, increasing in acceleration phase. Normally, CVT work in high efficiency area, namely transmission ratio *i*<1.

4 Conclusion

This paper established the MB-CVT transmission efficiency test-bed by engine. The range of transmission ratio is 0.8-2.367 in experiment. The range of transmission ratio i=0.407-0.8 is too difficult to obtain for engine, because the speed of engine is limited. With the decrease of transmission ratio, the CVT efficiency increases first, and then decreases. The range of efficiency is nearly 45%-89% in increases part(i>1), the range of efficiency is nearly 85%-89% in decrease part(i<1), the efficiency is the maximum when transmission ratio is 1. Normally, CVT work in high efficiency area, namely transmission ratio i<1. The results from experiment are in agreement with those others, whereby demonstrating that the established test-bed is rational and that the analyses are reliable.

Appendices

η_{z}	MB-CVT efficiency
η_{zc}	transmission efficiency of bearing
η_{lzq}	transmission efficiency of coupling
η_{wxj}	transmission efficiency of universal joint
η_{cl}	transmission efficiency of gear
η_{xxjg}	transmission efficiency of planetary gear selector mechanism
i_e	Transmission ratio of final gear(the value is 6.02)
i_b	the rate of between two half axles as a output and one as a output shaft (the value is 0.5)
i	The range of transmission ratio between the driver and driven pulley in theory (the value is 2.367-0.407)
i_d	transmission ratio of planet gear system (the value is 1.7857 in reverse, the value is 1 in drive)
i_D	The range of the total transmission ratio in drive (the value is 7.125-1.225)
i_R	The range of the total transmission ratio in reserve (the value is 12.723-2.188)
i'	realistic transmission ratio
n_1	input speed
n_2	output speed
T_1	input torque
T_2	output torque
MB-CVT	metal belt continuously variable transmission
ECM	engine control module
PCM	powertrain control module
transmission ratio	the pulley speed ratio
D	drive
R	reserve

Acknowledgments

This research was supported by the Doctor Fund of Xi’an University of Science and Technology (2012QDJ027) and Cultivation Fund of Xi’an University of Science and Technology (201330).

References

[1] Yang Y L, Qin D T, Sun D Y, Yang W, Li P J, Hu M, Zhang J S, Yang Z B 2002 Experimental investigation into the performance of metal-belt type continuously variable transmission *Chinese Journal of Mechanical Engineering* **5** 38 *(In Chinese)*

[2] Deng Tao, Sun D Y, Qin D T, Luo Y 2012 Simulation and test of integrated control for continuously variable transmission system *Automotive Engineering* **1** 32

[3] Xue D L, Yang Kai, Cheng J J, Wu J 2012 Analysis and study on the clamping force of metal V-belt CVT *Automotive Engineering* **10** 34

[4] Micklem J D, Longmore D K, Burrows C R 1994 Modelling of the steel pushing V-belt continuously variable transmission *Proc IMechE, Part C: J. Mechanical Engineering Science* **1** 208

[5] Kim P, Ryu W S, Kim H, Hwang, S-H, Kim, H-S 2008 A study on the reduction in pressure fluctuations for an independent pressure-control-type continuously variable transmission *Proc IMechE, Part D: J. Automobile Engineering* **5** 222

[6] Carbone G, Mangialardi L, Mantriota G 2005 The influence of pulley deformations on the shifting mechanism of metal belt CVT *Journal of Mechanical Design* **1** 127

[7] Srivastava N, Haque I 2007 Transient dynamics of the metal V-belt CVT: Effects of pulley flexibility and friction characteristics *Journal of Computational and Nonlinear Dynamics* **1** 2

[8] Srivastava N, Haque I 2005 On the transient dynamics of a metal pushing V-belt CVT at high speeds *International Journal of Vehicle Design* **1** 37

[9] Srivastava N, Haque I 2009 A review on belt and chain continuously variable transmissions (CVT): Dynamics and control *Mechanism and Machine Theory* **1** 44

[10] Kong L, Parker R G 2008 Steady mechanics of layered, multi-band belt drives used in continuously variable transmissions (CVT) *Mechanism and Machine Theory* **2** 43

[11] Guebeii M, Micklem J D, Burrows C R 1992 Maximum transmission efficiency of a steel belt continuously variable transmission *American Sociaty of Mechanical engineers Conference on Power transmission and gearing Phoenix Arizona* **1** 43

[12] Sun D Z, Tan Z J, Guo D Z, Cheng N S 2002 Analysis of transmission efficiency for metal pushing V-Belt type CVT *Journal of Northeastern University (Natural Science)* **1** 23

[13] Liao J, Sun D Y, Qin D T 2003 Efficiency analysis of steel pushing V-belt continuously variable transmission in theory *Journal of Chongqing University* **3** 26

[14] Zhang W, Liu K, Zhou C G, Zhang H Y, Zhang B F 2010 Research on friction power loss of metal belt continuously variable transmission *China Mechanical Engineering* **11** 21

[15] Narita K, Priest M 2007 Metal-metal friction characteristics and the transmission efficiency of a metal V-belt-type continuously variable transmission *Proc IMechE, Part J: J Engineering Tribology* **1** 221

[16] Akehurst S, Vaughan N D, Parker D A, Simner D 2004 Modelling of loss mechanisms in a pushing V-belt continuously variable transmission. Part 1: torque losses due to band friction *Proc IMechE, Part D: J. Automobile Engineering* **11** 218

[17] Akehurst S, Vaughan N D, Parker D A, Simner D 2004 Modelling of loss mechanisms in a pushing V-belt continuously variable transmission. Part 2: pulley deflection losses and total torque loss validation *Proc IMechE, Part D: J. Automobile Engineering* **11** 218

[18] Akehurst S, Vaughan N D, Parker D A, Simner D 2004 Modelling of loss mechanisms in a pushing V-belt continuously variable transmission. Part 3: belt slip losses *Proc IMechE, Part D: J. Automobile Engineering* **11** 218

[19] Kobayashi D, Mabuchi Y, Kato Y 1998 A study on the torque capacity of a metal pushing V-belt for CVTs *SAE paper* No. 980822

[20] Zhang W, Liu K, Zhou C G 2010 Research on pulley deformation of metal belt continuously variable transmission *SAE Technical Paper* No. 01-1977

Authors

Wu Zhang, born on August 28, 1985, Shaanxi Province

Current position: a lecturer in the School of Mechanical Engineering of Xi'an University of Science and Technology, China
University studies: B.S. in Mechanical Engineering from Hu Bei University of Arts and Science in 2006; his M.S. in Vehicle engineering from Xi'an University of Technology in 2009 and his Ph.D. in Vehicle engineering from Xi'an University of Technology in 2012
Scientific interest: continuously variable transmission and mechanical transmission
Publications: 12 papers have been published and 7 patents have been authorized

Wei Guo, born on September 1, 1955, Shaanxi Province

Current position: professor in the School of Mechanical Engineering of Xi'an University of Science and Technology, China.
University studies: B.S. in Mechanical Engineering from Xi'an University of Science and Technology in 1982; his M.S. in Mechanical engineering from Xi'an University of Science and Technology in 1988.
Scientific interest: coal mine machinery and mine vehicle.
Publications: 70 papers and 2 monograph s have been published, and 9 patents have been authorized.

Chuanwei Zhang, born on October 20, 1974, Shaanxi Province

Current position: professor in the School of Mechanical Engineering of Xi'an University of Science and Technology, China.
University studies: B.S. in Mechanical Engineering from Xi'an University of Science and Technology in 1998; his M.S. in Mechanical engineering from Xi'an University of Science and Technology in 2001 and his Ph.D. in Mechanical engineering from Xi'an Jiaotong University in 2006.
Scientific interest: coal mine machinery and electric vehicle.
Publications: 30 papers and 3 monograph s have been published, and 6 patents have been authorized.

Yizhi Yang, born on April 11, 1985, Shaanxi Province

Current position: lecturer in the College of Humanities and Foreign Languages of Xi'an University of Science and Technology, China.
University studies: B.S. in School of english studies from Xi`an International Studies University in 2006; her M.S. in School of english studies from Xi`an International Studies University in 2009.
Scientific interest: English teach theory and english translation.
Publications: 4 papers have been published.

Yu Zhang, born on April 20, 1983, Shaanxi Province

Current position: engineer in the 41st Research Institute, The 6th Academy of China Aerospace Science And Industry Corporation, China.
University studies: B.S. in Mechanical Engineering from Hu Bei University of Arts and Science in 2006; his M.S. in Vehicle engineering from Northwestern Polytechnical University in 2009.
Scientific interest: dydrodynamics and mathematics.
Publications: 4 papers have been published.

A method of reliability modelling based on characteristic model for performance digital mock-up of hypersonic vehicle

Zhang Feng$^{1, 2*}$, Xue-Hui Feng$^{1, 2}$

^{1} *School of automation, Northwestern Polytechnical University, Xi'an 710072, China*

^{2} *School of Information Engineering, Yulin University, 719000, Yulin, China*

Received 3 March 2014, www.tsi.lv

Abstract

In order to grasp the complexity of the hypersonic vehicle dynamic characteristics, create its reliability control model, for the mathematical model of hypersonic vehicles is highly nonlinear and strong coupling, introduced the object-oriented modelling method, design a neural network, Petri control algorithm based on characteristics model, the mathematical model of the nonlinear is transformed into the equivalent linear model with control design requirements. And through the appropriate transform, design of the hypersonic vehicle dynamic inversion control system, building performance prototype reliability model based on Petri net, can stabilize the system, get decoupled affect purposes. Simulation results show that the performance digital mock-up reliability model is high accuracy, robustness, anti-jamming capability, has a good dynamic and steady-state performance.

Keywords: hypersonic vehicle, performance digital mock-up, characteristic model, reliability, flight control

1 Introduction

Nowadays, the hypersonic vehicle model and control has become a hot topic in the field of aerospace control, however, due to the lack of wind tunnel testing ground equipment, and flight test validation of hypersonic atmospheric properties is difficult, it is difficult to estimate; aircraft aerodynamic parameters cannot be experimentally means of access. Existing hypersonic aerodynamic parameters used in the model vehicle are basically simulated by the software platform for the try to simulate the real flight environment; the model also added a wide range of variants of unknown parameters and uncertainties. Hypersonic aircraft simulate these nonlinear, uncertainties, strong coupling characteristics of control theory and engineering presents a great challenge for the control of hypersonic aircraft, domestic and foreign scholars have used almost all kinds of advanced control theory [1].

The purpose of modelling is to in-depth analysis system, to facilitate control system design, based on the principles of regulation and modern control theory, are generally required before performing the first controller design modelling, based on the analysis of dynamic characteristics of an object, create the accused Mathematical model of the object. According to the mathematical model of the object dynamics controller design method in control theory and practical applications has played a significant role, however, with the continuous development of science and technology and production, the controlled object structure more complex, precise Dynamic modelling is becoming increasingly difficult; hand controller design as simple as possible in meeting the performance requirements of the situation. Academician Wu Hongxin is based on the above reasons, in twentieth century put forward the idea of characteristic modelling. The basic idea of characteristic modelling is: for high order complex object, in order to meet the control performance requirements, how to design a low order controller, that is to say to seek a kind of how objects modelling becomes simple, in order to facilitate the method of low order controller design [2].

2 Modelling method of object-oriented characteristic

2.1 FEATURE MODELLING CHARACTERISTIC

Traditional mathematical model of the system is usually use the time of continuous or discrete description and use mathematical formulas strictly controlled object depicts the static and dynamic performance. The characteristic model is introduced to describe the complex system characteristic quantity is unable or difficult to accurate mathematical modelling, using feature key points of system and key information are described, from another point of view, generally reflect the static and dynamic performance of the system. For a better understanding of characteristic modelling, here are a few basic concepts.

Define the characteristics of the model includes two points: First, it describes the dynamic behaviour and characteristics information, characterizing feature model is noteworthy here is that the whole system and not just the controlled object, including environmental information and control tasks belong to feature models:

* ***Corresponding author*** e-mail: tfnew21@sina.com

second, the characteristics of the system model to quantitatively characterize the key points and critical information, the system can be broadly reflect the static and dynamic performance. Third, though the model is accurate modelling features, but do not lose the characteristic information systems [3]. From a control standpoint, the system accurately characterizes the performance and behaviour of the static and dynamic mathematical model is a special form of feature model, which is characterized by the model complete content.

Feature model can be divided into two categories: deterministic model and intelligent characteristic model [4]:

(1) The deterministic model. It is mainly for those with high order complex mathematical equations describing the physical mechanism of clear. This kind of system is generally available lower order time-varying differential equations describing the form feature model.

(2) Intelligent characteristic model. Smart features complex object model is mainly aimed at those not currently using explicit mathematical equation, and with uncertainties. Clear but difficult to describe kinetic equation object from the object physical mechanism for those accused of starting a physical mechanism, establish a direct relationship between the characteristics of input and output control between the characteristic parameters of the model established by empirical research and expert object physical mechanism [4]. In addition, you can also other types of feature models, neural network identification model. Adaptive control structure based on characteristic model can be represented in Figure 1.

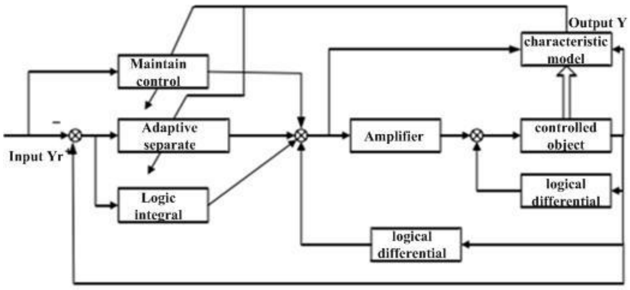


FIGURE 1 Adaptive control structure based on characteristic model

2.2 CREATE THE CHARACTERISTICS MODEL

Generally, when the unique characteristics of the model for its slow deterioration of equation. At present, the linear time-invariant systems feature modelling problem has been solved [5], for some special kind of linear time-varying systems, nonlinear systems, and multiple linear time-varying systems [6, 7], has also been proven that they can use features of the model deterioration or differential equations of second order equations when expressed. These features are derived in many models have been described in detail in the literature, this article only direct quote conclusions as needed. Some features of the model for single-input single-output systems have proven to be represented by the formula (1) as shown in the second-order differential equation.

$$\begin{aligned} m(u+1) &= f_1(u)m(u) + f_2u(u-1) \\ &+g_1(u)x(u) + g_2x(u-1) \end{aligned}, \tag{1}$$

where $m(u)$ is output, $x(u)$ is the amount of control, the equation is a characteristic parameter f_1, f_2, g_1, g_2, and the scope thereof can be determined in advance.

When the object is a minimum-phase system, or some non-minimum phase system, in order to simplify the engineering convenience, characterized model can also be used instead of the following form

$$m(u+1) = f_1(u)m(u) + f_2u(u-1) + g_1(u)x(u)\,. \tag{2}$$

For some multiple-input multiple-output system, characterized by second-order differential mode has proven to be used as output decoupling type equations expressed

$$\begin{aligned} U(u+1) &= \tilde{F}_1(u)U(u) + \tilde{F}_2(u)X(u-1) \\ &+\tilde{G}_1(u)X(k) + \tilde{G}_2(u)X(u-1) \end{aligned}. \tag{3}$$

Set the desired output is $U_r\left(u\right)$, the actual output is $U\left(u\right)$, the control amount of $X\left(u\right)$, $\tilde{F}_1(u)$, $\tilde{F}_2(u)$, $\tilde{G}_1(u)$, $\tilde{G}_2(u)$ is a characteristic parameter estimates. Identification of characteristic parameters shown by (4) Feature Model

$$\begin{aligned} u(k+1) &= f_1(k)u(k) + f_2(k)u(k-1) + g_1(k)x(k) \\ &= \Phi^T(k-1)\theta(k) \end{aligned}, \tag{4}$$

where

$$\Phi(k-1) = \left[u(k-1) \quad u(k-2) \quad x(k-1)\right]^T$$

$$\theta(k)\left[f_1(k) \quad f_2(k) \quad g_1(k)\right]^T$$

In order to achieve adaptive control, you must select the appropriate line identification method to estimate the parameters of f_1, f_2, g_1. Many online identification methods available, each method has its own characteristics, and total factor when combined with adaptive control when there is a question of the applicability. Applicability refers to the so-called identification accuracy, convergence speed computation, memory footprint, and in conjunction with the controller after effects can meet the performance requirements.

2.3 RELIABILITY CONTROL MODEL BASED ON PETRI NETS

Petri net is a kind of network information flow model, including the conditions and events with two types of nodes, in the conditions and events for the nodes of the two graphs and state information representing tokens distribution, and triggered rules of the event driven state evolution according to certain, which reflects the dynamic operation process of the system. Under normal circumstances, described the event nodes with small rectangular, called change; described the Condition node with a small circle, called the library [8]. Not between two nodes and two base nodes have directed arc is connected, and between nodes and base nodes can have directed arc connecting, which made up of two sub graph is called network. Some of the library network node on the number of dots tokens, which Petri network. This paper proposes a control algorithm based on fuzzy neural Petri net, the algorithm is described as follows:

A fuzzy neural Petri net is defined as nine tuple: $FNPN = (P, T, F, C, W, \mu, \alpha, \beta, M)$, where

$P = \{p_1, p_2, ..., p_m\}$ library is a finite set of the total;

$T = \{t_1, t_2, ..., t_n\}$ is a finite set of changes;

$F \subseteq (P \times T) \cup (T \times P)$ is a finite set of arcs total;

$C = \{X, Y, G\}$ is a set of propositions, wherein, $X = \{x_1, x_2, ..., x_n\}$ is the input proposition, $Y = \{y_1, y_2, ..., y_m\}$ intermediate proposition, $G = \{g_1, g_2, ..., g_k\}$ is the conclusion proposition ;

$W = \{w_1, w_2, ..., w_n\}$ is the input arc on the right set of values;

$\mu = \{\mu_1, \mu_2, ..., \mu_n\}$ is a finite set of fuzzy trust degree;

$\alpha = P \to C$ is the place to proposition mapping;

$M = P \to [0,1]$ is a mapping. Figure 2 shows the reliability of the model based on Petri nets.

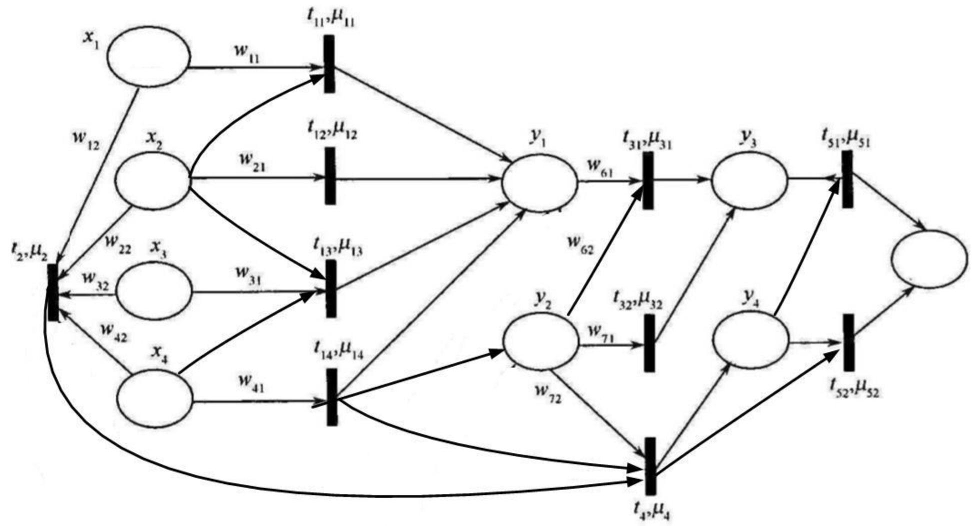


FIGURE 2 Reliability of the model based on Petri nets

For system reliability analysis of Petri network model, based on the 2/3 system as an example, assume that each component is equipped with a device, the reliability of the mock-up system analysis of Petri net model as shown in figure 3, Wherein p_1, p_2, p_3 denote the three components in a good state, as long as these three libraries library has two tokens, the system can work, there are tokens with p_{sc} indicates that the system is in good condition; p_{1f}, p_{2f}, p_{3f} are respectively the three members in the running state, the same library as long as these three libraries have two tokens, then the system is in the failure state; There are token representation with p_{sf} system failure; changes t_{1f}, t_{2f}, t_{3f} denote failure process three members, with a time characteristic, therefore, represented by a rectangular frame; similarly, changes of t_{1r}, t_{2r}, t_{3r}, respectively, during the operation of the three members; The other six changes are immediately change, change that as long as there are tokens in the input library and the corresponding input arc suppression library no token, the changes triggered immediately.

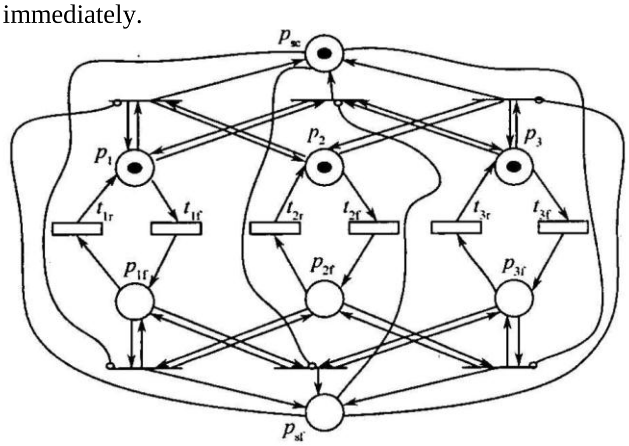


FIGURE 3 Reliability of the mock-up system analysis of Petri net model

2.4 RELIABILITY MODELLING BASED ON NEURAL NETWORKS

Neural networks are an important part of intelligent control, it has to learn, adapt, self-organization, arbitrary function approximation and massively parallel processing capabilities, particularly suitable for nonlinear control systems and uncertainty of the system. Involved in a wide range of neural network control, the paper selected for analysis only two typical networks. Figure 4 shows the single neuron adaptive control structure diagram. Single neuron adaptive controller control algorithm is as follows:

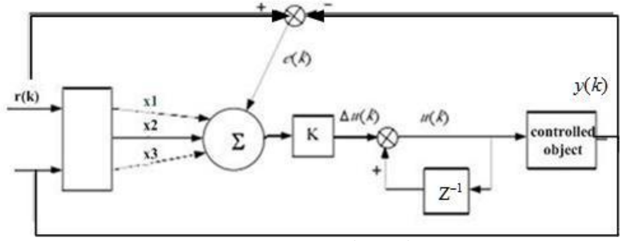


FIGURE 4 Instance of data fusion

$$x(k) = x(k-1) + K\sum_{i=1}^{3} w_i(k)u_i(k)\,, \tag{5}$$

where K is a proportional coefficient of neurons, $K > 0$, $w_i(k)$ is the weight coefficient, i = 1, 2,3. The controller is achieved by adjusting the weighting coefficients adaptive network, self-organizing features, the use of a learning rule to adjust the weights supervision, while adding items into supervised learning algorithm $z(k)$, then the weights of the neural network learning algorithm can be used to represent the formula (6):

$$\begin{aligned} w_1(i) &= w_1(i-1) + \eta z(i)x(i)u_1(i) \\ w_2(i) &= w_2(i-1) + \eta z(i)x(i)u_2(i)\,, \\ w_3(i) &= w_3(i-1) + \eta z(i)x(i)u_3(i) \end{aligned} \tag{6}$$

$$\begin{aligned} x(k) &= x(k-1) + K\sum_{i=1}^{3} w_i(k)u_i(k) \\ &= x(k-1) + Kw_1(k)z(k) + Kw_2\left[z(k) - e(k-1)\right]. \\ &+ Kw_3(k)[z(k) - z(k-2)] \end{aligned} \tag{7}$$

Observation equation (5) to (7) shows that the single neuron self-selection using the controller performance and K values have a great relationship. Petri nets and its role in controlling the proportion of adjustable parameters similar response when the system is slow, by increasing K, faster response times, but K is too large, it will produce a large amount of overshoot, and may even make the system does not stable.

3 Simulation and performance analysis

In this paper, the reliability of the model shown in Figure 2, for example, the reliability of the application block diagram of a network system of the bridge shown in Figure 5, the edges of the set order of the reliability R_a, R_b, R_c, R_d, R_e, can be obtained by the method of system reliability truth table degrees are:

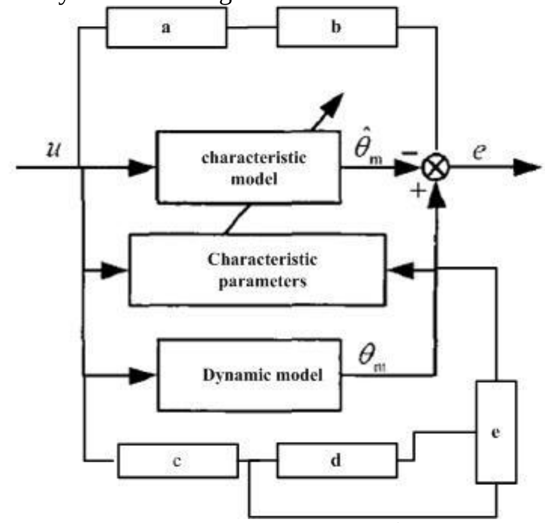


FIGURE 5 Reliability block diagram of the system

$$\begin{aligned} R_s &= (1-R_a)(1-R_b)R_cR_d(1-R_e) + \\ &(1-R_a)(1-R_b)R_cR_dR_e(1-R_e)R_bR_c(1-R_d)R_e \\ &+ R_aR_bR_cR_d(1-R_e) + R_bR_cR_dR_e \\ &+ R_a(1-R_b)(1-R_c)R_dR_e \end{aligned}. \tag{8}$$

Randomly generated 100 groups of component reliability value (0.85~1), the type (8) to calculate the system reliability as the sample value.

The model shown in Figure 2 as the reliability of the reliability of the estimated model bridge system, wherein, $x_1 \sim x_5$ represents successively member a, b, c, d, e reliability of 2.3 with the learning algorithm, using Java programming and a set of initial values are given on the assumption that the expert knowledge of the situation: $\omega_{11} = \omega_{21} = \omega_{31} = \omega_{41} = \omega_{51} = \omega_{61} = \omega_{71} = 0$, $\omega_{12} = \omega_{22} = \omega_{32} = \omega_{42} = \omega_{52} = \omega_{62} = 1$. All were taken to a trust, the step size chosen 0.0001, with 100 sets of sample data values and the trust of the right training, the training results shown in Figure 6.

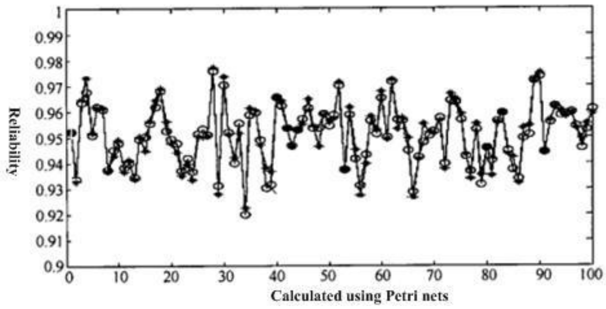


FIGURE 6 The results of training samples

Because the feature model needs to be calculated in the position loop in real time, so the combination of servo system, the sampling period in the process of simulation for Δt = 10ms. Because the system is open-loop state, so given the control value should not be too high, the simulation time should not be too long, make the simulation time in Δt = 40s. The given initial value is: $\tilde{\theta}(0) = 10^{-4} \times [1111]^T, P(0) = 10^7 \times I_{5\times5}$.

The simulation results are shown in figure 7.

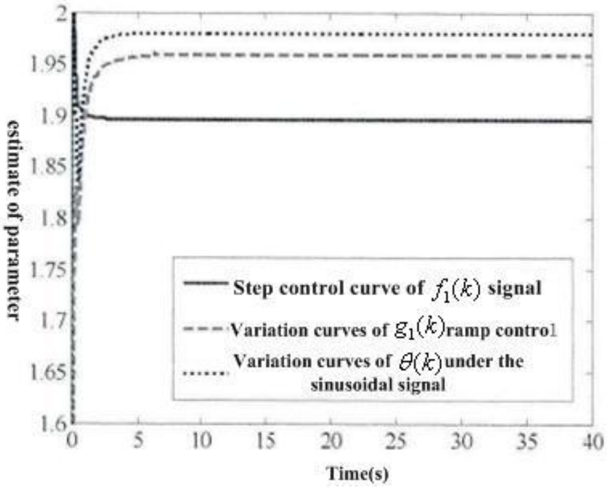


FIGURE 7 The curve under the three input control signal

4 Conclusions

Based on the existing research on the basis of characteristics of the model control theory, the amount of features and characteristics of the model are analysed, finishing, introduces the features of the model based on full coefficient adaptive control features, and gives the general design ideas and design steps. Petri nets control study and a number of neural network control is the essence of the model-based control features, the extension of the scope of this theory. Finally, the reliability of the design of the particular track motor control algorithm, the algorithm consists of two parts, the first part is fed inverse dynamics and analytical method are combined to solve the nominal pressure command and gesture commands based on feedback from the learning section Neurons adaptive control to correct the error. Simulation results show that the designed controller has a good track tracking capability.

Acknowledgments

This work is partially supported by the Defence basic research projects of **#A0420131501**, Research and Cooperation Project for Department of Yulin city of (**#Sf13-08, #2012cxy3-26**). Thanks for the help.

References

[1] Cai Yamei, Wang Liping 2010 Hypersonic Programs in USA and Key Technologies Analysis *Aerospace Manufacturing Technology* **12**(6) 4-7 *(In Chinese)*

[2] Buschek H, Calise A J 1997 Uncertainty Modelling and Fixed-Order Controller Design for a Hypersonic Vehicle Model *Journal of Guidance, Control, and Dynamics* **20**(1) 42-8

[3] Wilcox Z D, MacKunis W, Bhat S, et al. 2010 Lyapunov-Based Exponential Tracking Control of a Hypersonic Aircraft with Aerothermoelastic Effects *Journal of Guidance, Control, and Dynamics* **33**(4) 1213-24

[4] Wu H X, Meng B 2009 Review on the control of hypersonic flight vehicles *Adv Mech* **39** 756-65

[5] Zhang Z, Hu J 2011 Prediction based guidance algorithm for high-lift re-entry vehicles *Sci China Inf Sci* **54** 498-510

[6] Li X D, Xian B, Diao C, et al. 2011 Output feedback control of hypersonic vehicles based on neural network and high gain observer *Sci China Inf Sci* **54** 429-47

[7] Hirtz J, Stone R, Mcadams D, et al. 2002 A functional basis for engineering design:reconciling and evolving previous efforts *Research in Engineering Design* **13**(2) 65-82

[8] Su J, Renaud J E 1997 Automatic Differentiation in Robust Optimization *AIAA Journal* **35**(6) 1072-9

Authors

	Zhang Feng, born on June 26, 1980, in Shannxi Yulin **Current position, grades:** associate professor in Yulin University. **University studies:** MS degree in Computer science from Xidian University in 2009. **Scientific interest:** Cloud integrated manufacturing technology, the modeling of complex systems, the Internet of Things applications.
	Xue Hui-feng, born on June 16, 1964, in Shanxi Yuncheng **Current position:** professor in Northwestern Polytechnical University **University studies:** PhD degree in Water resource economics from Xi'an Polytechnic in 1995 **Scientific interest:** modelling of complex systems, Simulation and performance evaluation, management, systems engineering, energy and environmental systems engineering, computer control, intelligent control, network control

Research of key technology on self-propelled farmland levelling machine and hydraulic servo system simulation

Jiangtao Liu*, Jinggang Yi

Mechanical & Electronic Engineering College, Agricultural university of Hebei, BaoDing, 071001, China

Received 1 March 2014, www.tsi.lv

Abstract

According to the present situation of the farmland levelling, the equipment cost is high, maintenance is complex and its cost is high. The paper carries a research on the key technology of self-propelled farmland levelling machine. The key technology includes the levelling knife, the levelling part, the sundry separating device and the measurement and control system of the laser and inclination sensor. At the same time, the paper establishes the hydraulic servo system mathematical modal and utilizes MATLAB to analyse, revise and simulate for the system mathematical modal.

Keywords: farmland levelling machine, levelling knife, levelling part, inclination sensor, simulation

1 Introduction

China is a large agricultural country. The agriculture is the major water consumer and the surface irrigation occupies the dominant position in China's agricultural irrigation。According to the analysis, the field partial loss accounts for about 35% in the loss of irrigation water, so the field water-saving has the great potential. The cause of the field water loss includes that bedding block is too large, the land is not smooth, or the field exists a lot of sundries, such as waste plastic, hard straw, weeds, brick and tile, which the irrigation is not uniform and the deep seepage is serious. The research shows that when the land levelling error is less than 1~2cm, the inch water don't exposes the mud; the amount of shallow water irrigation can achieve the accurate water and the water saving is about 30~50%; it also can reduce fertilizer loss, improve the utilization ratio of the fertilizer. In the drought area, it can keep the moisture and improve the germination rate. At the same time, the levelling field can make the seeding depth uniformity and the seedling tidy and also make the crops get the required optimum water during the whole growth stage to improve the crop yield [1].

Since the 80's of 20 centuries, laser grader technology has attracted the wide attention from the scientific community and industry of china. Some large farms and enterprises imported the laser control grader to level the farmland [2].

Since the early 1990s, some schools and research institutions in China have also studied the laser grader. In 1996 Heilongjiang Academy of land reclamation sciences and Beijing Institute of Technology successfully developed agricultural laser grader of 1PTY-6. In 1997 Aviation Industry Corporation of China completed the project of the laser calibration grader [3]. In 2003 Northeast Agricultural University designed and developed the laser grader of 1PJY-3.0 [4]. Research mainly focused on the flat shovel. In 2007 South China Agricultural University designed a laser land leveller for paddy [5]. Since the late 1990s China Agriculture University devoted oneself to design and develop the farmland grader. The system adopts laser and the hydraulic system to level [2, 3, 6, 7].

In the nineteen seventies, The United States first applied the laser technology in agricultural grader, and had made the great economic benefit and social benefit [8]. America Spectral Precision Instrument Company successfully designed and developed the first set of the laser knife plate [3]. Because the laser knife plate levelling system had many unique technique effects and the great economic benefit, it obtained the fast development. In the 80's many foreign enterprises producing the grader is equipped with the laser levelling system, such as America's DRESSR, America's Spectra-Physics Company, America's TOPCON Laser Systems Company, German Boukema Company, Construction Machinery Company (Habaumag) and Swiss Firm Leica etc. In the 90's many developing countries also had used the laser land levelling technology, and achieved the good economic benefit, for example India, turkey and Pakistan etc. In American and Portugal, the use of the farmland levelling technique make the farmland irrigation uniformity improve from 17 to 20% and the crop yield increase by 7~ 31%; In India, the water saving is about 15 ~ 20%; in Turkey, the irrigation water efficiency is improved by 25 ~ 100%, the wheat yield is increased by 35 ~ 75%, the cotton yield is increased by 20 ~ 50% [2]. At present the grader has combined the advanced achievements in other fields in the developed industrial

* ***Corresponding author*** e-mail: liujiangtao2003@126.com

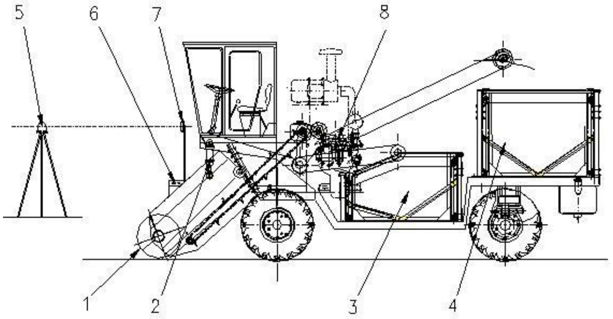
countries, led by the US, Europe and Japan. The advanced achievements include all wheel drive technology, laser automatic levelling device, electronic monitoring system etc. [7, 8, 9, 10, 11].

Based on the surface levelling machine that was successfully developed by the research group of the author without the sundries cleaning function [12, 13, 14, 15], the research group of the author studies and designs the self-propelled farmland levelling machine that not only can level but also clean the sundries. In addition to the recent studies of the research group, the domestic and foreign similar studies are the grader.

According to the request, this paper focuses on the levelling knife, the levelling part, the sundry separating device and the measurement and control system of laser and inclination sensor. At the same time, the hydraulic system mathematical modal is simulated through MATLAB.

2 Overall structure

Self-propelled farmland levelling machine includes levelling part, laser receiver, laser emitter, inclination sensor, hydraulic cylinder, sundry separating device, soil box and sundries box etc. The structure figure of self-propelled farmland levelling machine sees Figure 1. Levelling part installed in the front is connected with the frame by bolts. The hydraulic control system controls the levelling part to work on the plane that parallels with the datum plane. At the same time, the levelling part removes the mixed soil, and it is transported to the sundry and soil separator through the conveyor belt for the sundry and soil separation. Then the sundry and soil is respectively transported to the sundry box and soil box.



1. Levelling part 2. Hydraulic cylinder 3. Soil box 4. Sundry box
5. Laser emitter 6. Inclination sensor 7. Laser receiver
8. Sundry separating device
FIGURE 1 Self-propelled farmland levelling machine structure figure

3 Key technology

3.1 LEVELLING KNIFE

According to the working requirement, the levelling knife achieves two purposes. The first purpose is to cut the soil and collect the crushing soil containing the sundry and the second purpose is to ensure the soil surface roughness. In order to be able to efficiently cut the soil, using sliding mode; and in order to realize the broken soil collection function with the sundry, the levelling knife adopts a curved plates. Three-dimensional map of the levelling knife sees Figure 2.

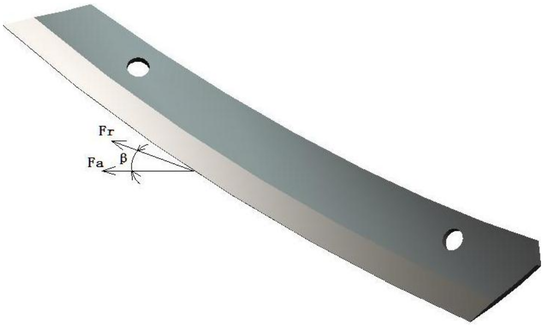


FIGURE 2 Three-dimensional map of the levelling knife

When self-propelled farmland levelling machine works, the point of the levelling knife firstly contacts with the soil, and then the edge contacts with the soil one by one, which changes the past scraper way and reduces the forward resistance.

Each piece of the levelling knife should ensure that the lowest position at any point in the blade is at the same altitude, or in the condition to keep the spindle levelling, the gyration radius is equal at each point on the edge. Because the spiral levelling knife has the helix angle (*β*) and the soil with the sundry is cut from the main forces in the normal direction of the blade, it is thrown in the same direction. Therefore, the helix angle can control the throwing direction of the crushing soil. The simulation and experiment results show that the helix angle is appropriate from 65° to 78°.

The absolute motion of the levelling knife is composed of two kinds of motion at work. One is the circular motion around the centre of the levelling shaft, another is the linear motion of the levelling knife with self-propelled farmland levelling machine. When self-propelled farmland levelling machine works two kinds of motion produces the effect for the levelling knife to generate the moving track of cosine cycloid. The moving track of cosine cycloid sees Figure 3. The moving equation of the levelling knife sees formula.1.

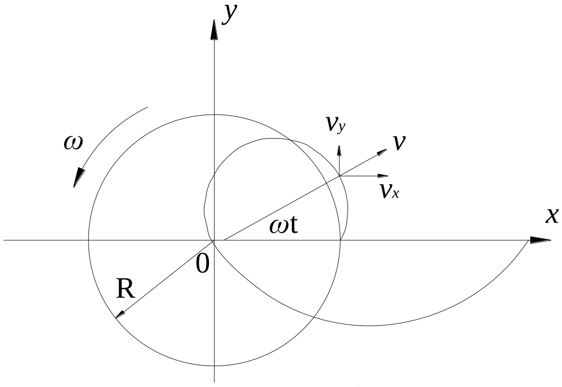


FIGURE 3 The moving track of cosine cycloid

$$\begin{cases} x = v_m t + R_i \cos\omega t \\ y = R_i \sin\omega t \end{cases}, \quad (1)$$

where v_m is the forward speed of self-propelled farmland levelling machine, ω is the angular speed of the levelling shaft, R_i is the rotating radius of the levelling knife.

The above equation is differentiated to obtain the speed of the levelling knife.

$$\begin{cases} v_x = v_m - R_i\omega \sin\omega t \\ v_y = R_i\omega \cos\omega t \end{cases}, \quad (2)$$

The levelling and cutting knife point speed is as follow:

$$v = \sqrt{v_m{}^2 + R_i{}^2\omega^2 - 2v_m R_i \omega \sin\omega t}\,. \quad (3)$$

3.2 LEVELLING PART

The spiral levelling knifes are uniformly and symmetrically installed on the levelling shaft. The three-dimensional map of the levelling part sees Figure 4. When self-propelled farmland levelling machine works, the levelling knifes cut into the soil in turn to realize the continuous and stable cutting and restrain the shock in cutting process. Because of the helix angle (β) the broken soil and the sundry is thrown along the direction of the vertical edge tangent. Therefore, it converges the middle symmetry plane in the thrown process and then is transported.

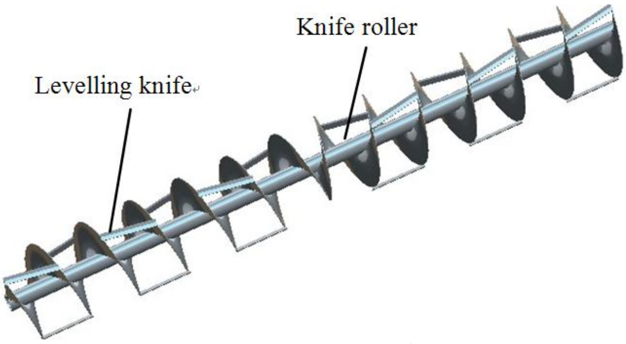
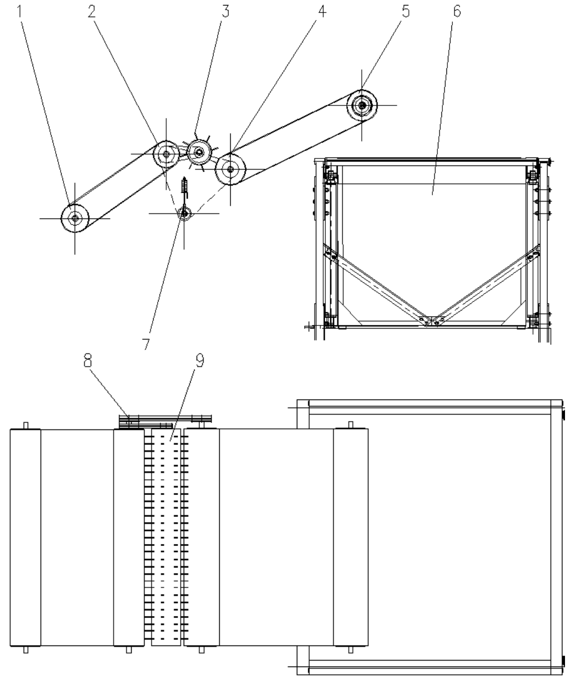


FIGURE 4 The three-dimensional map of the levelling part

The levelling parts configuration directly relates to smoothly cut the complex soil (containing the sundry) and reduce the power consumption. The research adopts the combination-levelling mode of the roller and cutter. The levelling knifes are uniformly arranged and welded in the spiral knife roller. The inclination angle of the levelling knife and the helix angle of the knife roller are equal. The theoretical analysis and practical experiment shows that the above arrangement mode can make the levelling knife easily cut the soil and reduce the power consumption.

3.3 SUNDRY SEPARATING DEVICE

At present, the grader can only level the soil and cannot clean up the sundry. In order to make up for the current grader flaw, the research group designs the sundry separating device. It is installed in the farmland-levelling machine and can effectively separate the soil and sundry. The device installs the separating roller of the soil and sundry with a certain amount of spring tooth in the middle in order to effectively separate the soil and sundry. Both ends of separating roller are respectively installed a conveyor belt. The front conveyor belt transports the soil with the sundry and the back conveyor belt transports the separated sundry to the sundry box. The sundry separating device sees Figure 5.



1, 2, 4, 5. Conveyor belt wheel 3. Spring tooth 6. Sundry box
7. Tension wheel 8. Power belt 9. Soil and sundry separating roller
FIGURE 5 Sundry separating device

3.4 MEASUREMENT AND CONTROL SYSTEM

The measurement and control system mainly includes inclination sensor, laser emitter, photo-electricity sensor, laser receiver, levelling control system, hydraulic servo system and levelling execution part. The Measurement and control system structure sees Figure 6.

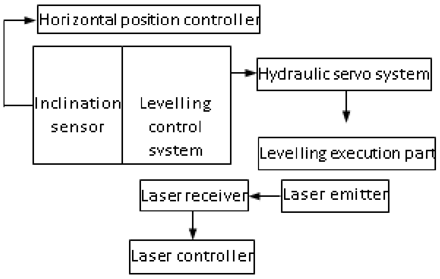


FIGURE 6 The system hardware structure

3.4.1 Laser working principle

Laser emitter emits a laser beam that can rotate 360° to scan and form the datum plane. Laser receiver is installed in the mast of the levelling part. Laser receiver receives the laser signal to transmit to the controller. If the above receiver receives the datum laser signal, which shows that the levelling part locates below the working plane and the correction signal improving the levelling part is transmitted to the hydraulic servo system, whereas if the under receiver receives the datum laser signal, which shows that the levelling part locates above the working plane and the correction signal reducing the levelling part is transmitted to the hydraulic servo system. After the hydraulic control system receive the correction signal from the levelling control system, the hydraulic servo system controls the levelling execution part to improve or reduce the levelling part to make the levelling part to work on the plane that parallels with the datum plane. When the middle receiver receives the datum laser signal, which shows that the levelling part levels. The laser working principle sees Figure 7.

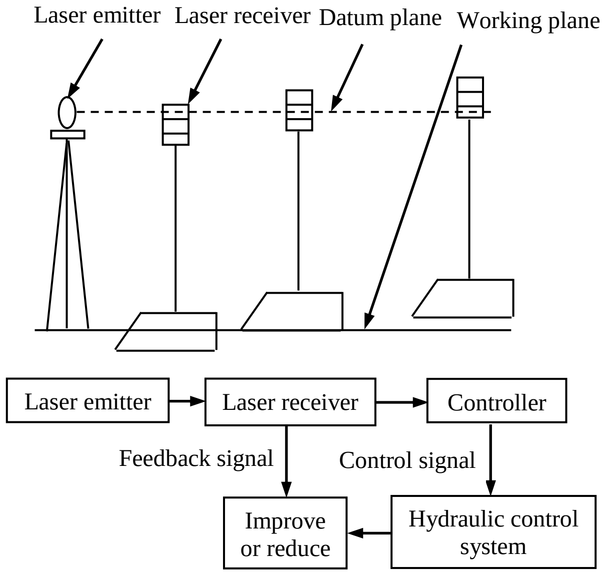


FIGURE 7 The laser working principle

3.4.2 Inclination sensor working principle

The paper uses the liquid pendulum inclination sensor to measure the inclination angle of the levelling part. The sensor is equipped with the conductive liquid in the glass shell, and has three platinum electrodes to connect with the external. Three electrodes are parallel to each other and have the equal distance. The conductive liquid of between two electrodes is equivalent to two resistors R1 and R2. When the levelling part levels, the electrode depth inserted into the conductive fluid is equal or R1 is equal to R2, and the control system doesn't output the signal, whereas when the levelling part inclines, the middle electrode depth inserted into the conductive fluid is fixed and the electrode depth inserted into the conductive fluid isn't equal on both sides or R1 isn't equal to R2.The control system outputs the signal. After the hydraulic control system receive the inclination signal from the levelling control system, the hydraulic servo system controls the levelling execution part to adjust the levelling part to make the levelling part to work on the plane that parallels with the datum plane. The inclination sensor working principle sees Figure 8.

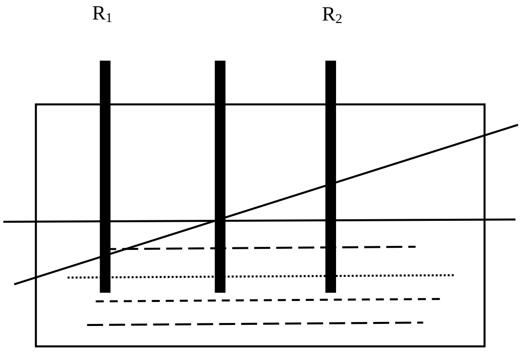


FIGURE 8 The inclination sensor working principle

4 Hydraulic Servo System Simulation

4.1 MATHEMATICAL MODEL OF HYDRAULIC SERVO SYSTEM

Self-propelled farmland levelling machine requires the high adjustment precision, the fast reaction and the easy parameter real-time feedback. So the hydraulic servo system adopts the closed-loop system of the valve control hydraulic cylinder to control the levelling execution part. The mathematical modal of the hydraulic servo system sees Figure 9.

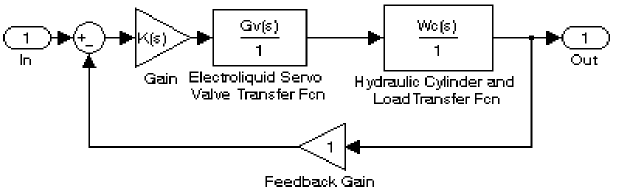

1
In
+
-
K(s)
Gain
Gv(s)
1
Electroliquid Servo Valve Transfer Fcn
Wc(s)
1
Hydraulic Cylinder and Load Transfer Fcn
1
Out
1
Feedback Gain

FIGURE 9 The mathematical modal of the hydraulic servo system

4.2 SYSTEM ANALYSIS, SIMULATION AND ADJUSTMENT

Using the control system toolbox compiles the applied program to analyse the opened loop transfer function of the hydraulic servo system. Step and bode figure of adjusting front and back mathematical model is separately drawn. Sees Figure 10 and Figure 11.

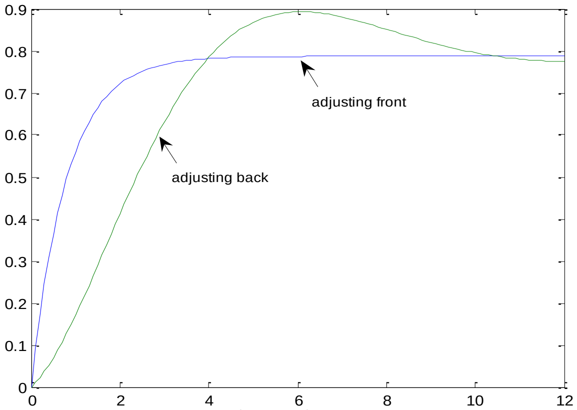


FIGURE 10 Step figure of mathematical model

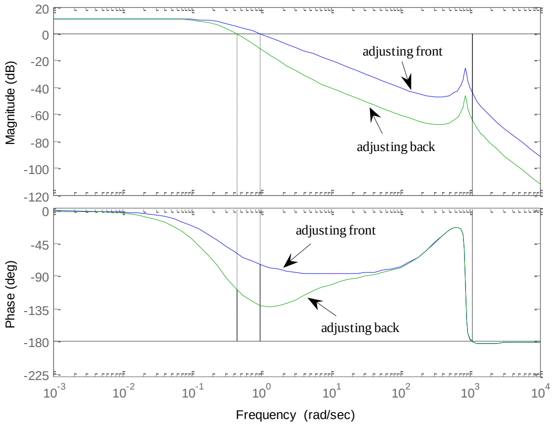


FIGURE 11 Bode figure of mathematical model

As we can see in Figure 10, the damping coefficient of adjusting front is approximate to one. The system has not the sigma. The system response is slow; the adjusting time is long and the fast reaction lags. Therefore, the control system need be adjusted.

After the system is adjusted, the adjusting function is as follow:

$$G_1(s) = \frac{0.08523s + 1}{1.023s + 1}, \tag{4}$$

According to Fig.10 after the system is adjusted, the damping coefficient is approximate to the optimal value. The adjusting time becomes short and the fast reaction moves up. And the sigma is small. According to Fig.11 after the system is adjusted, the control system is stable. The system can satisfy with the precision request of self-propelled farmland levelling machine.

6 Conclusions

According to the need of farmland levelling operation, the paper designs the levelling knife, the levelling part, the sundry separating device and the measurement and control system of the laser and inclination sensor. Self-propelled farmland levelling machine not only can level but also clean the sundries. At the same time, it greatly reduces the labour intensity of the farmland reclamation and improves the levelling efficiency. The hydraulic servo system simulation shows that the hydraulic control system is stable and reliable.

References

[1] Liu Jiangtao, Cui Baojian, Yi Jinggang 2012 Research of Control and Measurement System on Self-propelled Farmland Cleaning and Flatting Machine *Journal of Agricultural Mechanization Research* **2012**(6) 101-2 *(In Chinese)*

[2] HOU Ming-liang 2005 *Researehon Hydraulie Control Systemin Laser Land Leveling Maehine of Drag Type* Bei Jing: China Agriculture University 4-5 *(In Chinese)*

[3] AI Wei-zhong, JIANG Ping, SUN Song-lin, LUO Ya-hui 2010 Present situation and developing trend of the laser control techniques in farm land levelling in Hunan *Hunan Agricultural Machinery* **137**(2) 16-7 *(In Chinese)*

[4] Han Bao 2003 Development of 1JPY-3.0laser controlled compositive levelling machine *Transactions of the CSAE* **19**(3) 116-20 *(In Chinese)*

[5] Li Qing, Luo Xiwen, Wang Maohua, et al. 2007 Design of a laser land leveler for paddy field *Transactions of the CSAE* **23**(4) 88-93 *(In Chinese)*

[6] HOU Ming-liang 2008 Test Study of the Hydraulic Control System in Laser Levelling Machine *Transactions of the CSAE* (6) 24 *(In Chinese)*

[7] Liu Jiangtao, Cui Baojian, Jiang Haiyong, Yi Jinggang 2013 Measurement and Control System of Self-propelled Levelling Machine Based on Inclination Sensor and Laser *Sensors & Transducers Journal* **159**(11) 87-91 *(In Chinese)*

[8] Rajeev Ranjan, Mandal N K 2012 Study of Vibration Characteristics of a Multi Cracked Rotating Shaft Using Piezoelectric Sensor *Sensors & Transducers Journal* **147**(12) 45-52

[9] Davis G 1979 The Use of Laser in Lang Forming *Power Forming Magazine* (7) 12-5

[10] Zhang Yong-Jie, Chun-Feng L V, Yang Jin-Feng, Liu Wei-Wen, Zhao Hui 2012 Design of Cursor Magnets on the Wiedemann Effect in Magnetostrictive Linear Position Sensors *Sensors & Transducers Journal* **138**(3) 80-3 *(In Chinese)*

[11] SOETEDJO Aryuanto, NURCAHYO Eko, PRAWIDA Fiqih 2013 Photodiode Array for Detecting Laser Pointer Applied in Shooting Simulator *Sensors & Transducers Journal* **151**(11) 78-83

[12] CHEN Jing, YI Jinggang, JIANG Haiyong, XING Yazhou 2007 Foundation trench clearing machine *Construction machinery* (1) 101-2

[13] CHEN Jing, YI Jing-gang, JIANG Hai-yong, XING Ya-zhou, LIU Jiang-tao 2007 ZM3-1 Foundation Trench-leveling Machine *Journal of Agricultural Mechanization Research* (2) 105-8

[14] JIANG Hai-yong, YI Jing-gang, QI Xiaona, XING Ya-zhou, LIU Jiang-tao, CHEN Jing 2007 Design of cutting shaft for trench-levelling machine and the efficiency analysis *Construction machinery* (8) 65-7

[15] JIANG Hai-yong, QI Xiaona, LIU Jiang-tao, YI Jing-gang, XING Ya-zhou1, CHEN Jing 2007 Design of Cutting Shaft for Scraper Machine and the Efficiency Analysis *Journal of Agricultural Mechanization Research* (10) 46-8

Authors	
	Liu Jiangtao Liu, born on October 4, 1978, China **Current position, grades:** University teachers, lecturer **University studies:** Mechanical engineering **Scientific interest:** Integrative technique of mechanics-electronics-hydraulics **Publications:** 16 **Experience:** In 1999 I studied at mechanical & electronic engineering college, agricultural university of Hebei, mechanical engineering and Automation Specialty. In the university period, I won 6 scholarships, passed through the university English six levels, and served for 3 years as a monitor. I graduated from the mechanical engineering and automation specialty in agricultural university of hebei in 2003 and acquired the degree of bachelor of engineering. In 2003 I studied at agricultural university of hebei, agricultural mechanization engineering. I graduated from the agricultural mechanization engineering in agricultural university of hebei in 2006 and acquired the excellent graduate and the master degree of engineering. Since 2006 I taught at the agricultural university of hebei. In 2008 I was promoted to lecturer. I major in the research of mechanical engineering machinery, have published 16 papers, have presided over two research projects and participated in five international conferences for three times over an academic report.
	Jinggang Yi, born on October 2, 1962, China **Current position, grades:** Department dean, Professor **University studies:** Mechanical engineering, Agricultural engineering **Scientific interest:** CAD/CAM technology, Electromechanical integration technology **Publications:** 50 **Experience:** In 1980 I studied at mechanical & electronic engineering college, agricultural university of Hebei, tractor repair and manufacture specialty. I graduated from tractor repair and manufacture specialty in agricultural university of hebei in 1984 and acquired the degree of bachelor of engineering. Since 1984 I taught at the agricultural university of hebei. In 2005 I was promoted to professor. I major in the research of mechanical engineering and agricultural engineering machinery, have published 50 papers, have presided over ten research projects and participated in sixteen international conferences for ten times over an academic report.

Comfort and energy-saving control of electric vehicle based on nonlinear model predictive algorithm

Manli Dou[1, 2*], Chun Shi[1, 2], Gang Wu[1, 2], Xiaoguang Liu[1, 2]

[1] *School of Information Science and Technology, University of Science and Technology of China, China*

[2] *National Electric Vehicle System Integration Engineering Research Center, China*

Received 1 February 2014, www.tsi.lv

Abstract

This paper develops a control-oriented drivability model for an electric vehicle and a nonlinear model predictive optimization algorithm for an electric vehicle. A cost function is developed that considers the tracking error of setting value and the variation of control volume. Longitudinal ride comfort and energy-saving is also considered. Simulations show that the developed control system provides significant benefits in terms of fuel economy, vehicle safety and tracking capability while at the same time also satisfying driver desired car following characteristics.

Keywords: Nonlinear Model Predictive, Comfort, Energy-saving, Electric Vehicle

1 Introduction

The legislation on reduction of fuel consumption and CO2 emissions has created a large interest in electric vehicle (EV) technologies. Electric vehicles have developed by world's major car manufacturers in recent years. The EVs have several advantages over vehicles with internal combustion engines, such as energy efficiency and environmental friendliness, and is seen to the right way to solve the energy and environment problems [1].

During development and calibration phases of an EV control system, it is of crucial importance to assess and optimize vehicle drivability.

Ride comfort is classified to three categories: vertical, horizontal, and longitudinal vibration. The variation in longitudinal acceleration has great influence on ride comfort [2]. Therefore, longitudinal ride comfort is one of the most crucial features to most advanced vehicle control systems.

The energy of battery carried by electric cars is limited. In the motion the process vehicles, charge and discharge processes always happen. The value of battery discharge current will directly affect the actual capacity of the battery; smaller discharge current makes the battery emit more energy [3]. Therefore, control the value of the charge current can save energy.

An important advantage of model predictive control (MPC) is its ability to cope with constraints on controls and states in an explicit and optimal way, and has experienced a growing success for slow complex plants. In the last decades, several developments have allowed using these methods also for fast system such as automotive application [4].

This paper will focus on development and implementation of a nonlinear model predictive optimization algorithm for an electric vehicle. It is organized as follows. In section 2, a nonlinear model for electric vehicle is introduced. In section 3, a model predictive optimization problem with constraints is constructed considering riding comfort, fuel economy and driver desired response. In section 4, the infeasibility issue is processed and the control law is numerically solved. In section 5, its success is demonstrated by simulations.

2 Nonlinear model for electric vehicle

A typical driving system of pure electric bus is mainly constituted by a traction electric machine (motor), a propeller shaft, a universal joint, a final drive and drive shafts [4]. Figure 1 shows the driving system configuration.

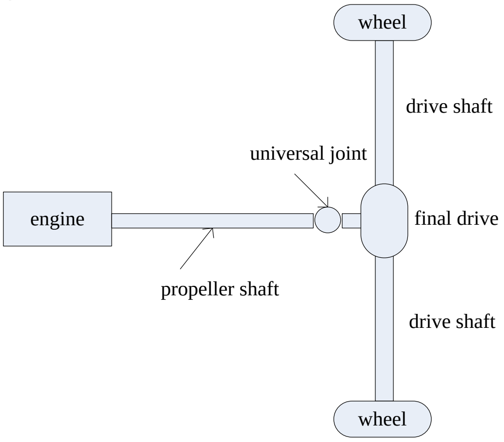


FIGURE 1 Driving system configuration

[*] ***Corresponding author*** e-mail: liuxg88@mail.ustc.edu.cn

From the mechanism model of the longitudinal dynamics of the vehicle, an input-output model can be deduced.

As the torque applied to the shaft T_{shaft} is generated by the torque on the motor shaft T_{em}, taking into account the main reduction gear ratio ς_{rd} and the transmission efficiency η, the relationship between T_{shaft} and T_{em} can be express as:

$$T_{shaft} = T_{em}\varsigma_{rd}\eta\,, \tag{1}$$

Ignoring the slip ratio of the tire and assume $v_{veh} = w_{wheel} r_{wheel}$, the driving force F_x can be express as:

$$F_x = \frac{T_{em}\varsigma_{rd}\eta - T_{br}}{r_{wheel}} - \frac{J_{wheel}\dot{v}_{veh}}{r^2{}_{wheel}} - F_{roll,fr}\,, \tag{2}$$

where r_{wheel} is the radius of the wheel, w_{wheel} is the wheel's angular velocity, v_{veh} is the vehicle's longitudinal velocity.

Using Newton's second law on the longitudinal direction, the following equation is obtained

$$\dot{v}_{veh} = \frac{1}{M_{veh}}(2F_{x,f} + 2F_{x,r} - \frac{1}{2}\rho_{air} C_d A v_{veh}^2 - M_{veh} g \sin(\gamma))\,, \tag{3}$$

where ρ_{air} is the air density, C_d is the vehicle's drag coefficient, A is the vehicle's frontal area, γ is the climbing angle.

Take (2) into (3), ignore the difference of the vertical forces acting on the front and rear wheels, the longitudinal dynamics of the vehicle can be expressed as [5]:

$$M_{veh}\delta\dot{v}_{veh} = \frac{T_{em}\varsigma_{rd}\eta - T_{br}}{r_{wheel}} - F_{roll,fr} - \frac{1}{2}\rho_{air} C_d A v_{veh}^2 - M_{veh} g \sin(\gamma)\,, \tag{4}$$

where the $\delta = 1 + \frac{\sum J_{wheel}}{r^2{}_{wheel}}$ is rotating mass conversion factor.

The vehicle is assumed to driven on straight roadway, so the road inclination γ is zero, and the rolling friction $F_{roll,fr}$ is assumed to be constant, the relationship between T_{em} and v_{veh} can be express as an input-output system:

$$a_0 + a_1 v_{veh}^2 + a_2\dot{v}_{veh} = b_1 T_{em}\,. \tag{5}$$

The input variable of the system is T_{em}, and the output variable is v_{veh}, and b_1, a_0, a_1, a_2 are parameters.

According to (5), the longitudinal dynamics of the vehicle is nonlinear, i.e. the relationship between the motor torque T_{em} of pure electric vehicle and the vehicle speed v_{veh} is nonlinear.

The discrete model can be obtained by discretization of (5) using forward difference scheme:

$$y(k+1) = -\theta_1 y^2(k) - \theta_2 y(k) + \theta_3 u(k) + \theta_4\,, \tag{6}$$

where y is v_{veh}, u is T_{em}, $\theta = [\theta_1, \theta_2, \theta_3, \theta_4]$ is the parameter vector.

The (6) shows the non-linear model is a Non-linear Auto Regressive with eXogenous inputs (NARX)model [6], and as the model depends on its parameters in linear way, it can be treated as a linear-in-the-parameters model.

The linear-in-the-parameters model takes the form:

$$y(k+1) = \theta x(k) + \varepsilon(k+1)\,, \tag{7}$$

where $x(k) = [-y^2(k), -y(k), u(k), 1(k)]'$.

Suppose N data samples $\{x(k), y(k)\}_{k=1}^{N}$ are used for model identification, equation (8) can be formulated as:

$$Y = \theta X + \Xi\,, \tag{8}$$

where $Y = [y(1),...,y(k)]'$, $X = [x(1),...,x(k)]'$, $\Xi = [\theta_1,...,\theta_2]$. Then use least-squares (LS) method to determine the parameters.

3 Construction of nonlinear predictive optimization problem

Model predictive control (MPC), also referred to as moving horizon control or receding horizon control, is a control strategy in which the applied input is determined on-line at the recalculation instant by solving an open-loop optimal control problem over a fixed prediction horizon into the future [7].

The basic overall structure of a NMPC control loop is shown in Figure 2. Based on the applied input u_t and the measured outputs y_t, estimate model predictions outputs y_t is obtained. This estimate is fed into the NMPC controller, which computes a new input applied to the system. Often an additional set-point calculation model is added to the overall loop to produce setting sequence w_t.

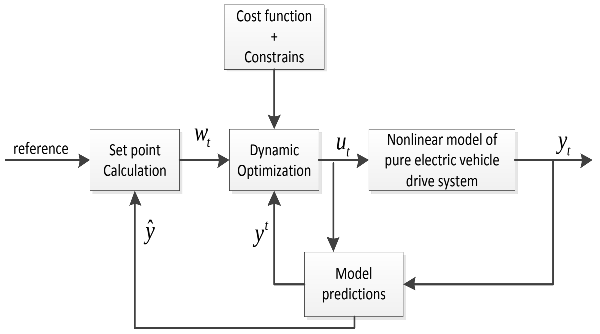


FIGURE 2 NMPC control loop

The set-point target calculation model is used to explain the driver's driving intentions. In the conventional vehicle driving behaviour, when driver depresses the accelerator pedal, the power output of the engine increases, and vehicle speed increases. In order to maintain driving habits unchanged, the pedal opening degree should correspond to the acceleration demand. So pedal opening degree α is mapped to acceleration value a as:

$$a = f(\alpha), \tag{9}$$

and the speed set point curve of vehicle is obtained as:

$$w_t(k) = v_0 + a * k\,, \tag{10}$$

where v_0 is current vehicle speed, and k is sampling time point.

When vehicle is driven on the road, acceleration and deceleration occurs frequently according to the driver's instructions start and stoplights, and this result in vibration, which has a great impact on the ride comfort. Therefore, improving the comfort of the vehicle during acceleration and deceleration, and can greatly improve the overall comfort of the vehicle.

Therefore, the maximum value of the vehicle acceleration and deceleration should be limit as:

$$\Delta y_{n\_\max} \le \Delta y_{t+i} \le \Delta y_{p\_\max}\,, \tag{11}$$

where y is the vehicle speed, and Δy stands for vehicle acceleration when $\Delta y \ge 0$ and deceleration when $\Delta y < 0$, $\Delta y_{n\_\max}$ and $\Delta y_{p\_\max}$ are the maximum value of the vehicle acceleration and deceleration.

As the energy carried by the pure electric vehicle is limited energy, efficient use of battery power on the extension of a pure electric vehicle driving range is very important. Driving range is one of the economic indicators of electric vehicle.

The actual capacity of the battery is discharged under certain conditions of the actual release of the battery power, typically less than the theoretical capacity and the rated capacity. The actual capacity of the battery is affected by the value of discharging current, the ambient temperature, the aging of battery and so on [8].

Taking into account the effect of the battery capacity ratio, low current discharge rate can reduce the loss energy and so that the battery can be more energy.

In pure electric vehicle, when the motor is in electric state, battery release energy, and the current flows to the motor from the battery; when the motor is in generation state, battery is charged, and the current flows to the battery from the motor. The current is proportional to motor torque, so the motor torque should be limited as:

$$u_{n\_\max} \le u_{t+i} \le u_{p\_\max}\,, \tag{12}$$

where u is the motor torque, $u_{n\_\max}$ and $u_{p\_\max}$ is the minimum value and maximum value of the motor torque.

The cost function design in predictive optimization problem should make the tracking error of actually speed and speed set point curve, and considering the control value should not change too intense, the cost function can be expressed as [9]:

$$J = E\{\sum_{t=0}^{p-1}[y_{t+i} - w_{t+i}]^2 + \lambda[u_{t+i} - u_{t-1}]^2\}\,, \tag{13}$$

where λ is the weighting factor.

Therefore, we have a predictive optimization problem modified by constraint expressed as

$$\begin{gathered} \min J = E\{\sum_{t=0}^{p-1}[y_{t+i} - w_{t+i}]^2 + \lambda[u_{t+i} - u_{t-1}]^2\}, \\ s.t \begin{cases} Y = \theta X + \Xi & . \\ u_{n\_\max} \le u_{t+i} \le u_{p\_\max} \\ \Delta y_{n\_\max} \le \Delta y_{t+i} \le \Delta y_{p\_\max}. \end{cases} \end{gathered} \tag{14}$$

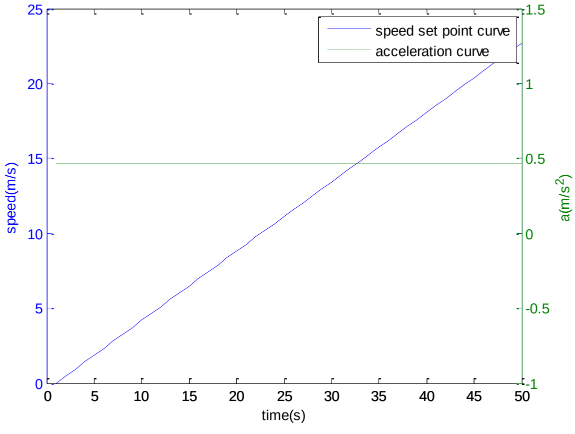
Model predictive control is formulated as a repeated solution of a horizon open loop optimal control problem subject to system dynamics and input and state constraints. Based on measurements obtained at time t, the controller predicts the dynamic behaviour of the system over a prediction horizon t_p in the future according to (9), and determines the input such that a predetermined the cost function is minimized. In the designed algorithm control horizon t_c and prediction horizon t_p is the same [10].

4 Simulation and analysis

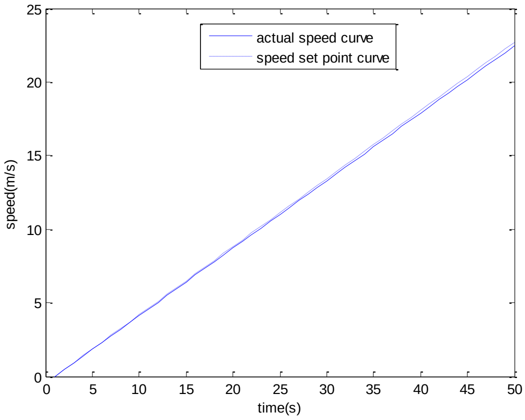
The simulation and verification of the designed algorithm is implemented in the MATLAB/Simulink environment using a variable-step solver that is suited for stiff dynamic systems.

Battery capacity is represented by SOC values and calculated by the MATLAB / Simulink battery model. The weighting factor λ is set to 0.2. All the input, output, controlled variables are normalized to [-1,1].

Time for electric bus accelerate s from start to 50km/h is about 30 seconds, so the acceleration of $0.463m/s^2$. When the accelerator pedal value α is set to maximum, acceleration value a is $0.463m/s^2$, assume the current vehicle speed v_0 is zero, the speed set point curve is shown in Figure 3(a). Actual speed and the set speed is shown in Figure 3 (b).



(a)

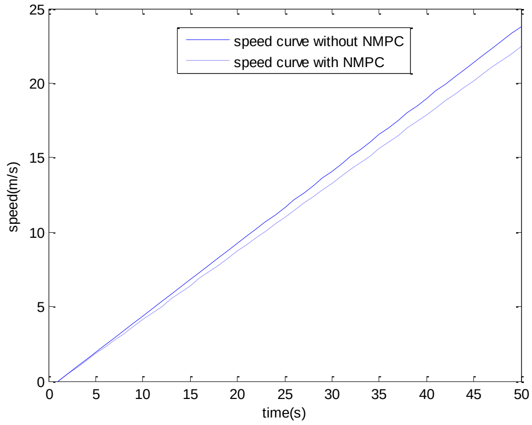


(b)

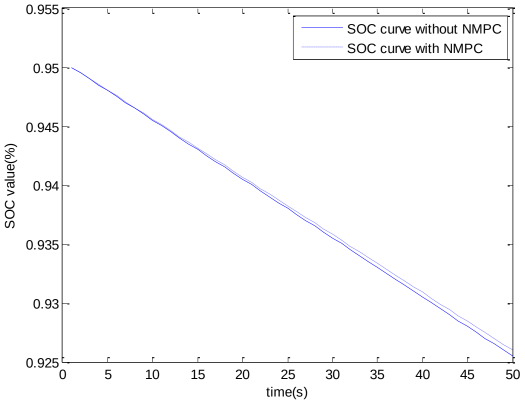
FIGURE 3 The simulation results :a) speed set point curve .b) actual speed curve and the set speed curve

In Figure 3(b), the simulation actual speed curve and the set speed curve is shown as solid line and dot-dash line, and it can be seen that the actual speed following the set point well.

Figure 4(a) shows when no control algorithm is implemented; the speed is a little higher than the speed when control algorithm is implemented. Figure 4(b) shows the SOC of the battery in the two conditions, the initial SOC value is set to 95% from figure 4(b) it can be seen that the designed algorithm reducing the change of SOC value.



(a)



(b)

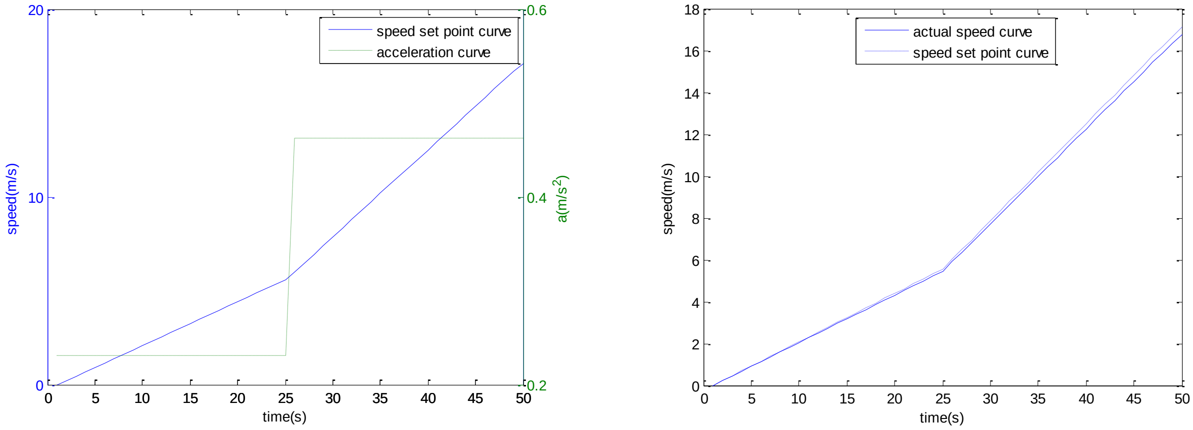
FIGURE 4 The simulation results :a) speed curve comparison .b) SOC curve comparison

When the accelerator pedal value α is set to half of the maximum for several time before change to the maximum, the initial vehicle speed is zero, the speed set point curve is shown in Figure 5(a). Actual speed and the set speed is shown in Figure 5(b), the simulation actual speed curve and the set speed curve is shown as solid line and dot-dash line.

In Figure 5(b), when accelerator pedal changes, the actual speed still following the set point well.

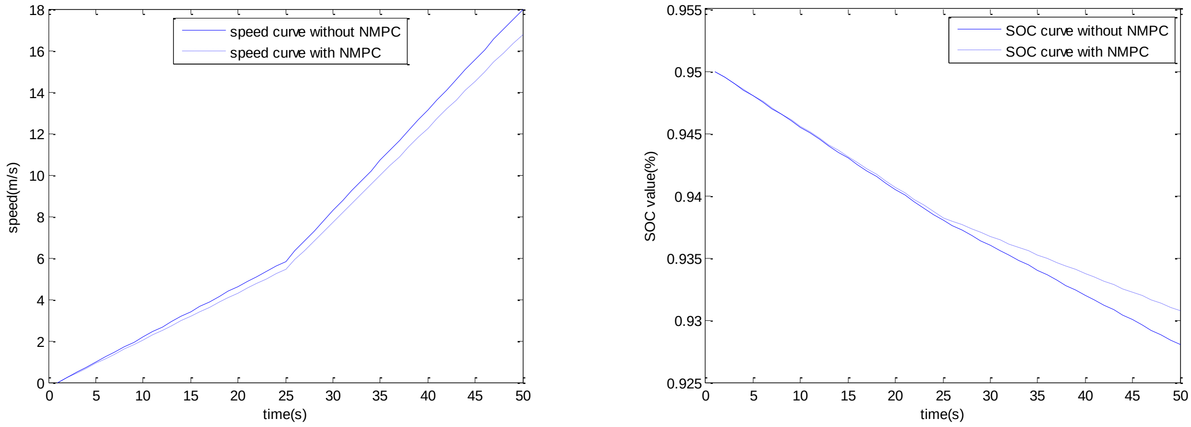
Figure 6(a) shows when no control algorithm is implemented; the speed is a little higher than the speed when control algorithm is implemented. Figure 4(b) shows the SOC of the battery in the two conditions, the initial SOC value is set to 95% from Figure 4(b) it can be seen that the designed algorithm reducing the change of SOC value.

Figure 6(a) shows when no control algorithm is implemented; the speed is changed rapidly than the speed when control algorithm is implemented. Figure 4(b) shows the SOC of the battery in the two conditions, the initial SOC value is set to 95% from figure 4(b) it can be seen that the designed algorithm reducing the change of SOC value significantly.



(a) (b)

FIGURE 5 The simulation results :a) speed set point curve .b) actual speed curve and the set speed curve



(a) (b)

FIGURE 6 The simulation results: a) speed curve comparison .b) SOC curve comparison

5 Conclusions

This paper develops a control-oriented drivability model for an electric vehicle and a nonlinear model predictive optimization algorithm for an electric vehicle.

According to the structure of the mechanism model, an input - output model is developed. As the input - output model is nonlinear; it can be described as Non-linear Auto Regressive with eXogenous inputs (NARX) model.

Then a nonlinear model predictive optimization algorithm is implemented with a cost function which is developed that considers the tracking error of setting value and the variation of control volume, and longitudinal ride comfort and energy-saving is also considered.

Simulations show that the developed optimization algorithm is energy-saving and improves the ride comfort.

Since the algorithm only considers the situation when the vehicle is driven on a straight flat road, therefore only in the direction of the longitudinal movement has been optimized. Future work will focus on the effect of lateral movement and the safety factor on the developed optimization algorithm.

Acknowledgments

The authors would like to thank Anhui Ankai Automobile Co., Ltd and National Electric Vehicle System Integration Engineering Research Centre.

References

[1] Chengqun Li, Xing Sun, Xiaolei Dong 2010 *Advanced Materials Research* **139-141** 2243-6

[2] Wang Guo-Quan, Yang Wen-Tong, Xu Xian-Feng, Yu Qun 2003 *Journal of Shanghai Jiaotong University* **37**(11) 1772-5 *(in Chinese)*

[3] Li S G, Sharkh S M, Walsh F C, Zhang C N 2011 *IEEE Transactions on Vehicular Technology* **60**(8) 3571-85

[4] Kraus T, Ferreau H J, Kayacan E, Ramon H, De Baerdemaeker J, Diehl M, Saeys W 2013 *Computers and Electronics in Agriculture* **98** 25-33

[5] Koprubasi K 2008 *Modelling and control of a hybrid-electric vehicle for drivability and fuel economy improvements* PhD Thesis The Ohio State University

[6] Parra O J S, Martinez N O D, Puente F J 2010 *NISS2010 4th International Conference on New Trends in Information Science and Service Science* 754-9

[7] He Lin, Li Liang, Yu Liangyao, Mao Enrong, Song Jian 2012 *Proceedings of the Institution of Mechanical Engineers Part D: Journal of Automobile Engineering* **226**(8) 1016-25

[8] Chen Bo-Chiuan, Wu Yuh-Yih, Tsai Hsien-Chi 2014 *Applied Energy* **113** 1764-74

[9] Li Kang, Peng Jian-Xun, Irwin G W 2005 *IEEE transactions on automatic control* **50**(8) 1211-6

[10] Zhang Ridong, Xue Anke, Wang Shu-Qing 2011 *Industrial and Engineering Chemistry Research* **50**(13) 8110-21

Authors

	Dou Manli, born in 1988,Anhui, China **Current position, grades**: Ph.D.student **University studies:** University of Science and Technology of China **Scientific interest:** automotive electronics, advanced control and optimization **Publications:** 5 papers published in international journals and conferences **Experience:** Engaged in automotive electronics and advanced control and optimization of theoretical and applied research. Participated in several research and development projects. Several papers published in international journals and conferences
	Shi Chun, born in 1980, Jiangsu, China **Current position, grades:** lecturer, doctor **University studies:** University of Science and Technology of China **Scientific interest:** automotive electronics, advanced control and optimization **Publications** : more than 10 papers published in international journals and conferences **Experience:** Engaged in automotive electronics and advanced control and optimization of theoretical and applied research, participated in more than ten research and development projects. Several papers published in international journals and conferences
	Wu Gang, born in 1964, Jiangsu, China **Current position, grades:** Professor, Ph.D. supervisor **University studies:** University of Science and Technology of China **Scientific interest:** process control, advanced control and optimization **Publications :** more than 100 papers published in international journals and conferences **Experience:** Engaged in industrial process control and advanced control and optimization of theoretical and applied research, including predictive control, adaptive control, system identification, constraint control, intelligent control, on-line operation optimization, Participated in more than ten research and development projects, won four ministerial-level prize and the first global Chinese community Intelligent Control and Intelligent Automation Conference Best Paper Award, Papers published in international journals and conference papers nearly one hundred.
	Liu Xiaoguang, born in 1988, Anhui, China **Current position, grades:** engineer, master **University studies:** University of Science and Technology of China **Scientific interest:** automotive electronics, advanced control, precision optics **Publications** : 5 papers published in international journals and conferences **Experience:** Engaged in automotive electronics and advanced control and optimization of theoretical and applied research, participated in several research and development projects. Several papers published in international journals and conferences.

Force-fight problem in control of aileron's plane

Ying Zhang*, **Zhaohui Yuan**

School of automation, Northwestern Polytechnical University, Xi'an, China

Received 1March 2014, www.tsi.lv

Abstract

In order to reduce or eliminate the force-fight phenomenon of single feedback loop of redundant-channels, modelling the whole control system on the basis of analysing the structure of aileron, the correctness of the model is verified by experiments. Simulation results show that set the dead band of the valve which control the feedback loop smaller is conducive to the decrease of system's fighting-force; for every reduce in the difference of the two valves' overlap of 0.01mm, the fighting-force decreases one time; when the driving speed is more than 50mm/s, system abstains smaller fighting-force. Therefore, the optimization of structure parameters can reduce fighting-force effectively. When the parameters of valves and driving speed is restricted, another method of using a bypass orifice to connect the two cavities of the cylinder is proposed to solve the problem, simulation results shows that fighting-force reduce 2000N for every increase in the orifice's diameter of 0.1mm when using the fixed orifice, and using the variable orifice can abstain a small fighting-force and meanwhile reduce the wastage of hydraulic oil.

Keywords: single feedback loop, force-fight phenomenon, difference of the two valves' overlap, driving speed of motor, bypass orifice

1 Introduction

Cross-linking and interference problem between channels of redundant steering gears that is multiple steering gears driving a comprehensive shaft which is used on aircrafts and space vehicles is called force-fight phenomenon [1, 2]. The redundant steering gear is the key component of a flight control system, in order to improve its reliability and security, the main control surface usually adopts parallel actuators. Due to the error accumulation during the process of manufacture and installation of valves and actuators, the displacement of each actuator is usually different, coupled with the large torsional stiffness of the control shaft there will be force-fight between the actuators which is easy to cause fatigue and failure of the structure [3-5], therefore the flight control system needs to take effective measures to reduce or eliminate this phenomenon. Boeing-777 adopts the differential pressure transducer to transmit fighting-force to the instruction set to equilibrate the force to reduce force-fight [6]. In F/A-18 the two channels share a main control valve, this kind of design can ensure the synchronism of the two actuators by controlling the precision of the spool [7]. In Su30 a throttle valve is used to connect the two cavities to reduce force-fight [8]. Usually the actuators are designed to work independently and each has a feedback channel in common control structure of rudder surface, as long as the consistency of all the actuators is controlled within a small range the fighting-force will not be very large [9-13], but for the certain control structure studied in this paper, the two parallel branches share only one feedback channel to simplify the structure, the fighting-force will be very large and it is easy to cause structural damage if one control valve is shut off but another one is still open [14-20]. On the bases of analysing the characteristics of the single feedback loop control structure, using MATLAB/SIMULINK to model and simulate the whole system precisely incorporating nonlinearities and the pressure loss of pipelines, the accuracy of the model is verified by experiments and several factors which have great influence on the force-fight are analysed. A method of adopting a bypass orifice is used to reduce the fighting-force and the selection of the orifice's diameter is analysed. To reduce oil consumption, the variable orifice is designed, simulation results show that this method can effectively reduce the fighting-force and the wastage of oil.

2 Structure and working principle

The schematic diagram of the system is shown in Figure 1. This system mainly consists of driving motor, mechanical transmission part, overlap valves, actuator cylinders, pipelines and a shaft. Driving motor is used to control the movement of the joystick. The two throttles of the overlap valve are connected to the cavities of the actuator cylinder through the pipelines. The head port of actuator is hinged-supported on the airframe and the piston rod of rod port is hinged-supported on the shaft.

During of flight the aileron deflects with the pull and push of the joystick controlled by the driving motor. When the joystick is pushed forward, the two plate valves open clockwise and the high-pressure oil flows into the left cavity of cylinder which will make the rod push the shaft rotates clockwise. This is a single feedback loop system, since the feedback channel is connected to one

* ***Corresponding author*** e-mail: zhangying_1102@163.com

end of the shaft next to the piston rod 1 so the feedback signal is decided by the torque on the shaft produced by piston rod 1 and has nothing to do with piston rod 2.

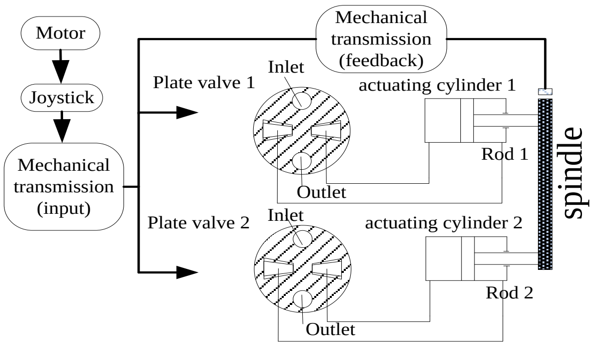


FIGURE 1 Schematic diagram

The feedback angle will make the plate valves rotate anticlockwise until both of them are closed. The difference of the two actuators' displacements will produce torsional moment on the shaft, and in turn, there will be counterforce on the two piston rods equal and opposite. The system balance will be reached until the differential pressure of cylinder's two cavities and the force on the piston rod are equal. The system's fighting-force is defined as half of the difference of the reaction force on each piston rod caused by the torsion of shaft.

3 Mathematical model

3.1 FLOW EQUATION OF PLATE VALVE

The structure of the overlap valve is shown in Figure 2.

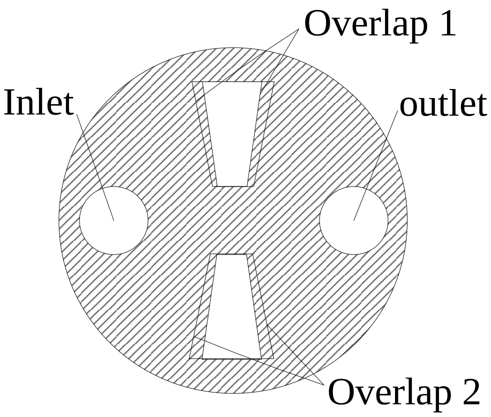


FIGURE 2 Overlap valve

The overlap valve consists of two discs, the one at bottom is show in Figure 2 which consists of an inlet and an outlet (mutual isolated) and two trapezoidal throttles, another disc covers on it which has two larger trapezoidal covers, the difference of the areas of the covers and throttles are called the overlap. The overlap of the plate valve is set to f_0, f is the open area of the throttle corresponding to the rotation of valve. The input and output flow rates of the two throttles are calculated as:

$$\begin{cases} Q_{11} = c_V (f - f_0)\sqrt{2(p_s - p_1)/\rho} \\ Q_{12} = c_V (f - f_0)\sqrt{2(p_2 - p_0)/\rho} \end{cases} (f > f_0) \quad , \qquad \begin{cases} Q_{11} = 0 \\ Q_{12} = 0 \end{cases} (f < f_0) \tag{1}$$

where Q_{11},Q_{12} is the input and output flow rate of plate valve respectively, Cv is the coefficient of flow, P_1 and P_2 is the output and input pressure respectively, ρ is the density of oil, P_0 is the return pressure, P_s is the source pressure.

3.2 CONTINUITY EQUATION OF ACTUATOR

The actuator is asymmetric, so the effective area of the two cavities is not the same, the actuator works in two ways as show in Figure 3.

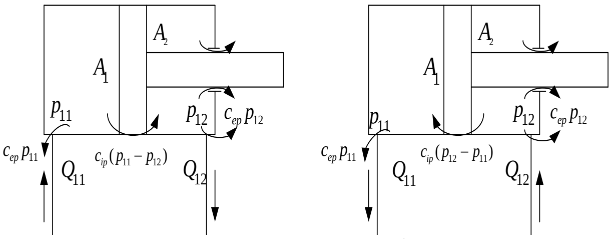


FIGURE 3 Working diagram of actuator

The continuity equation of actuator is presented as:

$$\begin{cases} Q_{11} = A_1 \dfrac{dx_{t1}}{dt} + \dfrac{(x_{01} + x_{t1})A_1}{E_y}\dfrac{dp_{11}}{dt} + C_{ip}(p_{11} - p_{12}) + C_{ep}(p_{11} - p_0) \\ Q_{12} = A_2 \dfrac{dx_{t1}}{dt} - \dfrac{(x_{02} - x_{t1})A_2}{E_y}\dfrac{dp_{12}}{dt} + C_{ip}(p_{11} - p_{12}) - C_{ep}(p_{12} - p_0) \end{cases} , \tag{2}$$

where P_{11} is the oil pressure of head port, P_{12} is the oil pressure of rod port, Q_{11} is the flow rate which flow into head port, Q_{12} is the flow rate which flow out from rod port. A_1 is the effective area of head port, A_2 is the effective area of rod port, x_{01} is the initial displacement of head port, x_{02} is the initial displacement of rod port, *Cip* and *Cep* are the internal and external leakage coefficient respectively, Ey is the modulus of elasticity, x_{t1} is displacement of the rod.

In the course of work the head port is hinged-supported, cause the fixed stiffness of cylinder is considered, the connection part may have certain deformation during the dynamic process which means that the cylinder body itself has certain displacement and it is set to x_c, so the rod's displacement relative to cylinder is set to (x_t-x_c), so the continuity equation is expressed as:

$$\begin{cases} Q_{11} = A_1 \dfrac{d(x_{t1} - x_{c1})}{dt} + \dfrac{(x_{01} + x_{t1} - x_{c1})A_1}{E_y}\dfrac{dp_{11}}{dt} + C_{ip}(p_{11} - p_{12}) + C_{ep}(p_{11} - p_0) \\ Q_{12} = A_2 \dfrac{d(x_{t1} - x_{c1})}{dt} - \dfrac{(x_{02} - (x_{t1} - x_{c1}))A_2}{E_y}\dfrac{dp_{12}}{dt} + C_{ip}(p_{11} - p_{12}) - C_{ep}(p_{12} - p_0) \end{cases} . \tag{3}$$

3.3 DYNAMIC EQUILIBRIUM EQUATION

3.3.1 Equilibrium equation of piston rod

The output force of piston rod that is the pressure differential of the two cavities should be balanced with the load, which including the inertia force of rod and load, viscous damping force, friction and elastic force which is caused by the deformation of the piston rod. The dynamic equilibrium equation of piston rod is presented as:

$$p_{11}A_1 - p_{12}A_2 = m_t \frac{d^2 x_{t1}}{dt^2} + B_t \frac{d(x_{t1} - x_{c1})}{dt} + K_{s2}x_{l1} + f_L + F_L \,, \quad (4)$$

where m_t is the equivalent mass of both piston and load, B_t is the viscous damping coefficient of piston and load, K_{S2} is the joint stiffness between piston rod and load, x_{l1} is the deformation of the piston rod, f_L is the frictional resistance of piston rod, F_L is the aerodynamic load.

3.3.2 Equilibrium equation of cylinder

The force applied to the actuator cylinder include the hydraulic pressure of the two cavities, the inertia force of the cylinder, viscous damping force and the force caused by the deformation of the hinged-supported part. The equilibrium equation of cylinder can be expressed as:

$$p_{11}A_1 - p_{12}A_2 = -m_c \frac{d^2 x_{c1}}{dt^2} - B_t \frac{d(x_{c1} - x_{t1})}{dt} - K_{s1}x_{c1} \,, \quad (5)$$

where m_c is the mass of cylinder, K_{s1} is the fixed stiffness where the cylinder is hinged-supported.

3.4 FRICTION LOSS OF PIPELINE

Considering the friction loss of the pipeline from plate valve to cylinder, the relationship of the pressure in the two cavities and the input/output pressure of the plate valve can be expressed as:

$$\begin{cases} p_{11} = p_1 - \Delta p_{11} \\ p_{12} = p_2 + \Delta p_{12} \end{cases} , \quad (6)$$

where Δp_{11} is the friction loss on the way the hydraulic oil flow into the actuator cylinder from plate valve, Δp_{12} is the friction loss on the way the hydraulic oil flow back to plate valve from cylinder.

The friction loss not only include the pressure loss when oil passes through straight pipeline but also include the pressure loss when oil passes through the bended places of pipeline, the two condition have different formulations.

The formulation of the friction loss of straight pipeline is presented as:

$$\Delta p_\lambda = \lambda \frac{l}{d} \frac{\rho v^2}{2} \,, \quad (7)$$

where Δp_λ is the friction loss, λ is the frictional resistance factor, l is the length of the pipeline, d is the diameter of the pipeline, v is the average velocity of flow in pipeline.

This formulation can be not only applied in laminar flow but also applied in turbulent flow, the value of the resistance factor λ is different in the two cases. In the process of calculation, the Reynolds number is real-time monitored to confirm the value of resistance factor λ, which can be expressed as:

$$\begin{cases} \lambda = 75/\mathrm{Re} & (\mathrm{Re} < 2320) \\ \lambda = 0.3164/\mathrm{Re}^{0.25} & (\mathrm{Re} > 2320) \end{cases} . \quad (8)$$

When the pipeline is bended the friction loss should be considered too, for circular tube the pressure loss of the bended place in pipeline is presented as:

$$\Delta p_\xi = \xi \frac{\rho v^2}{2} = \frac{7.878\rho}{\pi^2 d^4} Q^2 \,, \quad (9)$$

Where ΔP_ξ is called local pressure loss, ξ is the local resistance factor, Q is the flow rate when the oil flows through the bended places of pipeline.

3.5 TORTIONAL MOMENT ANALYSIS

The torsional moment of the shaft when the output displacements of the two actuators are different can be calculated as:

$$M_d = |\theta_1 - \theta_2| K_s \,, \quad (10)$$

where M_d is the torsional moment, (θ_1-θ_2) is the torsion angle of the shaft, K_s is the torsional rigidity.

According to the torsional moment calculated above, the counterforce and the deformation of piston rod can be presented as:

$$F = M_d / a = K_{s2} x_l \,, \quad (11)$$

where a is the effective arm, F is the counterforce on the piston rod.

4 Analysis of simulation results

The main simulation parameters are shown in Table 1.

TABLE 1 Main simulation parameters

Oil pressure	**Return pressure**	**Fixed stiffness**	**Joint stiffness**	**Piston quality**
21MPa	**0MPa**	**1e8N/m**	**1e8N/m**	**2kg**
External leakage coefficient	Internal leakage coefficient	Bypass orifice diameter	Oil density	Actuator quality
1e-18	**1.6e-13**	**0.5mm**	**850kg/m^3**	**12kg**
External radius of trapezoidal	Internal radius of trapezoidal	Piston diameter	Rod diameter	Pipe diameter
15mm	**7mm**	**0.1m**	**0.043m**	**10mm**

4.1 ASYMMETRY OF PIPELINE

The asymmetry of pipeline is an obvious factor which will cause the desynchrony of the two channels. To analyse the influence of this factor on force-fight, assume that the two valves have the same dead band and the zero offset is ignored. Set the external pipeline's length of valve 1 and valve 2 to 1.8m and 0.75m respectively. The simulation results of the fighting-force caused by the asymmetry of pipeline are presented in Figure 4 when the input displacement of joystick is ±10mm (step).

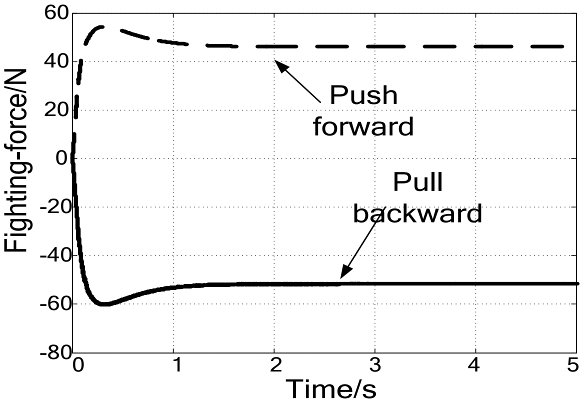


FIGURE 4 Fighting-force of asymmetric pipeline

Simulation results show that, when the difference of the pipeline's length is 1.05m, the fighting-force is very small. So the influence of the asymmetry of pipeline on force-fight can be neglected.

4.2 DIFFERENCE OF DEAD BAND

The overlap valve is widely used in practical applications, but the manufacture error and wear will both lead to the difference of the two valves' dead band inevitably. Without consideration of zero offset, set the difference of two valves' dead band to 0.01~0.04mm, and the dead band of valve 1 is 0.04mm which is smaller than valve 2's. The simulation results of fighting-force are presented in Figure 5.

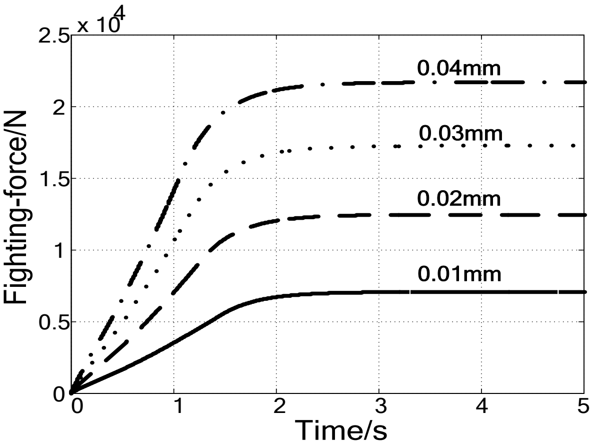


FIGURE 5 Valve 1's dead band is smaller

In Figure 6, valve 2's dead band is set to 0.04mm which is smaller than valve 1's and other conditions is the same as before.

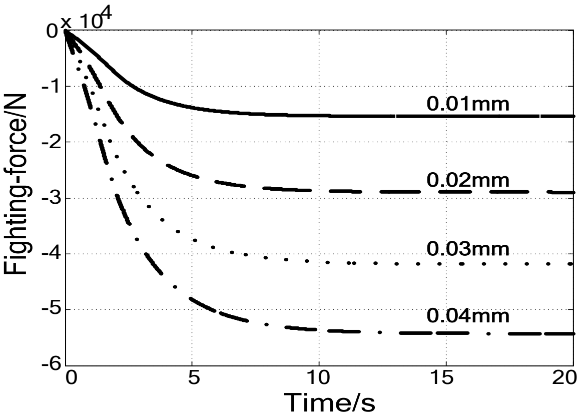


FIGURE 6 Valve 1's dead band is larger

List the data of fighting-force in the two cases above in Table 2.

TABLE 2 Fighting-force changes with the difference of valve's dead band

Difference of dead band/mm (valve 1 is smaller)	**0.01**	**0.02**	**0.03**	**0.04**
Fighting-force/N	8430	14800	20400	25600
Difference of dead band/mm (valve 1 is larger)	0.01	0.02	0.03	0.04
Fighting-force /N	15460	29030	41910	54350

As show in Table 2 when the dead band of valve 1 is smaller, for every increase in the difference of the two valves' dead band of 0.01mm the fighting-force increase around 5000N; when the dead band of valve 1 is larger for every increase in the difference of the two valves' dead band of 0.01mm the fighting-force increase around 12000N.

The fighting-force is obviously different in the two cases because the feedback channel is controlled by valve 1. In the former case, valve 1 is always open until both of the valves are closed so the feedback process is very quick; In the latter case after valve 1 is closed, rod 2 drags rod 1 to produce feedback signal until the two valves are closed so the feedback process is very slow that is why the fighting-force is very large .

4.3 SPEED OF DRIVING MOTOR

For the joystick is controlled by motor, the input displacement signal is impossible to achieve the given value immediately, so it is necessary to analysis the influence of the driving speed on system's force-fight.

The zero offset is ignored and the dead band of valve 1 and valve 2 is 0.04mm and 0.052mm respectively, the joystick is pushed forward by 10mm (step). The simulation results are presented in Figure 7 when the driving speed is set to 10mm/s, 20mm/s, 50mm/s and 100mm/s.

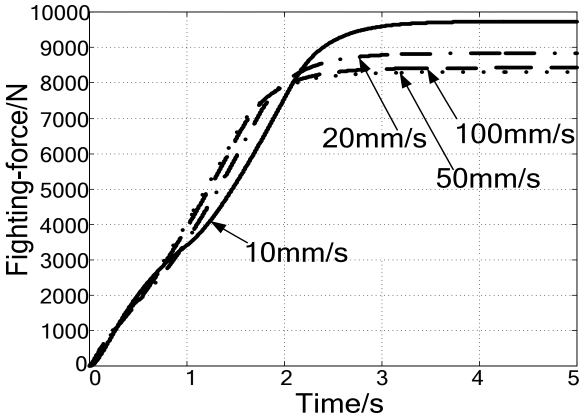


FIGURE 7 Fighting-force of different driving speed

According to the simulation results the fighting-force decrease with the increase of the driving speed but when the speed is more than 50mm/s, the fighting-force remain basically unchanged.

4.4 ADOPT BYPASS ORIFICE

When the parameters of valves and driving speed is restricted, adopting the bypass orifice to connect the two cavities of the actuator cylinder to reduce fighting-force.

4.4.1 Adopt fixed orifice

Adopt a fixed orifice whose diameter is 0.05mm and length is 8mm to connect the two cavities. When there is no aerodynamic load the dead band of the two valves is 0.04mm and 0.052mm respectively, the simulation results are presented in Figure 8 when the input displacement is 10mm (step).

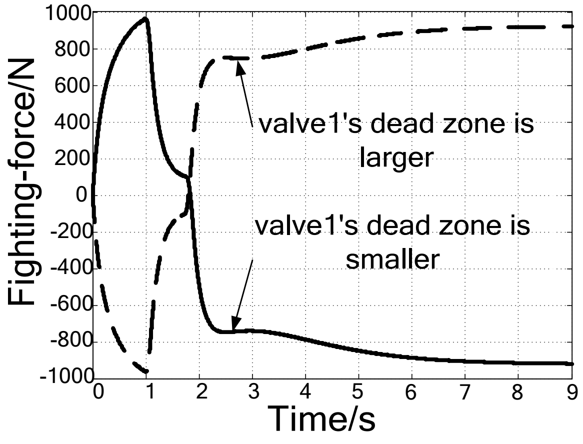


FIGURE 8 Fighting-force when adopt fixed orifice

Simulation results show that in the two cases the fighting-force are both very small, and the fighting-force in the two cases are basically the same, that is because the two valves are always open to supply the flow rate when the bypass orifice is adopted.

The influence of the diameter of the orifice is analyzed through simulation below. The valve 1's dead band is 0.04mm and the valve 2's dead band is 0.052mm, and the diameter of bypass orifice is set to 0.2mm~0.5mm. The simulation results are presented in Figure 9 when there is no aerodynamic load.

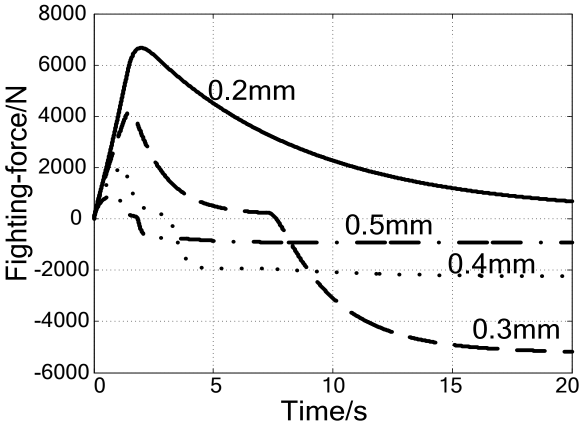


FIGURE 9 Different diameter of bypass orifice

The simulation results show that, the fighting-force decrease obviously with the increase of the diameter. The fighting-force reduce 2000N for every increase in the diameter of 0.1mm, and the stabilizing process is slow when the diameter is small.

4.4.2 Adopt variable orifice

To reduce the oil consumption, adopt the variable orifice, the area of the variable orifice can be obtained as:

$$f = \pi r^2 (1 - \frac{\Delta p}{p_i}), \tag{12}$$

where P_i is the threshold differential pressure.

With the increase of ΔP the area of the orifice will decrease, and when $\Delta P \geqq P_i$ the orifice closes. This design is to reduce ΔP of the two cavities during the initial procedure to reduce the fighting-force, and keep a certain differential pressure during the stabilizing process to guarantee the response speed of system.

The largest diameter of the orifice is 0.5mm, and its length is 8mm, the threshold differential pressure is 5MPa, the dead band of the two valves is 0.04mm and 0.052mm respectively. Simulation results of fighting-force are presented in Figure 10.

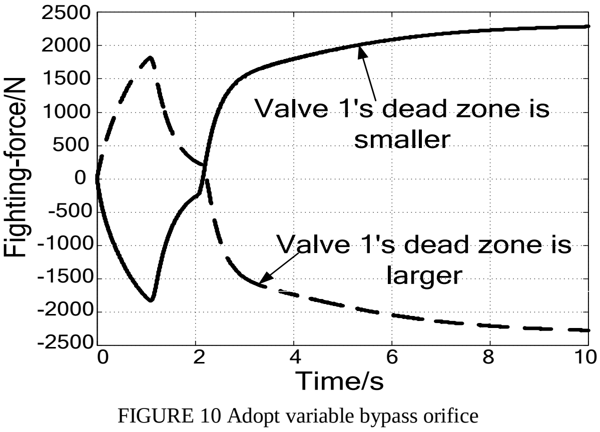


FIGURE 10 Adopt variable bypass orifice

Simulation results show that the fighting-force is around 2400N, which is a little bigger than adopting the fixed orifice, but is relatively small. For the area of the orifice is variable the wastage of hydraulic oil is reduced and the stabilizing process is accelerated.

5 Experimental verification

In the test, the input displacement signal is the sinusoidal waves of frequency 0.05Hz~0.9Hz and amplitude 5mm, the dead band of valve 1 and valve 2 is 0.04mm and 0.052mm respectively and there is no aerodynamic load. In the case of without bypass orifice, the test and simulation results are listed in Table 3.

TABLE 3 Simulation results contract with test results

Frequency/Hz	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**
FF of simulation/KN	20.9	11.2	5.72	3.88	2.95
FF of test/KN	21.3	11.7	6.1	4.09	3.17
Frequency/Hz	0.5	0.6	0.7	0.8	0.9
FF of simulation/KN	2.38	2	1.73	1.52	1.36
FF of test/KN	2.52	2.23	1.97	1.78	1.54

Simulation and test results of adopting bypass orifice of 8mm diameter and 0.5mm length are listed in Table 4.

TABLE 4 Simulation results contract with test results when adopt bypass orifice

Frequency/Hz	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**
FF of simulation/KN	1.51	1.35	1.01	0.1	0.97
FF of test/KN	1.83	1.62	1.13	1.08	1.35
Frequency/Hz	0.5	0.6	0.7	0.8	0.9
FF of simulation/KN	0.94	0.916	0.891	0.866	0.84
FF of test/KN	1.42	1.16	1.10	0.99	0.90

The data show that the simulation error remains below 8%, which means this model can be used as an accurate simulation platform to solve the force-fight problem of aileron. For the inconsistent of the clearance in mechanical transmission is ignored in the model, the simulation results are always smaller. Simulation and test results show that when the frequency of input signal is low the method of adopting the bypass orifice can reduce the fighting-force greatly, with the increase of the frequency the adopting of bypass-orifice has relatively small impact on force-fight. When the frequency reaches 1Hz the fighting-force reduce to 1000N without any measurement.

6 Conclusions

(1) The asymmetry of pipelines has little impact on the force-fight, so it is unnecessary to consider this factor in the pipeline layout;

(2) The difference of two valves dead band has great influence on force-fight, so it is important to reduce this difference. Set the dead band of the valve which control the feedback loop smaller is conducive to the decrease of system's fighting-force;

(3) Fighting-force will be great if the driving speed is too small, so this speed is recommended to be more than 50mm/s;

(4) Adopting bypass orifice can reduce the fighting-force greatly and the fighting-force will decrease with the increase of the orifice's diameter;

(5) Adopting the variable orifice can reduce the wastage of hydraulic oil at the same time of obtaining a small fighting-force.

Acknowledgments

I would like to extend my sincere gratitude to my tutor, for his instructive advice and useful suggestions on my paper. Special thanks should go to my friends who have put considerable time and effort into their comments on the draft. This work was financially supported by National Natural Science Foundation of China (61374032/2013).

References

[1] Wang Zhanlin, Qiu Lihua, Zhang Xiaosha, et al 1991 Research on the synchronous drive of the redundant channels for hydraulic actuator *Journal of Astronautics* **12**(2) 9-14

[2] Annaz F Y 2005 *Smart Materials and Structures* **14**(6) 1227-38

[3] Maré J-C 2007 Comparison of Hydraulically and Electrically Powered Actuators with reference to Aerospace Applications. *Hydraulics and Pneumatics Conference*

[4] Cochoy O, Hanke S, Carl U B 2007 *Aerospace Science and Technology* **11**(2) 194-201

[5] O'brien M 1998 *Aircraft Engineering and Aerospace Technology* **70**(5) 364-6

[6] van Den Bossche D The A380 flight control electrohydrostatic actuators achievements and lessons learnt. 25th international congress of the aeronautical sciences, Hamburg, Germany, 3–8 September 2006 *session 7.4.1* 1–8

[7] Straub H H, Creswell R 1991 *Aerospace Technology Conference and Exposition* Long Beach: CAUnited States 31-40

[8] Wang Zhengjie, Guo Shijun, Li Wei 2008 *WSEAS Transactions on Systems and Control* **3**(10) 869-78

[9] Nagai Kiyoshi, Dake Yuichiro, Shiigi Yasuto, et al 2010 Design of redundant drive joints with double actuation using springs in the second actuator to avoid excessive active torques *2010 IEEE International Conference on Robotics and Automation (ICRA)* Anchorage, AK 805-12

[10] Zhang J, Gao F, Yu H, et al 2011 *Journal of Mechanical Engineering Science* **225**(12) 3031-44

[11] Patra K C 2011 *Engineering fluid mechanics and hydraulic machines* Oxford UK: Alpha Science International

[12] Karam W, Mare J-C 2009 *Aerospace Eng Aircraft Technol* **81**(4) 288–98

[13] Attar B 2008 *Realistic modelling in extreme conditions of electro-hydraulic servovalve used for aerospace guidance and navigation* France: PhD Thesis, INSA Toulouse

[14] Fu Y L, Liu H S, Pang Y, et al 2010 Force fighting research of dual redundant hydraulic actuation system. 2010 *international conference on intelligent system design and engineering application, Beijing, China, 13–14 October 2010* China: Beihang University 762–6

[15] Jacazio G, Gastaldi L 2008 Equalization techniques for dual redundant electro hydraulic servo actuators for flight control systems. *Bath/ASME Symposium on Fluid Power and Motion Control (FPMC 2008)* Bath, 2008 543-57

[16] Qi H T, Mare J-C, Fu Y L 2009 Force equalization in hybrid actuation systems *7th international conference on fluid power transmission and control, Hangzhou,China, 7–10 Apr 2009* China: Zhejiang University 342–7

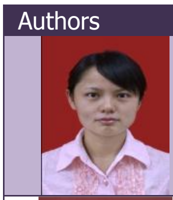
[17] Cheng Hui, Yiu Yiu-Kuen, Li Zhang 2003 *IEEE/ASME Transactions on Mechatronics* **8**(4) 483-91

[18] Mare J-C, Moulaire P 2001 The decoupling of position controlled electrohydraulic actuators mounted in tandem or in serie *Seventh Scandinavian International Conference on Fluid Power* Linkoping 93-108

[19] Steffen T, Davies J, Dixon R, Goodall R M 2007 Using a Series of Moving Coils as a High Redundancy Actuator *Preprint of the IFAC Conference for Advanced Intelligent Mechatronics* (AIM) Zurich

[20] Ifrim A M 2013 *Electrotehnica, Electronica, Automatica* **61**(2) 79-84

Authors



Ying Zhang, born on November 2, 1984, Hubei, China

Current position, grades: PhD studies in Detection Technology and Automatic Equipment
University study: M.Sc. in Mechanical Engineering from Northwestern Polytechnical University.
Research interests: servo control and design of hydraulic system

Zhaohui Yuan, born in March, 1964, Anhui, China

Current position, grades: full Professor in School of automation, Northwestern Polytechnical University
University study: M.Sc. in Transmission and Control of fluid and PhD in Detection Technology and Automatic Equipment
Research interests: different aspects of servo control systems

Gene selection for cancer classification using the combination of SVM-RFE and GA

Xiaobo Li[*]

Department of Computer Science and Technology, College of Engineering, Lishui University, Lishui 323000, China

Received 1 March 2014, www.tsi.lv

Abstract

Gene selection is a key research issue in molecular cancer classification and identification of cancer biomarkers using microarray data. Support vector machine recursive feature elimination (SVM-RFE) is a well known algorithm for this purpose. In this study, a novel gene selection algorithm is proposed to enhance the SVM-RFE method. The proposed approach is designed to use the combination of SVM-RFE and genetic algorithm (GA). The performance of the proposed model is validated on a binary and a multi-category microarray gene expression datasets. The results show that the proposed gene selection method is able to elevate the performance of SVM-RFE, which extracts much less number of informative genes and achieves highest classification accuracy.

Keywords: cancer classification, gene selection, support vector machine recursive feature elimination (SVM-RFE), genetic algorithm (GA), microarray data

1 Introduction

Recent advances in cancer genomic would provide opportunities for personalized cancer medicine [1]. Cancer is a systemic and complex disease with highly heterogeneity, which remains a key obstacle for accurate diagnosis and treatment of cancers. There exist different pathways between patients with tumours, and it is prone to over-treatment or ineffective treatment of the patients if the same type of treatment is used to treat a certain type of cancer. One typical example is the anti-cancer drug Trastuzumab, which is an antibody that interferes with the human epidermal growth factor (HER2) receptor, and is only effective in patients that HER2 is over-expressed [2]. At the same time, personalized cancer medicine underlines the need for molecular classification of tumours and identification of reliable tumour biomarkers to predict tumour subtypes.

Nowadays high throughput technologies such as microarray technology allow for monitoring of thousands of gene expression values simultaneously, and have been successfully conducted in molecular classification of tumours and identification of tumour biomarkers [3]. However, the sample size of microarray data is typically small (less than 100), and the number of genes is large (generally more than 10,000). The key issue that needs to be addressed is to select a smaller number of informative genes from the thousands of genes measured in microarray, which are subsequently used to accurately classify tumour samples [4, 5].

Support vector machine recursive feature elimination (SVM-RFE) is a well known algorithm for this purpose proposed by Guyon et al. [6]. SVM-RFE algorithm has recently attracted many researchers since it can obtain satisfactory results with microarray gene expression data [5, 7-11]. This paper proposes a novel SVM-RFE based gene selection algorithm by combining a genetic algorithm (GA) with SVM-RFE criteria. The proposed method, which is referred to as SVM-RFE/GA, can be divided into two stages: in the first stage, a ranking list for the original gene set is yielded by SVM-RFE criteria, and top n ranking genes are retained as “candidate gene set”; in the second stage, a genetic algorithm is applied on the candidate gene set, in order to search an optimal minimum gene set. Experimental results demonstrate the feasibility and effectiveness of the proposed method. Section 2 describes SVM-RFE and GA methods, and gives a detailed description of the SVM-RFE/GA model. Section 3 demonstrates the experimental results. Section 4 analyses and discusses the results. Section 5 concludes this work.

2 Method

2.1 SVM-RFE

Support vector machine (SVM) is a superior classification model for sparse classification problems such as microarray gene expression data. Due to the high dimensionality of feature space in microarray data, linear SVM is adopted in this work. For a liner SVM, the margin width is defined as:

$$w = \sum_{i=1}^{n} \alpha_i c_i x_i \ , \tag{1}$$

[*] ***Corresponding author*** e-mail: oboaixil@126.com

$$m\arg in\ width = 2/\|w\|, \tag{2}$$

where n is the number of support vectors.

SVM-RFE is a type of embedded gene selection method [4]. SVM-RFE is a backward elimination procedure, which iteratively removes each feature, which is of the least importance to the SVM classifier. The objective function J in SVM-RFE is:

$$J = (1/2)\|w\|^2. \tag{3}$$

The Optimal Brain Damage (OBD) algorithm [12] approximates the change in *J* caused by removing each feature by expanding *J* in Taylor series to second order:

$$\Delta J(i) = \frac{\partial J}{\partial w_i}\Delta w_i + \frac{\partial^2 J}{\partial w_i^2}(\Delta w_i)^2. \tag{4}$$

The first order can be neglected at the optimum of *J*, and the second order becomes

$$\Delta J(i) = (\Delta w_i)^2. \tag{5}$$

The change in weight $\Delta w_i = w_i$ correlates with removing i th feature from the classifier, so $(w_i)^2$ is used as the ranking criterion in SVM-RFE. The feature with the smallest $(w_i)^2$ is eliminated since it has the smallest effect on classifier.

The detail of SVM-RFE algorithm is described as follows:

```
Inputs: initial gene set $I = \{1;2;...n\}$, ranked gene set $O = \{\ \}$.
While ( $I$ is not null)
Train the linear SVM classifier
Calculate the ranking criteria $r_i = (w_i)^2$ for all genes in $I$.
Choose the gene with the smallest ranking score: $g = \arg\min\{r_i\}$
Update ranked gene set $O$ and $I$ : $O = O \cup g$ , $I = I - g$
End While
Output: ranked gene set $O$
```

2.2 GENETIC ALGORITHM

Genetic algorithms (GA) [13-15] is a type of wrapper gene selection method, which is based on the principle of natural selection and genetics. GA is a globally adaptive probabilistic search algorithm, drawing on the biological mechanisms of fittest evolution and natural selection, and the genetic mechanisms of recombination and mutation. GA starts from an initial solution of randomly generated population, and the population contains a certain number of encoded individuals. Based on the principle of survival of the fittest, the evolution of each generation would produce more and better approximate solutions. In each generation, each individual is evaluated by the fitness function in the solution domain. The more fit individuals are retained and then modified with genetic operators of crossover and mutation, producing a new population representative of the new solution sets. This process loop is executed until a predetermined termination condition has been reached.

The main components of our GA are described as follows.

Representation of individual. Each individual is encoded by a N -bit binary vector, where N is the size of genetic space. The bit “1” represents a selected gene, and the bit “0” means the opposite.

Fitness Function. Each individual is evaluated by a support vector machines (SVM) classifier, i.e., SMO classifier in WEKA [16].GA is designed to minimize the classification error rate.

Genetic Operators. The genetic operations are performed by Roulette wheel selection, single-point crossover, and bit flip mutation.

2.3 THE SVM-RFE/GA MODEL

Among the thousands of genes detected by microarray technology, there exist four categories of genes for cancer classification [10]: (1) informative genes, which are important for cancer classification and may play a significant role in tumour development; (2) redundant genes, which may be related with cancer and function similarly to informative genes but they are not so significant for cancer classification; (3) irrelevant genes, which have no influence on cancer classification and are irrelevant to cancer; and (4) noisy genes, which have negative effects and their existence may decrease cancer classification performance. The gene selection methods are developed to obtain the first class while removing the next three classes of genes.

SVM-RFE eliminates “worst” gene at each step, generating a ranking for the genes based on their “importance” to the classifier. SVM-RFE has achieved an outstanding performance in cancer classification. However, it ignores the interaction between the genes. A two-stage strategy is proposed to overcome this deficiency. In the first stage, SVM-RFE is applied on the initial gene sets to generate ranking for the genes, and top n ranking genes are kept as “candidate gene set”. The first stage is considered as a prefiltering process, which is designed to remove redundant, irrelevant and noisy genes while retaining informative genes. In the second stage, since the genes in the candidate gene set may highly correlate with each other, a genetic algorithm is utilized to search an optimal minimum gene set in the solution space.

Based on the spirit of Structural Risk Minimization [6], nested gene subsets are defined by the ranking algorithm, and it is possible to select best gene subset by changing the parameter of n: the number of genes. In more detail, the parameter n is varied to select top n ranking genes, generating m incrementally nested gene subsets: $GS_1 \subset GS_2 \subset ... GS_m$. A genetic algorithm is further applied to search the optimal minimum gene subset from the given input space.

3 Experimental Results

3.1 DATA SET

The performance of the SVM-RFE/GA model is validated on both binary and multi-category microarray gene expression datasets. Table 1 summarizes the number of classes, the number of genes, the number of samples and the reference in each dataset.

TABLE 1 The two-class and multi-class gene expression datasets

Dataset	Platform	No. of Classes	No. of Genes	No. of Samples	Reference
Prostate	Affy U95Av2	2	12600	102	[17]
NCI60	Affy Hu6800	9	7129	60	[18]

The prostate dataset is a two-class gene expression dataset, which contains 52 tumor and 50 normal of prostate cancer samples, and it can be obtained from (http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi). The NCI60 dataset is a multi-class gene expression dataset, which can be downloaded from (http://www.broadinstitute.org/mpr/NCI60/). The data set contains 60 samples in 9 tumor types, and the two samples of prostate cancer were excluded from this study, since two samples are not enough for the classification issue.

3.2 EXPERIMENTAL PLATFORM

The experiments were conducted on the WEKA [16] (http://www.cs.waikato.ac.nz/ml/weka/) platform. The SMO classifier was used to execute the classification task, and the polynomial kernel function (PolyKernel) was chosen. The penalty parameter C of the classifier was set to 100. The performance of the SMO classifier was evaluated based on 10-fold cross-validation. The parameters of GA were set as follows: crossover probability = 1, mutation probability = 0.02, maximum generations = 50, and population size = 30.

Pre-processing procedure was performed on the experimental data: the housekeeping genes were removed, with 12,533 gene expression values remained in the prostate dataset and 7,071 gene expression values remained in the NCI60 dataset; The gene expression values were standardized to have a mean of 0 and a standard deviation of 1.

3.3 EXPERIMENTAL RESULTS

In the first stage, SVM-RFE algorithm generates ranked gene set, in which genes rank in descending order. Generally, a smaller subset of 50-100 genes is kept as informative genes in previous study [5]. Here a subset of top 100 ranking genes was retained. To test the performance of SVM-RFE algorithm, the number of genes was reduced from 100 to 1, and the gene with the lowest rank score was eliminated at each step. The performance of the classifiers was assessed using 10-fold cross-validation method. The classifier achieved 100% prediction accuracy with initial 100 genes in both datasets. As shown in Figure 1, in both datasets, the prediction accuracy maintains the highest accuracy when the gene number is reduced. In prostate and NCI60 datasets, the classifiers obtained 100% accuracy with minimum number of 9 and 80 genes, respectively. It was observed that the two-class dataset could obtain satisfactory classification results with less number of genes than the multi-category dataset. In NCI60 dataset, the 10-fold cross-validation accuracy did not exceed 90% when the gene number was less than 36. However, in prostate dataset, the classifier obtained 100% accuracy with minimum number of 9 genes.

To combine SVM-RFE with GA, a different number of top n ranking genes were chosen from the SVM-RFE algorithms as the candidate gene set, where n was set to 10, 20, 30, 50. Since the genetic algorithm is a randomly search model, 5 trials were executed on each candidate gene set, and the results were then averaged.

When Top-10 genes were searched from prostate cancer dataset (Table 2), the genetic algorithm was capable of finding smallest size of subset and achieves 100% classification accuracy. The average subset size of 5.4 genes is less than SVM-RFE method while it needs 9 genes to obtain the same accuracy.

When Top-50 genes were searched from NCI60 cancer dataset (Table 3), the genetic algorithm was capable of achieving 100% classification accuracy. The average subset size of 28 genes is much less than SVM-RFE method while it needs 80 genes to obtain the same accuracy.

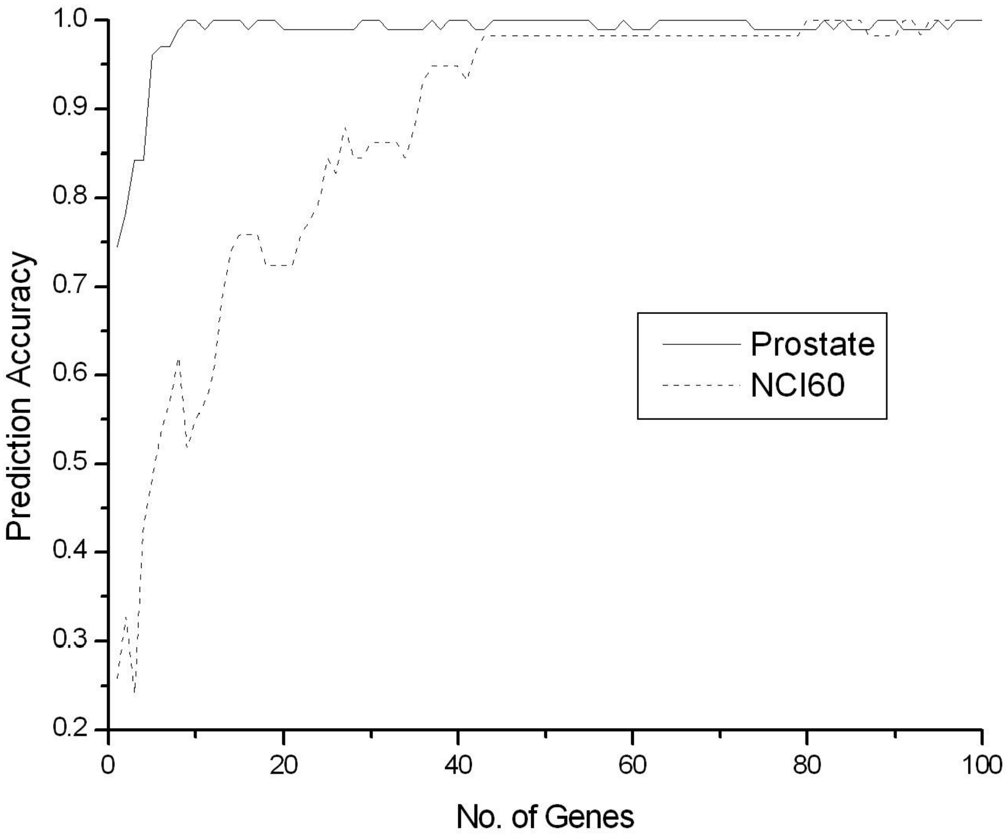


FIGURE 1 The 10-fold cross-validation prediction accuracy in prostate and NCI60 datasets when the number of genes was reduce from 100 to 1

It is observed that the choice of n is a critical problem in GA. When n is too small, the classifier is not able to obtain highest prediction accuracy. On the contrary, when n is too large, GA is possible to be trapped into local optimization, resulting in a larger number of selected genes.

TABLE 2 10-fold accuracy of the SVM-RFE/GA model on prostate cancer data set

Top *n* genes	Average accuracy (%)	Average subset size
10	100	5.4
20	100	7.0
30	100	8.0
50	100	13.2

TABLE 3 10-fold accuracy of the SVM-RFE/GA model on NCI60 cancer data set

Top *n* genes	Average accuracy (%)	Average subset size
10	65.8	6
20	84.6	13.8
30	94.1	20
50	100	28

The gene subset which can achieve highest prediction accuracy with minimum number of genes is defined as the “minimum gene subset”. In prostate cancer dataset, a gene subset selected from Top-10 genes contains minimum number of genes (n=5) while achieving 100% prediction accuracy, and the 5 selected genes are shown in Table 4.

TABLE 4 The selected genes of the minimum subset from prostate cancer dataset

Probe Set ID	Gene Symbol	Gene Title
32786_at	JUNB	jun B proto-oncogene
40282_s_at	CFD	complement factor D (adipsin)
41223_at	COX5A	cytochrome c oxidase subunit Va
41504_s_at	MAF	v-maf musculoaponeurotic fibrosarcoma oncogene homolog (avian)
863_g_at	SERPINB5	serpin peptidase inhibitor, clade B (ovalbumin), member 5

In NCI60 cancer dataset, a gene subset selected from Top-50 genes contains minimum number of genes (n=26) while achieving 100% prediction accuracy, and the 26 selected genes are shown in Table 5.

The results of the SVM-RFE/GA model were compared with some other algorithms in two aspects: the prediction accuracy and number of selected genes. In prostate cancer dataset (Table 6), only the SVM-RFE/GA model and SVM-RFE algorithm can achieve 100% prediction accuracy, but SVM-RFE/GA algorithm selects less number of genes.

The performance of SVM-RFE/GA model is more prominent in NCI60 cancer dataset (Table 7), the SVM-RFE/GA algorithm is capable of achieving 100% prediction accuracy, using much less number of genes (n=26) than SVM-RFE algorithm (n=80).

TABLE 5 The selected genes of the minimum subset from NCI60 cancer data set

Probe Set ID	Gene Symbol	Gene Title
AF005775_at	CFLAR	CASP8 and FADD-like apoptosis regulator
AF006041_at	DAXX	death-domain associated protein
D00017_at	ANXA2	annexin A2
D11327_s_at	PTPN7	protein tyrosine phosphatase, non-receptor type 7
D31888_at	RCOR1	REST corepressor 1
HG1869-HT1904_at	/	/
HG2147-HT2217_at	/	/
L38932_at	BECN1	beclin 1, autophagy related
L41349_at	PLCB4	phospholipase C, beta 4
M13929_s_at	MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
M37033_at	CD53	CD53 molecule
M59807_at	IL32	interleukin 32
M69181_at	MYH10	myosin, heavy chain 10, non-muscle
M90366_at	ZP2	zona pellucida glycoprotein 2 (sperm receptor)
M93036_at	EPCAM	epithelial cell adhesion molecule
U14577_s_at	MAP1A	microtubule-associated protein 1A
U49250_at	TBR1	T-box, brain, 1
U65785_at	HYOU1	hypoxia up-regulated 1
U95090_at	PRODH2	proline dehydrogenase (oxidase) 2
X03656_rna1_at	CSF3	colony stimulating factor 3 (granulocyte)
X12492_at	NFIC	nuclear factor I/C (CCAAT-binding transcription factor)
X52947_at	GJA1	gap junction protein, alpha 1, 43kDa
X75315_at	RBM38	RNA binding motif protein 38
X91247_at	TXNRD1	thioredoxin reductase 1
Y09305_at	DYRK4	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 4
Z18859_rna1_at	GNAT2	guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 2

TABLE 6 Results comparison of the SVM-RFE/GA model with some other algorithms in prostate dataset

Method	Prediction accuracy (%)	No. of Genes	Reference
SVM-RFE/GA	100	5	This study
SVM-RFE	100	9	[6]
SVM-RFE With MRMR	98.29	10	[11]
TSP	95.10	2	[19]
K-TSP	91.18	2	[19]

TABLE 7 Results comparison of the SVM-RFE/GA model with some other algorithms in NCI60 dataset

Method	Prediction accuracy (%)	No. of Genes	Reference
SVM-RFE/GA	100	26	This study
SVM-RFE	100	80	[6]
GA/SVM	87.93	27	[20]
GA/MLHD	85.37	13	[21]

4 Discussions

Gene selection has been a key research issue in microarray data analysis. Gene selection method aims to eliminate noisy, irrelevant and redundant genes, which can not only reduce the computational burden of the classifier, but also improve the classification accuracy of the classifier. In another aspect, the selected informative gene set, which contains least amount of genes, is much easier to be validated in subsequent molecular biology experiments.

As a type of embedded gene selection algorithm which is well coupled with support vector machine classifier, SVM-RFE achieves satisfactory results in the classification of microarray gene expression data. However, it ignores complementary relationship between genes. As a type of wrapper gene selection algorithm, GA selects genes nonlinearly by generating gene subsets randomly, which is efficient in detecting nonlinear relationships among genes. However, GA suffers instability in selecting genes and is prone to be trapped into local optimal solutions as the size of gene subset increases. In this study, a two-stage model is proposed to overcome these limitations.

Gene selection method eliminates noisy, irrelevant and redundant genes, in order to obtain highest classification accuracy with smallest number of genes. The performance of the SVM-RFE/GA model is validated on a binary and a multi-category microarray gene expression datasets. The SVM-RFE/GA model is superior to those previous studies in two aspects: firstly, the SVM-RFE/GA model can obtain highest prediction accuracy, and secondly, the number of selected genes is much less than theirs.

The selected genes in the minimum gene subsets are reported to be associated with cancer development. Out of the 5 selected genes from the prostate dataset (Table 4), JUNB [22] and SERPINB5 [23] were reported to be associated with prostate cancer. JunB is an upstream regulator of p16 and plays a key role in prostate cancer development [22]. Evidence shows that overexpression of SERPINB5 correlates with decreased prostate cancer metastasis [23].

Among the 26 selected genes of the minimum gene subset from NCI60 dataset (Table 5), a fraction of genes are found to have direct associations with cancers. CFLAR plays a role in inhibiting both apoptotic and necroptotic cell death [24]. Death domain-associated protein DAXX promotes ovarian cancer cell proliferation and chemoresistance [25]. Up-regulation of ANXA2 is reported to be associated with poor prognosis in human non-small cell lung cancer [26]. Evidence shows that overexpression of Beclin 1 may inhibit cell growth in colorectal cancer [27]. MYC plays a role in cell cycle progression, apoptosis and cellular transformation [28]. IL32 is expressed in different types of cancer [29]. EPCAM is reportedly to be strongly expressed and associated with breast cancer progression and metastasis

[30]. TXNRD1 is shown to be associated with poor prognosis in breast cancer [31].

5 Conclusions

In summary, this paper presents a model to combine GA with SVM-RFE, in order to enhance the performance of SVM-RFE. The performance of the SVM-RFE/GA model is validated on a binary and a multi-category microarray gene expression datasets. The SVM-RFE/GA model outperforms SVM-RFE algorithm, by taking advantage of both embedded and wrapper approaches. Compared with many previous gene selection algorithms, the SVM-RFE/GA model is capable of finding much smaller sized subsets of informative genes and achieving highest classification accuracy. Many selected genes by SVM-RFE/GA are reportedly associated with cancer, suggesting that SVM-RFE/GA model is an effective tool for molecular cancer classification and identification of cancer biomarkers using microarray data.

Acknowledgments

This study was supported by a grant from the National Natural Science Foundation of China (Grant No. 61373057), and a grant from the Zhejiang Provincial Natural Science Foundation of China (Grant No. Y1110763).

References

[1] Chin L, Andersen J N, Futreal P A (2011) *Nat Med* **17**(3) 297-303

[2] Ong F S, Das K, Wang J, Vakil H, Kuo J Z, Blackwell W L, Lim S W, Goodarzi M O, Bernstein K E, Rotter J I, Grody W W 2012 Personalized *Expert Rev Mol Diagn* **12**(6) 593-602

[3] Golub T R, Slonim D K, Tamayo P, Huard C, Gaasenbeek M, Mesirov J P, Coller H, Loh M L, Downing J R, Caligiuri M A, Bloomfield C D, Lander E S 1999) *Science* **286**(5439) 531-7

[4] Inza Y I, Larranaga P 2007 *Bioinformatics* **23**(19) 2507-17

[5] Li X, Peng S, Chen J, Lu B, Zhang H, Lai M 2012 *Biochemical and Biophysical Research Communications* **419**(2) 148-53

[6] Guyon I, Weston J, Barnhill S, Vapnik V 2002 *Machine Learning* **46**(1-3) 389-422

[7] Duan K B, Rajapakse J C, Wang H Y, Azuaje F 2005 *IEEE Transactions on Nanobioscience* **4**(3) 228-34

[8] Zhang X G, Lu X, Shi Q, Xu X Q, Leung H C E, Harris L N, Iglehart D J, Miron A, Liu J S, Wong W H 2006 *BMC Bioinformatics* **7** 197 (13 pages) doi:10.1186/1471-2105-7-197

[9] Zhou X, Tuck D P 2007 *Bioinformatics* **23**(9) 1106-14

[10] Tang Y C, Zhang Y Q, Huang Z 2007 *IEEE-ACM Transactions on Computational Biology and Bioinformatics* **4**(3) 365-81

[11] Mundra P A, Rajapakse J C (2010) *IEEE Transactions on Nanobioscience* **9**(1) 31-7

[12] Le Cun Y, Denker J, Solla S, Touretzky D S 1990 Optimal brain damage *Advances in Neural Information Processing Systems* Morgan Kaufmann 598-605

[13] Tan F, Fu X, Zhang Y, Bourgeois A 2008 *Soft Computing* **12**(2) 111-20

[14] Nicoletta D, Barbara P 2009 An evolutionary method for combining different feature selection criteria in microarray data classification *Journal of Artificial Evolution and applications* **2009** 1-10

[15] Cannas L, Dessi N, Pes B 2011 A Hybrid Model to Favor the Selection of High Quality Features in High Dimensional Domains *Intelligent Data Engineering and Automated Learning - IDEAL 2011* Berlin Heidelberg: Springer 228-35

[16] Mark H, Eibe F, Geoffrey H, Bernhard P, Peter R, Ian H W 2009 *SIGKDD Explor Newsl* **11**(1) 10-8

[17] Singh D, Febbo P G, Ross K, Jackson D G, Manola J, Ladd C, Tamayo P, Renshaw A A, D'Amico A V, Richie J P, Lander E S, Loda M, Kantoff P W, Golub T R, Sellers W R 2002 *Cancer Cell* **1**(2) 203-9

[18] Staunton J E, Slonim D K, Coller H A, Tamayo P, Angelo M J, Park J, Scherf U, Lee J K, Reinhold W O, Weinstein J N, Mesirov J P, Lander E S, Golub T R 2001 *Proc Natl Acad Sci USA* **98**(19) 10787-92

[19] Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D 2005 *Bioinformatics* **21**(20) 3896-904.

[20] Peng S H, Xu Q H, Ling X B, Peng X N, Du W, Chen L B 2003 *Febs Letters* **555**(2) 358-62

[21] Ooi C H, Tan P 2003 *Bioinformatics* **19**(1) 37-44

[22] Konishi N, Shimada K, Nakamura M, Ishida E, Ota I, Tanaka N, Fujimoto K 2008 *Clin Cancer Res* **14**(14) 4408-16

[23] Luo J L, Tan W, Ricono J M, Korchynskyi O, Zhang M, Gonias S L, Cheresh D A, Karin M 2007 *Nature* **446**(7136) 690-4

[24] Silke J, Strasser A 2013 *Sci Signal* **6**(258) pe2

[25] Pan W W, Zhou J J, Liu X M, Xu Y, Guo L J, Yu C, Shi Q H, Fan H Y 2013 *J Biol Chem* **288**(19) 13620-30

[26] Jia J W, Li K L, Wu J X, Guo S L 2013*Tumour Biol* **34**(3) 1767-71

[27] Chen Z, Li Y, Zhang C, Yi H, Wu C, Wang J, Liu Y, Tan J, Wen J 2013 *Dig Dis Sci* **58**(10) 2887-94

[28] Nair R, Roden D L, Teo W S, McFarland A, Junankar S, Ye S, Nguyen A, Yang J, Nikolic I, Hui M, Morey A, Shah J, Pfefferle AD, Usary J, Selinger C, Baker L A, Armstrong N, Cowley M J, Naylor M J, Ormandy C J, Lakhani S R, Herschkowitz J I, Perou C M, Kaplan W, O'Toole S A, Swarbrick A 2013 *Oncogene* (in print)

[29] Guenin S, Mouallif M, Hubert P, Jacobs N, Krusy N, Duray A, Ennaji M M, Saussez S, Delvenne P (2013) *Molecular Carcinogenesis* n/a-n/a

[30] Martowicz A, Rainer J, Lelong J, Spizzo G, Gastl G, Untergasser G 2013 *Mol Cancer* **12** 56

[31] Cadenas C, Franckenstein D, Schmidt M, Gehrmann M, Hermes M, Geppert B, Schormann W, Maccoux LJ, Schug M, Schumann A, Wilhelm C, Freis E, Ickstadt K, Rahnenfuhrer J, Baumbach JI, Sickmann A, Hengstler J G 2010 *Breast Cancer Res* **12**(3) R44

Author

Xiaobo Li

Current position, grades: full-time Assoc. Professor, Ph.D.
University studies: B.Sc. in Microelectronics (1990) from Nankai University (China), Master of Engineering (Research) (2004) from The University of Sydney (Australia) and Ph.D. in Pathology and Pathophysiology (2012) from Zhejiang University (China)
Research interests: different aspects of bioinformatics, machine learning and data mining

Dynamic analysis of ball-screw with rotating nut driven

Shigang Mu[*]

Department of Mechanical and Electrical Engineering, DeZhou University, DeZhou 253023, PR China

Received 1 March 2014, www.tsi.lv

Abstract

There is a certain degree difference between the static and operation condition for the high-speed Ball-screw with Rotating Nut. Therefore, this paper establishes a dynamic model of a preload-adjustable ball-screw with rotating nut by means of lumped-parameter and analyses the effects of changeable table position and work piece mass on the first three axial modes of the free vibration. A high-speed feeding system is modelled and its nature characteristics when the feeding system is in static, low and high rotate state. The results show that, at low speed state, the dynamics of the feeding system is the same as stationary state, and in high-speed conditions, the dynamics is quite different with the static state. The natural frequencies are notably changed with the position change of the table movement. The research lays an important theoretical foundation for developing this novel feed drive system.

Keywords: Ball screw, Dynamic analysis, Modal analysis, Frequency response

1 Introduction

The increasing demands that precision and engineering applications place on positioning systems has prompted an investigation into the ball screws. The reciprocating ball screw mechanism is a force and motion transfer device [1-3].Chin Chung Wei [4, 5] developed theoretical analyses of the kinematics of a single-nut double-cycle ball screw. Huang and Ravani [6] used the concept of medial axis transform (MAT) to analysis the contact stresses between ball-screw and ball-nut, in their analysis we can get normal forces, contact angles and contact stress in contact areas of ball-screw and ball-nut. These scholars study the dynamics of the traditional ball screw pair. The traditional screw transmission of ball screw pair, particularly long screw for large and heavy high-end NC machine, is quite heavier than nuts, the much more inertia force of ball screw rotation can lead to heat, deformation and serious energy consumption. The new-type nut-driven ball screw pair is the green product with advanced structure and promotion value which integrates rolling bearing functions and routine ball screw pair functions, during transmission, H. Weule [7] describes the advantages and characteristics of a Dynamic feed axis with ball screw drive and driven nut in comparison with the conventional electromechanical drive.

Currently, most literature analysis the feed system dynamics based on static state of, and ignored the influence of the speed of the feed system dynamics [8-11]. In fact, drive nut speed have great influence to feed the system dynamic characteristics, especially high speed feed condition. If you use the stationary state to the analysis of dynamic characteristics of high-speed feed cutting stability, will make a big error. This paper take a high-speed double nut drive type feed system as the research object. On the basis of actual working condition the lumped mass method is used to comparative analysis of the feed system quiescent state, low-speed operation dynamics differences in the state, the state of the high-speed operation. Related research conclusion can lay the foundation to a scientific analysis of the feed system dynamics characteristic and cutting stability.

2 Drive nut component axial stiffness calculation

2.1 NUT COMPONENT AXIAL STIFFNESS CALCULATION

Axial stiffness k_{nut} of nut components can be obtained through axial load exerted on the nut divided by amount of axial deformation:

$$k_{nut} = \frac{F_{axis}}{\delta_{axis}} . \quad (1)$$

In order to eliminate the axial clearance of ball screw and to improve the axial contact stiffness of ball screw pair, preloaded spacer is usually used to pre-tighten the double nut mechanism pretension, as shown in Figure 1. Given that normal force applied by each ball on nut A in ball screw pair to screw normal force of preload applied by preloaded spacer through nuts A and B to screw is P_p . When the axial working load is $\mathrm{F}_{axis} \neq 0$, nuts *A* and *B* are elastically deformed at the point of contact under joint action of axial load F_{axis} and normal force of preload P_p , and axial elastic deformation amount of nut *A* is equal to axial elastic deformation amount of nut *B*:

$$F_{axis} = \left(p_A - p_B \right) z \sin \alpha \cos \beta . \quad (2)$$

[*] ***Corresponding author*** e-mail: xiaff@126.com

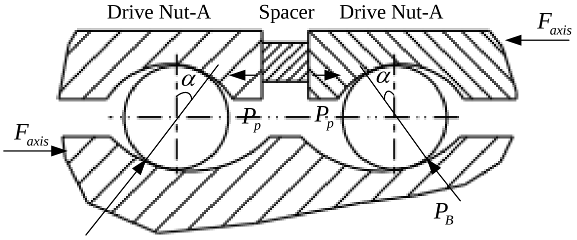


FIGURE 1 The ball screw vice work principle diagram of gasket type double nut prestressing structure

According to Hertz contact theory, elastic approach of two elastomers is proportional to power of their normal pressure $2/3^{th}$ power, and there

$$p_A^{2/3} - p_p^{2/3} = p_p^{2/3} - p_B^{2/3} . \tag{3}$$

When the known axial working load is F_{axis} and preload is P_p, values of P_A and P_B are obtained through equations (3).

2.2 INFLUNCING FACTORS OF DYNAMIC CHARACTERISTICS INDUCED BY ANGULAR VELOCITY

Nut-driven ball screw pair will, during high-speed operation, produce centrifugal force, gyroscopic moment, axial stiffness softening and other phenomena.As shown in Figure 2, centrifugal force of steel ball j increases with rotating speed of screw ω under action of axial load F_a, and steel ball moves to the sides of nut.Under joint action of preload,centrifugal force and normal force of internal and external channels,steel ball j revolves around screw axis, with its radius of $d_0/2$ and angular velocity of ω_{mj}:

$$F_{cj} = \frac{\pi}{12}\rho d^3 d_0 \omega^2 \frac{\omega_{mj}}{\omega} , \tag{4}$$

where d is the diameter of steel ball and ρ is mass density of steel ball.

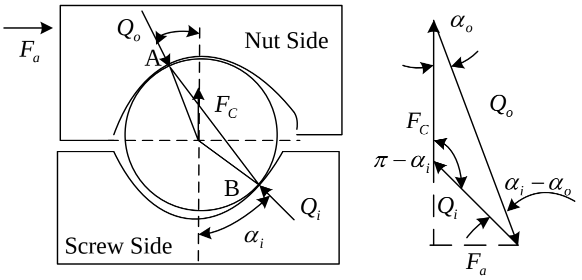
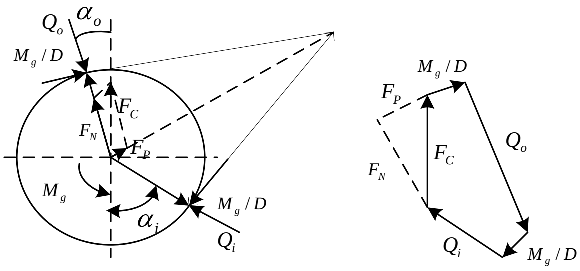


FIGURE 2 Loads acting on a ball at low-speed

At high-speed state,centrifugal force F_c is divided into F_N and F_P along the normal direction of lateral point of contact of nut and tangential direction of race. Under action of force F_N, steel ball extrudes the nut, resulting in the increase of lateral contact force of nut, decrease of lateral contact force of screw; under action of force F_P, steel ball moves upward, as shown in Figure 3.



FI

GURE 3 Loads acting on a ball at high-speed

At high-speed state, gyroscopic moment of steel ball is:

$$M_{gj} = J\omega^2 \frac{\omega_{bj}}{\omega}\frac{\omega_{mj}}{\omega}\sin\beta_j , \tag{5}$$

$$\beta_j = \arctan\frac{\sin\alpha_{oj}}{\gamma+\cos\alpha_{oj}} . \tag{6}$$

Expression for the relationship between angular velocity of driving nut ω and axial stiffness of driving nut components k_x is:

$$k_{nut} = \frac{k_o k_i \cos\alpha_o \cos\alpha_i}{k_o \cos\alpha_o + k_i \cos\alpha_i} . \tag{7}$$

When the feeding system is in high-speed rotation, axial stiffness of driving nut components will decrease gradually with increase of angular velocity, i.e. stiffness softening phenomenon.The reason for this phenomenon is that the centrifugal force leads to change in deformation of lateral raceway of screw of driving nut and contact area of lateral raceway of nut, resulting in the change of axial contact stiffness of driving nut.

3 The establishment of the dynamic model

3.1 ELASTIC MODEL STRUCTURE

In order to analysis the impact on dynamic characteristics of the system from elastic properties of ball screw, stiffness model is used for modeling of ball screw.The model of elastic structure system of nut-driven ball screw pair consisting of concentrated mass (inertia) and spring is shown in Figure 4.

In Figure 4, stiffness parameters of main transmission parts.synthetic axial stiffness k_a of ball screw and driving nut components, torsional stiffness of ball screw k_g, and axial stiffness between driving nut and working platform k_n. Main moments of inertia and quality parameters: moment of inertia of driving nut J_b mass of driving nut

m_n, and mass of working platform m_t. In consideration of synthetic torsional stiffness conditions, there will be certain angle difference between angular misalignment of driving nut θ_m input by the motor and angular misalignment of nut θ_n output by the motor. Under the action of axial force, screw and drive nut assembly will generate a certain amount of axial elongation; in consideration of synthetic axial stiffness of screw and driving nut components, axial displacement of nut of $(x_n - p\theta_n)$ is resulted, where p is the lead of screw.

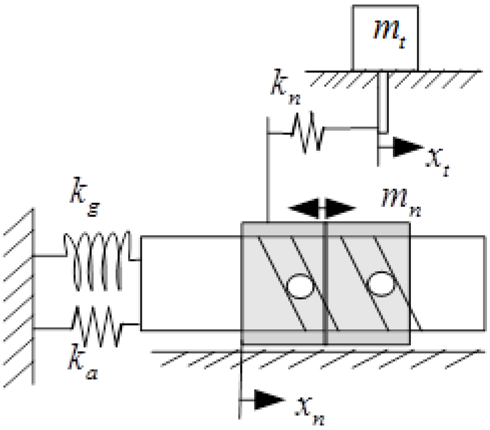


FIGURE 4 Modelling of the preload adjustable feed drive system with a lumped parameter system

Spring stiffness synthesis method can be used to work out axial stiffness of screw and synthetic axial stiffness of driving nutcomponents:

$$k_a = \left(\frac{1}{k_{screw}} + \frac{1}{k_{nut}} \right)^{-1} . \quad (8)$$

When determining the freedom degree of motion of moving parts, lagrangian energy method can be used for the construction of model for freedom outputs.First, based on speeds of transmission parts, the total kinetic energy of all moving parts can be obtained, which is:

$$T = \frac{1}{2} m_t \dot{x}_t^2 + \frac{1}{2} m_n \dot{x}_n{}^2 + \frac{1}{2} J_b \left(\frac{\dot{\theta}_m + \dot{\theta}_n}{2} \right)^2 . \quad (9)$$

The total potential energy of the system can be obtained simultaneously according to elastic deformation of the system, which is

$$V = \frac{1}{2} k_n (x_t - x_n)^2 + \frac{1}{2} k_a (x_n - p\theta_n)^2 + \frac{1}{2} k_g (\theta_m - \theta_n)^2 . \quad (10)$$

Taking $L = T - V$, lagrangian function about generalized coordinate and generalized force can be obtained, which is:

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_i} \right) - \frac{\partial T}{\partial q_i} + \frac{\partial V}{\partial q_i} = Q_i . \quad (11)$$

Taking freedoms of motion θ_m, θ_n, x_t and x_n as generalized coordinates q_i, $i = 1, \cdots, 4$, torque moment input T_m as generalized force, namely generalized coordinate matrix $q = (x_t \quad x_n \quad \theta_m \quad \theta_n)$:

$$Q = (F_x \quad 0 \quad T_m \quad 0)^T . \quad (12)$$

Make the lagrange equation represented in equation (11) further into the matrix form:

$$M\ddot{q} + Kq = Q . \quad (13)$$

As preload, working platform position and workpiece quality variy with the time during the processing, stiffness matrix and overall mass matrix of the system also vary with the time, which will lead to the change in natural frequency and modality in the dynamic system with the time.

3.2 BASIC PARAMETERS

This text cites double nut-driven ball screw pair as an example to construct the dynamic model for double nut-driven ball screw pair with adjustable preload.Table 1 lists geometric and physical parameters of the feeding system.

TABLE 1 List of the parameters used in the present analyses

Parameter	Unit	Value
nominal diameter	d_0/mm	41.4
screw length	L/mm	3000
helix angle	β/(°)	8.74
contact angle	α/(°)	45
ball's diameter	d_0/mm	6.35
worktable mass	m_L/kg	20.0
axial load	F_{axis}/N	500.0

4 Results and discussion

Immediate integration is one method to calculate the structural dynamic equation, and the commonly used method is Newmark method.Natural frequency and frequency response function of the system can be obtained via equation (14) by the use of Newmark integration method based on known conditions.

4.1 THE COMPARATIVE ANALYSIS OF NATURE FREQUENCY IN THE STATIC AND OPERATING STATE

Take the feeding system shown in Table 1 for an example,select stationary state ($\omega = 0$ rad/s), low-speed state ($\omega = 200$ rad/s and refer to Figure 4, no stiffness softening significantly occurs to the feed system) and high-speed state ($\omega = 3000$ rad/s,and at this time, stiffness softening occurs to the feeding system), to conduct dynamic analysis of the feeding system, and conduct comparative analysis on change in dynamic characteristics of the feeding system at such three states.

TABLE 2 The three kinds of operating state natural frequency contrast

Order number	**Natural frequency /Hz**		
	Static state	*Low speeds (200rad/s)*	*High speeds (3000rad/s)*
1 Order	208.7	206.4	157.2
2 Order	429.8	425.7	388.5
3 Order	823.9	815.6	764.3

It can be seen from Table 2 that natural frequency of spindle of the system at the low-speed state is basically the same as that at the stationary state, while the natural frequency at the high-speed state is very much different from that at the stationary state.It can be seen that for the spindle system, high speed induced bearing stiffness softening is an important factor that affects dynamic characteristics.

4.2 PRELOAD ON THE INFLUENCE OF THE DRIVE SYSTEM

Figure5 shows the frequency response function curve of working platform and axial acceleration of driving nut when the system is at $F_{axis} = 500.0N$ and preload is 40%. Figure 6 shows frequency response function curves of working platform and axial acceleration of driving nut when preload is respectively 40% and 35%.It can be observed that a serious resonance phenomenon occurs at $\omega_a = 157.2\text{Hz}$; this resonance characteristic is axial vibration of working platform due to synthesis of axial action of the parts, and the frequency value at this point is lower, therefore having a great impact on the system's processing accuracy.

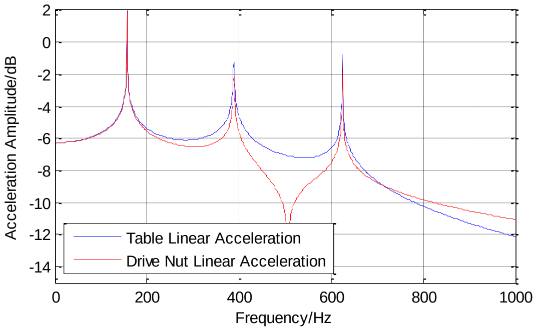


FIGURE5 Frequency responses of the ball screw acceleration and working table acceleration

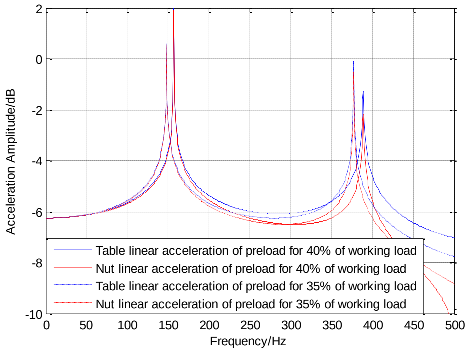


FIGURE6The resonant frequency shifts as the preload of the ball screw varies

4.3 STRUCTURE PARAMETERS ON THE INFLUENCE OF THE DRIVE SYSTEM

Figure 7 shows the impact of screw length, screw diameter, operating position of nut and other factors on the first-order natural frequency of the system.

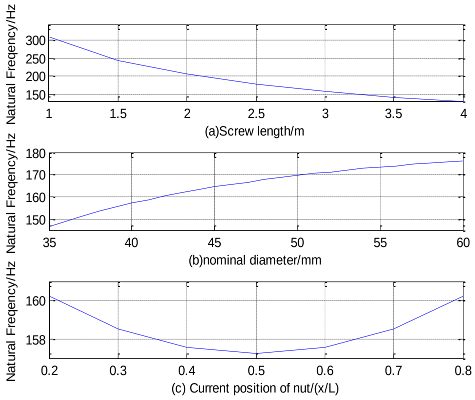


FIGURE7 The ball screw design parameters of the system's first natural frequency sensitivity analysis

This figure shows that with the increase in ball screw length, the first-order natural frequency of transmission system reduces; the first-order natural frequency of transmission system decreases; the bigger the screw diameter, the larger the first-order natural frequency, so increase of screw diameter is an effective way to improve the feed system's dynamic performance.

4.4 MOVING PARTS QUALITY CHANGE ON THE INFLUENCE OF THE DRIVE SYSTEM

Figure 8 shows the impact of frequency response function curve of mass variation of working platform vs. axial acceleration output when there is no workpiece on the working platform and the working platform is in the middle of screw.It shows that the mass of working platform has a great impact on the system's third-order natural frequency, but has little impact on the first-order and second-order natural frequencies.

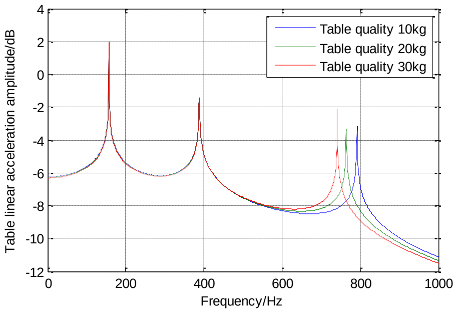


FIGURE 8 Effect of considering working table mass changes

Figure 9 and Figure 10 show the change in natural frequency of the system at the axial movement during the processing.Given that the quality of workpiece blank is 80kg, the mass of part after processing is 60kg.The natural frequency of each order varies linearly with change in mass of workpiece, but in general, although change in mass of workpiece has an impact on natural frequency of each order, which is less compared to that brought by the position of working platform.

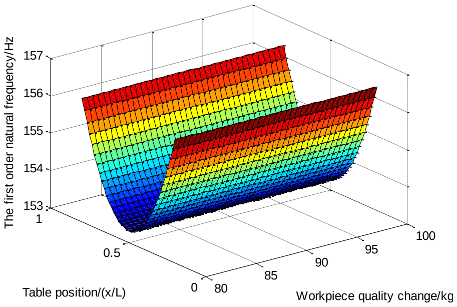


FIGURE 9 The quality and position to the first natural frequency influence

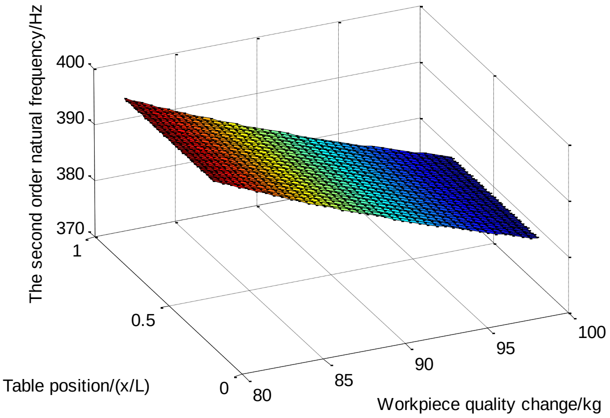


FIGURE 10 The quality and position to the second natural frequency influence

5 Conclusions

This text uses concentrated mass method and lagrange equation to construct the dynamic model for the transmission system with double nut driven ball screw pair under the action of preload. Take the high-speed nut-driven ball screw pair for an example, obtain natural frequency of the system and the corresponding vibration mode.The analysis results show that natural frequency of the system at the low-speed state is basically the same as that at the stationary state, while the natural frequency at the high-speed state is very much different from that at the stationary state.Natural frequency of the system is time-varying, and change in preload, design parameters and position of moving part during the processing, etc. has a remarkable impact on its natural frequency.The research results provide the important theoretical basis for development and manufacturing of such new feeding system as transmission system of nut-driven ball screw pair.

References

[1] Sun Yumin 2011 *Manufacturing Technology & Machine Tool* **7** 141-2

[2] Zhang Zuoying 2008 *Analysis and Experiment Study on Dynamic Performance of High-Speed Precision Ball Screw Mechanism* Jinan: Shandong University **6** *(in Chinese)*

[3] Altintas Y, Verl A 2011 *CIRP Annals-Manufacturing Technology* **60**(5) 779-96

[4] Wei C C, Lin J F 2003 *ASME Journal of Mechanical Design* **125**(10) 717-33

[5] Chin Chung Wei, Jen Fin Lin 2009 *Tribology International* **42**(13) 1816-31

[6] Huang H T, Ravani B 1997 *ASME J Mech Des* **119** 8-14

[7] Weule H, Frank T1997*Annals of the CIRP* **48**(1) 303-6

[8] Pislaru C 2004 *Proceedings of the Institution of Mechanical Engineers* **218**(2) 111-20

[9] Kim Min-Seok, Chung Sung-Chong 2005 *International Journal of Machine Tools & Manufacture* **45**(12) 1421-35

[10] Fujita T, Matsubara A, Kono D, Yamaji I 2010 *Precision Engineering* **34**(1) 34-42

[11] Zhou Y, Peng F Y, Cao X H 2011 *Mechanika* **17**(5) 523-8

Author

Shigang Mu, born in March 1976, DeZhou City Shandong Province

Current position, grades: PhD, researcher, Shandong University, China
University studies: doctoral degree on *Machinery and electronics* in Shandong University, China, in 2013
Scientific interest: mechanical structure dynamic analysis, modal analysis, vibration test and signal processing
Publications: 5
Experience: during the past three years the subject "Research on Dynamic Characteristics of High-speed Ball Screw with Nut Driven" (supervisor professor Xianying Feng) has been studied

Thermodynamic analysis of hydrogen production via zinc hydrolysis process

Ming Lv, Haiqiang Liu*

School of Mechanical Engineering, Hangzhou Dianzi University, 310018, Hangzhou, China

Received 1 March 2014, www.tsi.lv

Abstract

The thermodynamic studies were carried out for the hydrogen production via zinc hydrolysis. It is shows that it is reasonable to keep the temperature of zinc hydrolysis under 900 ºC. The system pressure has no notable thermodynamic influences on the hydrolysis reaction. The initial H_2O/Zn molar ratio should be controlled in a reasonable range. The concentration of steam in carrying gas in experiments should better be kept above 50%.

Keywords: hydrogen, hydrolysis, thermodynamics, zinc

1 Introduction

Two-step water splitting thermochemical cycles using metal-oxide redox pair are considered for the solar production of hydrogen [1]. Possible technical concepts of two-step cycles are discussed in paper [2] and [3], suggesting that the thermochemical cycles based on Zn/ZnO redox pair are very promising candidates. These cycles using the Zn/ZnO redox pair often proceed through two step: In the first, endothermic step, zinc oxide can be converted to zinc using high temperature solar heat by thermal dissociation [4, 5] or by carboreduction [2, 6]; The second, exothermic step, hydrogen is produced via the hydrolysis of zinc, which can be presented as follows:

$$Zn + H_2O = ZnO + H_2. \tag{1}$$

As one of the key steps in the thermo chemical cycle based on Zn/ZnO, only few information is available on the hydrolysis of zinc from published literature. In a series of thermo gravimetric analyses of commercial zinc powder and solar zinc powder in a temperature range of 350-500ºC, Weidenkaff et al.[5] found that the hydrolysis reaction proceeded faster for molten zinc and for zinc containing impurities, but a layer of ZnO prevented the reaction from reaching completion. The hydrolysis of submicron Zn particles in a temperature range of 330-360ºC was also studied by thermogravimetric analyses, and a fast surface reaction, corresponding to a mass increase of 2%, followed by a slow diffusion-limited reaction was observed [7]. The oxidation of liquid zinc with water vapor was studied by bubbling water vapor through bulk of liquid zinc at 450-500 ºC [8], and the results showed that the specific reaction rate increases as the water partial pressure increases, the main determining step of the hydrolysis reaction is the diffusion of reactants through the product zinc oxide layer. The oxidation of zinc vapor of about 750ºC and 800ºC with water vapor was also studied using a tubular aerosol flow reactor which features three temperature-controlled zones [9], up to 83% of zinc conversion could be obtained while the temperature of reaction zone is just below the Zn(g) saturation temperature. And several reaction parameters of the H_2 production by steam-quenching of Zn vapor were also studied in a hot-wall aerosol flow reactor, the results shows that high zinc conversions could get at a low quenching rate at the expense of low particle yield [10]. Detailed studies on the hydrolysis kinetics of zinc powder was also conducted by Vishnevetsky et al. [11], they found that the hydrolysis of zinc proceeded in two stage, and increasing the beginning temperature of the reaction is advantageous to the hydrolysis process, the reactivity of Solar Zinc is much higher than commercial zinc, which is agreed with paper [5]. Thermogravimetric analysis of the hydrolysis of zinc particles was also take out in a temperature range of 200-1000 ºC [12], two kinds of reaction mechanism of the zinc hydrolysis were revealed by Ming et al, in which the rate of hydrolysis reaction was limited by the evaporation of zinc and the diffusion of zinc through ZnO layer respectively.

In this paper, detailed thermodynamic studies were taken out for the hydrolysis of zinc. We focused on the thermodynamic influences of several important process parameters, including temperature, system pressure, initial H_2O/Zn molar ratio, and the concentration of steam in carrying gas.

2 Calculation methods

In this paper, the equilibrium composition of reacting mixture is calculated by the non-stoichiometric approach, in which the equilibrium composition is found by the direct minimization of Gibbs free energy. The famous

* ***Corresponding author*** e-mail: liuhaiqiang1980@126.com

computational thermo-chemistry software - FactSage was used for the calculation [13]. The thermodynamic parameters of reaction such as enthalpy change, Gibbs energy change were calculated by the REACTION module. The influences of temperature, system pressure and initial components ratio on the reaction equilibrium were studied by the EQULIB module.

The hydrogen yield ability of the zinc hydrolysis system was estimated by the hydrogen equilibrium yield ratio, which is defined as:

$$\mathrm{H_2\ yield[\%]} = n_{\mathrm{H_2,eq.}} / n_{\mathrm{Zn,initial}}, \quad (2)$$

where $n_{H2,eq.}$ is the hydrogen equilibrium molar amount, $n_{Zn,initial}$ is the initial zinc molar amount in system.

3 Results and discussion

The ΔH^0 and ΔG^0 of zinc hydrolysis reaction under different temperature at 1atm pressure were calculated. As shown in Figure 1, the hydrolysis of zinc is a moderate heat releasing reaction. When the reaction temperature is lower than 850°C, the stoichiometric reactants can be heated to the reaction temperature by the heat revealed by the reaction itself in ideal state. That means the zinc hydrolysis reaction can proceed auto-thermally. The reaction Gibbs energy change increases as the temperature increases. In ideal state, the zinc hydrolysis reaction can proceed at a temperature below 1223°C at 1 atm. According to previous kinetic studies of zinc hydrolysis [12], solid zinc will not react with water under low temperature as the restriction of chemical reaction kinetics. At normal state, the zinc hydrolysis reaction only proceeds at temperature higher than the melt point of zinc (417°C). So the reaction can be presented as follows:

$$\mathrm{Zn(l/g) + H_2O(g) \rightarrow ZnO(s) + H_2(g).} \quad (3)$$

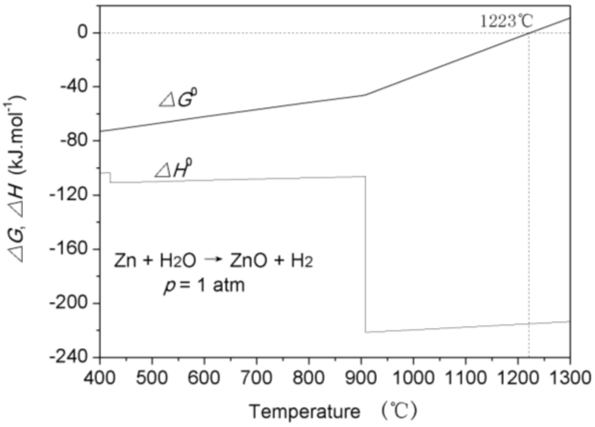


FIGURE 1 Temperature variations of ΔH^0 and ΔG^0 for the hydrolysis reaction of zinc at 1 atm

To study the thermodynamic influences of temperature and system pressure on zinc hydrolysis reaction, the equilibrium composition of initial 1 mol zinc and 1 mol water under different temperatures and different pressures are calculated. As shown in Figure 2, the calculation result shows that the hydrogen equilibrium yield ratio increases as temperature decreases. Obviously, under atmospheric pressure, the reaction temperature should be kept under 900°C to get a hydrogen equilibrium yield ratio higher than 90%.

When the temperature varies from 900°C to 1400°C, system pressure has great influence on the hydrogen equilibrium yield ratio. At the temperature of 1100°C, the hydrogen equilibrium yield ratio decreases from 85.7% to 17.8% while the system pressure increases from 0.1 atm to 10 atm. When the temperature is lower than 900C, as temperature decreases the influence of pressure on hydrogen equilibrium yield ratio decreases. When the temperature is lower than 700°C, the influence of system pressure is much small, as shown in Figure 2. From the sight of reaction thermodynamics, it is reasonable to control the reaction temperature under 900°C to get a high hydrogen production. Because the increase of pressure has little impact on reaction when the reaction temperature is below than 900°C, the zinc hydrolysis system pressure can be designed as atmospheric pressure, which is more favourable for practical utilization.

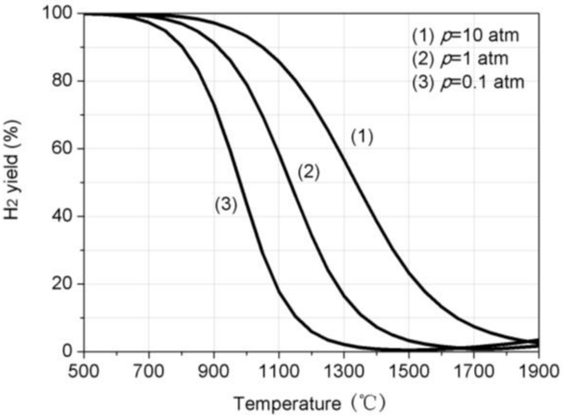


FIGURE 2 Variation of the H_2 equilibrium yield ratio as a function of the temperature and pressure (initial 1 mol Zn + 1 mol H_2O).

What's more, the result also shows that the increase of pressure can expand the high edge of zinc hydrolysis temperature window. However, it should be pointed out that water will split into hydrogen and oxygen at very high temperature. The thermodynamic calculation result shows that the low edge of water splitting temperature window is usually close to the high edge of zinc hydrolysis temperature window. As Figure 2 shows, under the pressure of 0.1 atm, the hydrogen equilibrium yield ratio decreases from 100% to 0 while the temperature increases from 500°C to 1500°C, the yielded hydrogen comes from the hydrolysis of zinc. Moreover, as the temperature increases above 1500°C at 0.1 atm, the hydrogen equilibrium yield ratio increases slowly because of the water splitting reaction.

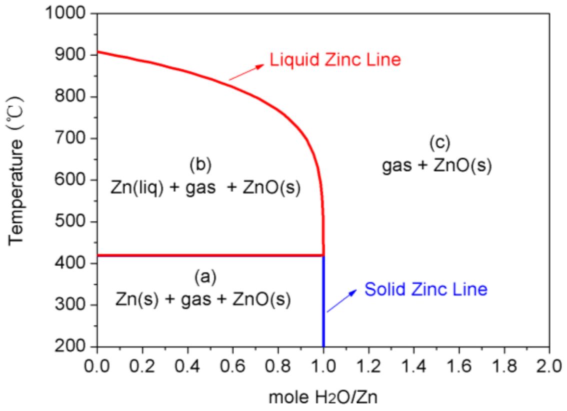


FIGURE 3 Variation of the Chemical equilibrium composition as a function of the temperature and initial H_2O/Zn molar ratio (p＝1 atm)

The influences of initial H_2O/Zn molar ratio were studied under atmospheric pressure. As Figure 3 shows, the diagram of equilibrium composition with different initial H_2O/Zn molar ratio and different reaction temperature can be divided into three areas. Under atmospheric pressure, when the H_2O/Zn ratio is lower than 1 and the temperature is lower than the melting point of zinc, in equilibrium state the zinc is in solid phase and the gas component is hydrogen, as show in Figure 3 area (a). The boundary of area (a) can be seen as the critical solid zinc phase line with a giving pressure and a giving temperature, which we called solid zinc line here. When the H_2O/Zn ratio is still kept lower than 1 and the temperature is between the melting point and the boiling point of zinc, as Figure 3 area(b) shows, for zinc the phase of gas and liquid coexist in equilibrium. And the gas components in area(b) are gaseous zinc and hydrogen. The boundary of area (b) can be seen as the critical liquid zinc phase line with a giving pressure and a giving temperature, which we called liquid zinc line here. When the H_2O/Zn ratio is higher than 1, as Figure 3 area(c) shows, there will be no liquid or solid zinc appear in equilibrium. The main components of gas in area(c) is hydrogen and steam at relative low temperature, while at relative high temperature gaseous zinc will represent as the incompletion of reaction. According to the previous studies, we care more about the reaction temperature in a range of zinc melting point to 900 ºC. So we focused on the liquid zinc line between area(b) and area(c). In this section of liquid zinc line, the x coordinate of each point represent the critical H_2O/Zn ratio at giving temperature and pressure for liquid zinc existence in equilibrium, which we called critical liquid zinc ratio in follows. And the y coordinate represent the critical temperature at giving initial H_2O/Zn ratio and pressure for liquid zinc existence in equilibrium, which we called critical liquid zinc temperature in follows. The critical liquid zinc ratio and the critical liquid zinc temperature at atmospheric pressure were giving in Table 1 and Table 2.

TABLE 1 Critical liquid zinc ratio at different temperatures (p=1 atm)

Temperatures (℃)	Critical liquid zinc ratio
500	0.998
600	0.985
700	0.921
800	0.697
900	0.08

TABLE 2 critical liquid zinc temperature at different initial compositions (p=1 atm)

Initial H_2O/Zn molar ratio	Critical liquid zinc temperature (℃)
0.2	886
0.4	859
0.6	823
0.8	767
1.0	420
1.2	/

We calculated the equilibrium composition under 800ºC and 1 atm as a example to study the influence of the initial H_2O/Zn molar ratio on zinc hydrolysis. As Figure 4 shows in the first stage, when the initial H_2O/Zn molar ratio is lower than the critical liquid zinc ratio (0.697 for 800 ºC), the content of equilibrium liquid zinc drops to zero quickly as the initial H_2O/Zn molar ratio increases.

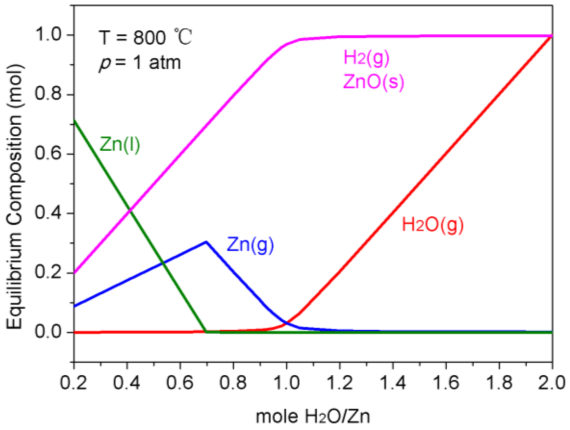
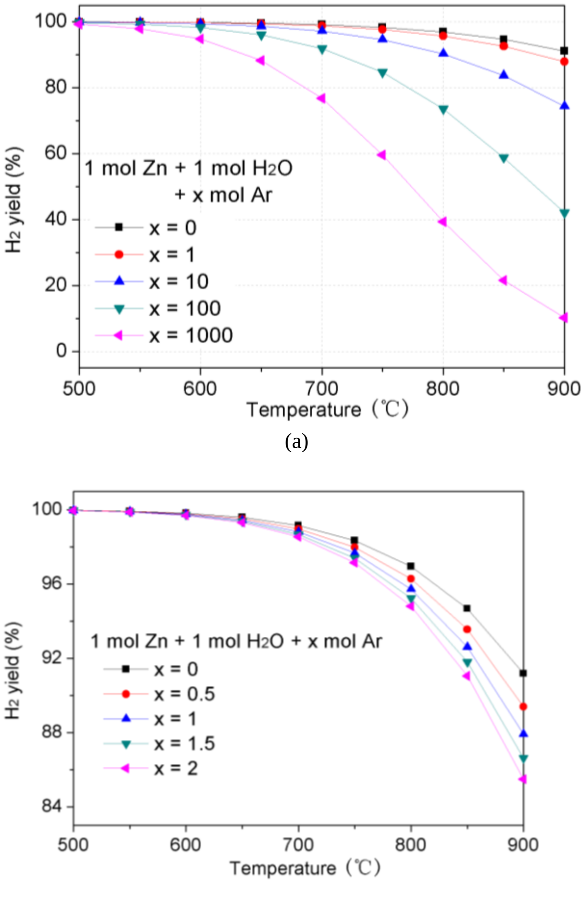


FIGURE 4 Variation of the equilibrium composition as a function of the initial H_2O/Zn molar ratio while T=800℃, p=1atm, the initial zinc molar amount was set constant at 1 mol

At the same time, the content of equilibrium gaseous zinc slowly increases linearly to a constant value. Therefore, it can conclude that the phase of zinc, which participates in hydrolysis reaction, is liquid. In the second stage, when the initial H_2O/Zn molar ratio increases from 0.697 to 1, the gaseous zinc begins to participate in hydrolysis reaction, and the content of equilibrium gaseous zinc decreases linearly to zero. Above all, we can conclude that the initial H_2O/Zn molar ratio should be kept larger than 1 to avoid the production of gaseous zinc and to improve the conversion of zinc. In addition, from sight of hydrogen concentration in equilibrium gas, the initial H_2O/Zn molar ratio should not be too large. Because the larger H_2O/Zn molar ratio will cause the larger waste of steam, which will bring the larger lost of raw material and energy. What's more, the larger H_2O/Zn molar ratio will be disadvantageous to the separation of hydrogen from product gases and to the recovery of exhausted water.



(a)

(b)

FIGURE 5 Variation of the equilibrium composition as a function of the initial H_2O/Zn molar ratio while T=800℃，p=1atm, the initial zinc molar amount was set constant at 1 mol

As in previous experimental studies of zinc hydrolysis, water steam was often brought into reactor via inert carrying gases [8, 12]. Therefore, the influence of inert carrying gas was also studied in this paper. We take argon as an example for inert carrying gas, as it is usually used in experiments as an inert content. The hydrogen equilibrium yield ratio at different temperatures with giving initial molar content of argon and 1 mol zinc and 1 mol water were calculated. As Figure 5 shows, the hydrogen equilibrium yield ratio at constant temperature decreases as the initial amount of argon increases. Higher the temperature is, larger the influence of inert gas amount appears on the reaction equilibrium. So it is suggested to use lower flow rate of inert carrying gas in experiments. In the temperature range of 500 ºC to 900ºC, as shown in Figure 5, the hydrogen equilibrium yield ratio dropped a little when the initial Ar/H_2O ratio varies from 0 to 2. Above all, the concentration of steam in carrying gas in experiments should better be kept above 50%, so that the relative error of thermodynamic equilibrium caused by the inert carrying gas can be controlled in 4%.

4 Conclusions

Detailed thermodynamic studies were carried out for the hydrogen production via zinc hydrolysis in this paper. The thermodynamic heat effect and several important process parameters of zinc hydrolysis reaction were studied. The study results show that the zinc hydrolysis can proceed auto-thermally. It is reasonable to keep the temperature of zinc hydrolysis under 900ºC to get a high zinc conversion. The system pressure has no notable thermodynamic influences on the hydrolysis reaction. The initial H_2O/Zn molar ratio should be controlled in a reasonable range. The concentration of steam in carrying gas in experiments should better be kept above 50% to get a relative precise experimental result.

Acknowledgements

Authors wish to acknowledge the sponsor of the National Natural Science Foundation of China (No.51209062 and No. 51305113), the Science and Technology Project of Zhejiang Province (No. 2012R10003 and No. 2010R50003), and the Zhejiang Provincial Natural Science Foundation of China (No. LQ12E05002).

References

[1] Steinfeld A, Kuhn P, Reller A, Palumbo R, Murray J, Tamaura Y 1998 Solar-processed metals as clean energy carriers and water-splitters *International Journal of Hydrogen Energy* **23**(9) 767-74

[2] Kraupl S, Steinfeld A 2005 Monte Carlo radiative transfer modeling of a solar chemical reactor for the co-production of zinc and syngas *Journal of Solar Energy Engineering, Transactions of the ASME* **127**(1) 102-8

[3] Abanades S, Charvin P, Flamant G 2007 Design and simulation of a solar chemical reactor for the thermal reduction of metal oxides: Case study of zinc oxide dissociation *Chemical Engineering Science* **62**(22) 6323-33

[4] Palumbo R, Lede J, Boutin O, Ricart E E, Steinfeld A, Moller S, Weidenkaff A, Fletcher E A, Bielicki J 1998 The production of Zn from ZnO in a high-temperature solar decomposition quench process - I. The scientific framework for the process *Chemical Engineering Science* **53**(14) 2503-17

[5] Weidenkaff A, Reller A W, Wokaun A, Steinfeld A 2000 Thermogravimetric analysis of the ZnO/Zn water splitting cycle *Thermochimica Acta* **359**(1) 69-75

[6] Adinberg R, Epstein M 2004 Experimental study of solar reactors for carboreduction of zinc oxide *Energy* **29**(5) 757-69

[7] Ernst F O, Steinfeld A, Pratsinis S E 2009 Hydrolysis rate of submicron Zn particles for solar H_2 synthesis *International Journal of Hydrogen Energy* **34**(3) 1166-75

[8] Berman A, Epstein M 2000 The kinetics of hydrogen production in the oxidation of liquid zinc with water vapor *International Journal of Hydrogen Energy* **25**(10) 957-67

[9] Wegner K, Ly H C, Weiss R J, Pratsinis S E, Steinfeld A 2006 In situ formation and hydrolysis of Zn nanoparticles for H_2 production by the 2-step ZnO/Zn water-splitting thermochemical cycle *International Journal of Hydrogen Energy* **31**(1) 55-61

[10] Melchior T, Piatkowski N, Steinfeld A 2009 H_2 production by steam-quenching of Zn vapor in a hot-wall aerosol flow reactor *Chemical Engineering Science* **64**(5) 1095-101

[11] Vishnevetsky I, Epstein M 2007 Production of hydrogen from solar zinc in steam atmosphere *International Journal of Hydrogen Energy* **32**(14) 2791-802

[12] Lv M, Zhou J H, Yang W J, Cen K F 2010 Thermogravimetric analysis of the hydrolysis of zinc particles *International Journal of Hydrogen energy* **35**(7) 2617-21

[13] Zhao Y L, Zhang Y M, Bao S X, Chen T J, Han J 2013 Calculation of mineral phase and liquid phase formation temperature during roasting of vanadium-bearing stone coal using FactSage software *International Journal of Mineral Processing* **124** 150-3

Authors

Lv Ming, born on February 26, 1982, Hunan China

Current position: Ph.D., lecturer of Department of Ocean Engineering in Hangzhou Dianzi University.
University studies: Zhejiang University
Scientific interest: major in clean energy production and artificial upwelling
Experience: Zhejiang University 2004/9-2009/12 Energy and Environment Engineering PhD research subjects engaged in clean energy production and artificial upwelling

Liu Haiqiang, born on April 19, 1980, JiangXi China

Current position: Ph.D., lecturer of Department of Ocean Engineering in Hangzhou Dianzi University
University studies: Zhejiang University
Scientific interest: Intelligent design and digital product design
Experience: Zhejiang University 2005/9-2010/9 Mechanical Manufacturing and Automation PhD research subjects engaged in theory and method of Product Data Management well known about PLM methodology and research integrated techniques of CAX/PDM, and focused on build the integrated product data model

Study on prediction of sintering drum strength under small sample lacking information

Qiang Song[1]*, Ai-min Wang[2]

[1]*Mechanical Engineering Department, Anyang Institute of Technology, Anyang 455000, Henan, China*

[2]*Computer and Information Engineering College, Anyang Normal University, Anyang 455000, Henan, China*

Received 1 March 2014, www.tsi.lv

Abstract

The paper provides a grey model and support vector machine algorithm and method for prediction of sinter drum strength based on the characteristics of large time delay, strong coupling, nonlinear, sintering process, put forward a kind of Combination forecasting model of drum strength based on grey model and support vector machine, the drum strength of sinter ore Laboratory values as output variables, the variables associated with the drum strength of sinter as input variables, using support vector machine powerful machine learning method and strong nonlinear fitting ability, so as to establish a stable, high precision of drum strength, the drum strength stronger generalization ability of the forecasting model, the method of the method has the high prediction accuracy, fast and convenient, and has great popularization and application value, and lay a good foundation for the green sintering technology of sintering.

Keywords: GM(1,1), LS-SVM, drum strength, Prediction

1 Introduction

The sintering process of iron and steel enterprises is the powdered iron material (such as Brazil, South Africa ore , India) with adding a certain proportion of the flux and fuel ignition and combustion, tiled in large sintering machine, produces a certain amount of fuel combustion with high temperature liquid phase, the other of un-melted particles bonded together, after cooling into porous block ore has a certain strength, as the blast furnace smelting of raw materials. The sinter is always the major raw materials for blast furnace at home and abroad, especially in China, the blast furnace sinter have accounted for more than 90%, the sinter output and quality directly affects the quality and quantity of indexes of iron-making and steelmaking. Therefore, the sintering production occupies an important position in China's iron and steel enterprises. At the same time, with China's accession to the WTO organization, the iron and steel industry in China has joined the ranks of international competition, adjusting the requires of the iron and steel industry structure and technological transformation, learn the advanced experience of foreign countries, sintering, iron-making raw material industry to accelerate the reduction of development have become an irreversible trend toward large-scale set, and it is imperative that domestic iron and steel enterprises to the international advanced iron and steel enterprises, to improve the control level of the existing sintering process to the advanced international level as soon as possible. From the control perspective, the sintering process are multivariable, nonlinear, large delay, strong coupling characteristics of complex controlled object, it is related to the temperature, pressure, speed, flow, a large number of physical parameters, including the complex process of physical change, chemical changes, and the distribution of gas in the solid material layer, the temperature field distribution etc. many aspects of the problem. Artificial traditional control method has been unable to meet the requirements of large sintering machine control; there is an urgent need for more accurate control method, stable to ensure normal operation of sintering production. Therefore, this paper uses the grey support vector machine algorithm to predict sinter tumbler strength, achieved satisfactory results.

1.1 SINTERING PROCESS

Sintering is a method that makes powdered materials (such as fine ore or preparation concentrate) into block mass under conditions involving incomplete fusion by heating to high temperature. Its production is sinter, which is irregular and porous. The following parts are usually included in sintering process: acceptance and storage of iron-containing raw materials, fuel and flux; crushing and screening of raw materials, fuel and flux; batching, mix-granulation, feeding, ignition and sintering of mix material; crushing, screening, cooling and size-stabilization of sinter. The flowchart is shown in Figure 1.

[*] ***Corresponding author*** e-mail: songqiang01@126.com

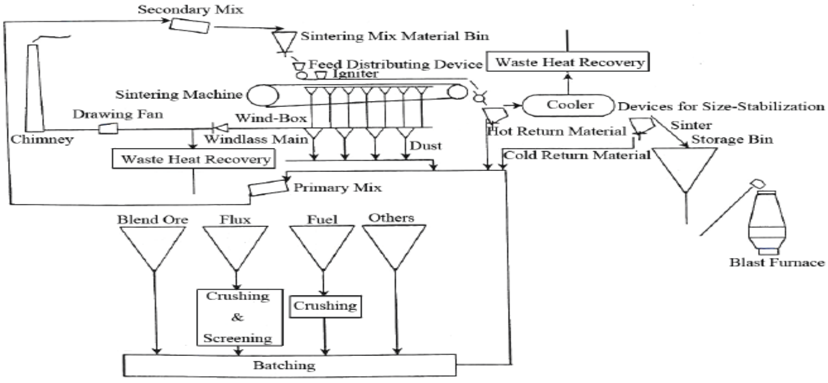


FIGURE 1 Sintering process

2 Grey model

2.1 GREY RESIDUAL ERROR CORRECTION MODEL

The term grey System is first published in a paper by Professor Deng Ju-Long, he presented the idea of "Control Problem of Unknown System" in his paper at the Sino-US Conference on Control System held in Shanghai in 1981. Presently, his paper on "Control Problem grey Systems" has been published in the Journal of System and Control Letters, nothing that the grey system theory is formal declaration on the international academia then. The basic concept of grey system theory is that all the nature for the existence of known information is white while unknown information is black, and the uncertainty information between the known (white) and the unknown (black) is grey. Grey system is mainly to dig out the nature of system under lack of information. It emphasizes information supplement to the system, and full use of white information have been identified through conducting systematic relational analysis and model construction. It makes the system state change from grey to white by prediction and decision methods to explore and understand the system [5-9]

A grey system is a system that is not completely known, i.e., the knowledge of the system is partially known and partially unknown. In recent years, grey models have been successfully employed in many prediction applications. The GM (1,1) model means a single differential equation model with a single variation. The modelling process is as follows: First of all, observed data are converted into new data series by a preliminary transformation called AGO (accumulated generating operation). Then a GM model based on the generated sequence is built, and then the prediction values are obtained by returning an AGO's level to the original level using IAGO (inverse accumulated generating operation).

A grey modelling algorithm is described as follows. (1) Suppose there is a set of discrete data that is unequal intervals as follows:

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), ..., x^{(0)}(\mathrm{n})) . \tag{1}$$

Accumulate the discrete data above once to get a new serial, that is

$$x^{(1)} = (x^{(1)}(1), x^{(2)}(2), ..., x^{(\mathrm{n})}(\mathrm{n})) . \tag{2}$$

(2) The GM (1,1) model can be constructed by establishing a first order differential equation for $x^{(1)}(\mathrm{t})$ as:

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = u . \tag{3}$$

The equation's general solution is: $x^{(1)}\left(t\right) = Ce^{-at} + \frac{\mu}{a}$

(3) The grey parameter a and u can be obtained by using the least square method:

$$\hat{\alpha} = \begin{bmatrix} a \\ u \end{bmatrix} = (B^T B)^{-1} B^T Y_N , \tag{4}$$

where $B = \begin{bmatrix} -0.5\left(x^{(1)}\left(2\right) + x^{(1)}\left(1\right)\right) & 1 \\ -0.5\left(x^{(1)}\left(3\right) + x^{(1)}\left(2\right)\right) & 1 \\ ... & ... \\ -0.5\left(x^{(1)}\left(n\right) + x^{(1)}\left(n-1\right)\right) & 1 \end{bmatrix}$,

$Y_N = \begin{bmatrix} x^{(0)}\left(2\right) \\ x^{(0)}\left(3\right) \\ ... \\ x^{(0)}\left(n\right) \end{bmatrix}$ grey parameters $\hat{\alpha}$ will be substituted into the time function, then:

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{u}{a} \right) e^{-ak} + \frac{u}{a} . \tag{5}$$

(4) Dealing $\hat{x}^{(1)}(k)$ for derivative and return to original equation then obtain

$$\hat{x}^{(0)}(k+1)=(1-e^{a})\left(x^{(0)}(1)-\frac{u}{a}\right)e^{-ak}. \tag{6}$$

(5) Calculating the difference of $x^{(0)}(k)$ and $\hat{x}^{(0)}(k)$ and the relative error $\varepsilon^{(0)}(k)=x^{(0)}(k)-\hat{x}^{(0)}(k)$.

The residual GM(1,1) model could be established to improve the predictive accuracy of the original GM(1,1) model. The modified prediction values can be obtained by adding the forecast values of the residual GM(1,1) model to the original $\hat{x}^{(0)}(k)$. However, the potency of the residual series depends on the number of data points with the same data, which is usually small when there are few observations. In these cases, the potency of the residual series with the same data may not be more than four, and a residual GM(1,1) model cannot be established. Here, we present an improved grey model to solve this problem.

2.2 RESIDUAL FORECASTING MODEL

To evaluate modelling performance, we should do synthetic test of goodness

$$C=\frac{s_2}{s_1}, \tag{7}$$

where $s_1=\frac{1}{n}\sum_{k=1}^{n}\left(x^{(0)}-\bar{x}^{(0)}\right)^2$, $s_2=\frac{1}{n}\sum_{k=1}^{n}\left(\varepsilon(k)-\bar{\varepsilon}\right)^2$ deviation between original data and estimating data:

$$\varepsilon^{(0)}=(\varepsilon(1),\varepsilon(2),...,\varepsilon(\text{n}))=$$
$$(\hat{x}^{(0)}(1)-\hat{x}^{(0)}(1),\hat{x}^{(0)}(2)-\hat{x}^{(0)}(2)-,...,\hat{x}^{(0)}(\text{n})-\hat{x}^{(0)}(\text{n})).$$

$$P=P\left\{\left|\varepsilon(k)-\bar{\varepsilon}\right|<0.674581\right\}. \tag{8}$$

The precision grade of forecasting model can be seen in Table. 1.

Finally, applying the inverse accumulated generation operation (AGO), we then have prediction values $\hat{x}^{(0)}(k)=\hat{x}^{(1)}(k)-\hat{x}^{(1)}(k-1)$.

TABLE 1 Precision grade of forecasting model

Precision grade	P	C
Good	0.95≤p	C≤0.35
Qualified	0.80≤p<0.95	0.35<C≤0.5
Just	0.70≤p<0.80	0.5<C≤0.65
Unqualified	p<0.70	0.65<C

2.3 ESTABLISH THE NEURAL NETWORK MODEL

This paper will change the variables related to the research index to carry on the estimation respectively, which will obtain a few estimated values as the input of BP neural network and adopt a hidden layer that the transfer function is the sigmoid function, while the output of grey neural network is real examination values of the alkalinity, the model structure shows Figure 1.

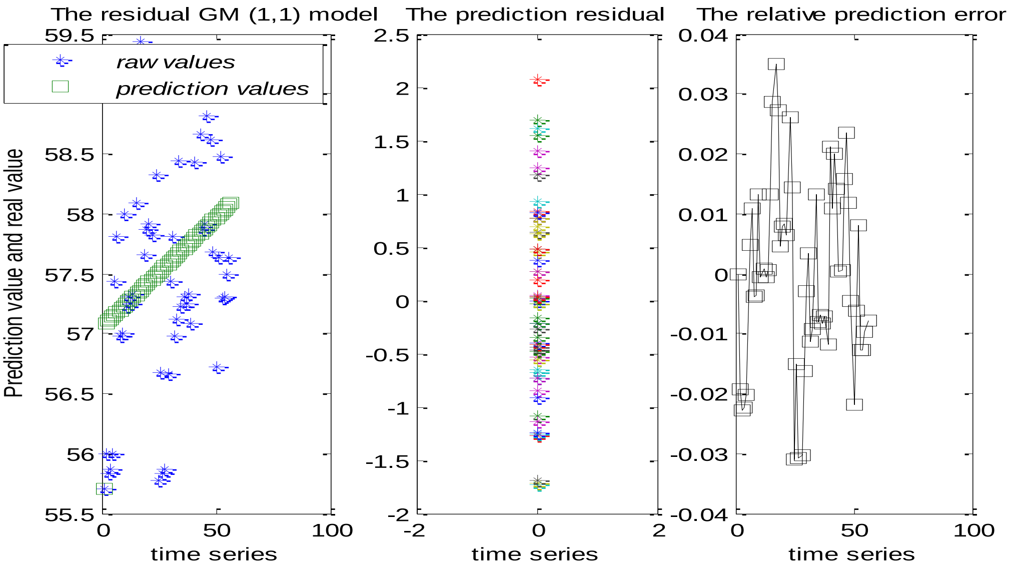


FIGURE 2 The prediction of Drum strength based on Grey residual error correction model

This paper adopts the grey neural network to predict the drum strength, aiming at this important output index, in the whole craft process, synthesizes the variables related to the drum strength make sure that ten important input variables as the input of GM(1,1), such as the layer thickness the trolley speed the first mixing water rate the mixing temperature the content of SiO2 in the mineral the content of CaO in the mineral the content of FeO in the mineral the second mixing water rate the proportion of CaO the proportion of Coal. This ten important variables(fifty-six datum) store in Excel database .In Matlab6.5,if we use the "import wizard", we may easily embed datum of Excel database into Matlab6.5,as long as we input the database name in Matlab window, we may employ the database respectively.

As can be seen, the grey theory to predict the residual correction, its accuracy is not improved, the grey model prediction accuracy of linear system is relatively high, but for the grey theory of nonlinear system, frequent changes in the still is not be desired, and therefore need to use other algorithms.

3 LS-SVM prediction model

3.1 LS-SVM PREDICTION model

The LS-SVM，evolved from the SVM，changes the inequality constraint of a SVM into all equality constraint and forces the sum of squared error(SSE) loss function to become an experience loss function of the training set [8]. Then the problem has become one of solving linear programming problems. This call be specifically described as follows. Given a training set $\{x_t, y_t\}_{t=1}^{N}$, with $x_t \in R^n$, $\mathrm{y_t} \in \mathrm{R}$, $x_t \in R^n$ is input vector of the first *t* samples, $\mathrm{y_t} \in \mathrm{R}$ is the desired output value of the first *t* corresponds to samples, N is the number of samples data, the problem of linear regression is to find a linear function y(x) that models the data. In feature space SVM models take the form:

$$y(x) = \omega^T \varphi(x) + \mathrm{b}\,, \tag{9}$$

where the nonlinear function mapping $\varphi(x): R^{\mathrm{n}} \to R^{n_{\mathrm{h}}}$ maps the high-dimensional space into the feature space.

Having comprehensively considered the complexity of function and fitting error, we can express the regression problem as the constrained optimization problem according to the structural risk minimization principle:

$$\min J(w,e) = \frac{1}{2} w^{\mathrm{T}} w + \gamma \frac{1}{2} \sum_{t=1}^{\mathrm{N}} e_t^2\,, \tag{10}$$

subject to the restrictive conditions, $y(x) = w^{\mathrm{T}} \varphi(x_t) + b + e_t$, for t = 1,…, N, where γ is margin parameter, and e_i is the slack variable for x_i.

In order to solve the above optimization problems, by changing the constrained problem into an unconstrained problem and introducing the Lagrange multipliers, we obtain the objective function:

$$L(w,b,e,\alpha) = J(w,e) - \sum_{\mathrm{t}=1}^{\mathrm{N}} \alpha_{\mathrm{t}} \{w^{\mathrm{T}} \varphi(x_t)\,, \tag{11}$$

where α_{t} is Lagrange multipliers. According to the optimal solution of Karush-Kuhn-Tucker(KKT) conditions, take the partial derivatives of (5) with respect to w, b and e respectively, and 1et them be zero, we obtain the optimal conditions as follows:

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \to w = \sum\limits_{t=1}^{N} \alpha_t \varphi(x_t) \\ \dfrac{\partial L}{\partial b} = 0 \to \sum\limits_{t=1}^{N} \alpha_t = 0 \\ \dfrac{\partial L}{\partial e_t} = 0 \to \alpha_t = \gamma e_t \\ \dfrac{\partial L}{\partial \alpha_t} = 0 \to w^T \varphi_t + b + e_t - y_t = 0 \end{cases}. \tag{12}$$

After elimination of $\mathrm{e_t}$ and w, the equation can be expressed as a linear function group:

$$\begin{bmatrix} 0 & \boldsymbol{1}^T \\ \boldsymbol{1} & \varphi(x_t)^T \varphi(x_t) + \boldsymbol{D} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix}, \tag{13}$$

where $\boldsymbol{y} = [y_1, \cdots, y_N]$, $1 = [1, \cdots, 1]$, $\alpha = [\alpha_1, ..., \alpha_N]$, $\boldsymbol{D} = diag[\gamma_1, \cdots, \gamma_{\mathrm{N}}]$. Select γ >0, and $\boldsymbol{\varphi} = \begin{bmatrix} 0 & \boldsymbol{1}^T \\ \boldsymbol{1} & \varphi(x_t)^T \varphi(x_t) + \boldsymbol{D} \end{bmatrix}$ guarantee matrix $\begin{bmatrix} \mathrm{b} \\ \boldsymbol{\alpha} \end{bmatrix} = \boldsymbol{\varphi}^{-1} \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix}$.

Finally, the LS-SVM regression model can be expressed as

$$y(x) = \sum_{t=1}^{N} \alpha_t \exp\{-\| x - x_t \|_2^2 / 2\sigma^2\} + b\,, \tag{16}$$

where σ is a positive real constant. Note that in the case of RBF kernel function, one has only two additional turning parameters σ and γ, which is less than standard SVM [12].

This LS-SVM regression leads to solving a set of linear equations, which is for many application in different areas. Especially, the solution by solving a linear system is instead of quadratic programming. It can decrease the model algorithm complexity and shorten computing time greatly. The LS-SVM algorithm software package is run in MATLAB 7.0.1 software.

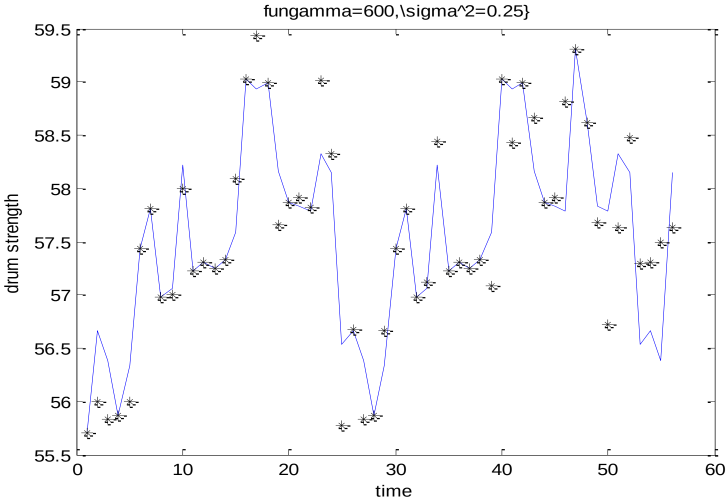


FIGURE 3 The prediction of Drum strength based on LS-SVM

3.2 COMBINATION FORECASTING MODEL.

We could see from the example that the difference value between forecasting value and practical value is large if the single model GM (1, 1) is employed to forecast. In order to make the forecasting result as close as the real value, literature [8] provide a combined grey neural network model. This kind of model could make the forecasting value as the input sample (learning sample) of the neural network, and make the real value as the target sample of the neural network. Adopting suitable structure and training on the neural network, then we could get a series of authority value threshold value, which correspond some corresponding crunodes [13 - 16].

This kind of combined method mainly use the characteristics such as three BP neural network layers carries the network which contains at least one concealed layer of S style and one linear output layer could approach any rational function. Through training, we endow the neural network the ability of simulating the relation between sequence data and sequence respectively. Meanwhile, literature provides the way that compensating the modelling error by neural network and blur logic, and doing some study training of network using heredity arithmetic. Thus, we bring forward a new combined model of grey neural network synthesizing the literature.

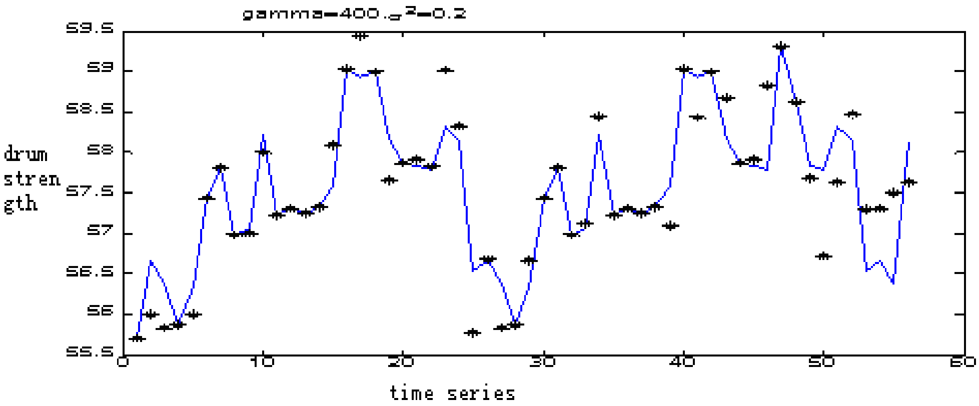


FIGURE 4 The prediction of Drum strength based on Combination

TABLE 2 The forecasting results of the three models

Model name	Mean relative error	Mean absolute error	Mean square error
Grey residual error correction model	2.09	0.65	2.888
LS-SVM	0.89	-0.042	9.02e-004
Combination forecasting model	0.03727	0.03269	8.66e-006

Figure.2-4 show the forecasting results of the three models respectively, respectively from Figure.2 ,it can be seen that although Grey residual error correction model has the better forecasting precision in drum strength the accuracy starts to decline obviously in some data points. In figure.3, the forecasting curve of LS-SVM is more close to observed curve than the curve of Grey residual error correction model from Figure.4, it can be seen that

LS-SVM as the highest forecasting accuracy. This combination method is better than the single method, which can get optimal combination prediction model from Table 1 it can be seen that three models present quite satisfactory forecasting results. By comparing the Mean relative error, mean absolute error and Mean square error of Combination forecasting model is smaller than that of Grey residual error correction model and LS-SVM. Moreover, combination-forecasting model has higher precise prediction than grey residual error correction

4 Conclusion

The original GM(1,1) model is a model with a group of differential equations adapted for variance of parameters, and it is a powerful forecasting model, especially when the number of observations is not large. In this paper, we have applied an improved grey GM(1,1) model by using a technique that combines residual modification with ls-svm. Our study results show that this method can yield more accurate results than the original GM(1,1) model and also solve problems resulting from having too few data, which may lead the same data residuals lower than four and violate the necessary condition of setting up a GM(1,1) model. The improved grey models were then applied to predict the drum strength in sintering process. Finally, through this study, our Combination forecasting model, is an appropriate forecasting method to yield more accurate results than the original GM(1,1) model and LS-SVM. In short, the Combination forecasting model is effective with the advantages of high precision, less requirement of samples and simple calculation. The grey neural network will be greatly applied and extended in future. However, we should admit that the combined grey neural network model is not perfect both in theory and practice. We will continue to do further study and some discussion based on the grey neural network model in future.

Acknowledgments

This thesis is greatly indebted to all who have instructed us in the course of our studies for the Master's degree, including professors WANG Ai-min, whose excellent lectures help brighten and broaden our vision to a great extent. In particular, we would like to extend our heart-felt thanks to our supervisor, professor YANG Cheng-zhi, for his meticulous instruction in both my study and life. We owe professor YANG Cheng-zhi the choice of the subject and the rectification of language errors in the thesis. Furthermore, our thanks are expressed to our roommates GUO Yi-jun and FENG Meng who help us with great encouragement, and their rich technology knowledge gives me great assistance in enlarging my point of view.

References

[1] FAN Xiao-hui, WANG Hai-dong 2002 *Mathematical model and Artificial Intelligence of sintering process* Central South University Press *(in Chinese)*

[2] Guo Wei-qiang, Yan Fei, Han Ning 2012 Wildfire Smoke Detection using Dynamic Features and Neural Network *JDCTA* **6(**11) 61-8 *(in Chinese)*

[3] SHI Meifeng, HE Zhongshi, LIU Xin, HUANG Meiyan 2012 A Distance Weight Object Tracking Method based on Combining Mean Shift and GM(1,1) *JDCTA* **6**(1) 318-25 *(in Chinese)*

[4] Zhipiao Liu, Qibo Sun, Shangguang Wang, Fangchun Yang 2012 The Performance Prediction of Cloud Service via JOGM(1,1) Model *AISS* **4**(5) 70-7 *(in Chinese)*

[5] Rui-lin Xu, Xin Xu, Xing-zhe Hou, Bo Zhu, Min-you Chen 2012 A Prediction Model for Wind Farm Power Generation based on Genetic-neural Network *JCIT* **7**(14) 11-9 *(in Chinese)*

[6] Deng J L 2002 *Grey Forecasting and grey Decision* Press of Hua Zhong University of Science and Technology *(in Chinese)*

[7] Rong-fei Ma 2011 Adaptive Error Control Mechanism based on GM(1,1) model and Cross-layer Design for Video Streaming over Wireless *JDCTA* **5**(8) 215-21 *(in Chinese)*

[8] Ching-Lien Huang, Yung Hui Chen, Tian-Long John Wan 2012 Optimization of Data Mining in Dynamic Environments based on a Component Search Neural Network Algorithm *JCIT* **7**(7) 216-23 *(in Chinese)*

[9] Shaio Yan Huang , An-An Chiu, Bau-Chi Wang 2012 Applying Intellectual Capital on Financial Distress Prediction Model in Taiwan Information Technology and Electronic Industry *IJACT* **4**(8) 270-80 *(in Chinese)*

[10] Xiaodong Wu, Han Wu, Guoqing Han, Yongsheng An 2012 Predicting the Solubility of Sulfur in Hydrogen Sulfide Using a Back-Propagation Neural Network *IJACT* **4**(8) 281-7 *(in Chinese)*

[11] Han-Chen Huang 2012 Using Artificial Neural Networks to Predict Restaurant Industry Service Recovery *IJACT_CST* **4**(10) 315-21 *(in Chinese)*

[12] Fuwei Zhang, Qin Su 2012 Application of Wavelet Neural Network into the Sustainable Development of Power Market *AISS* **4**(10) 269-77 *(in Chinese)*

[13] NIU Dong-xiao, LV Hai-tao, ZHANG Yun-yun 2008 Combination method of mid-long term load forecasting based on support vector machine within the Bayesian evidence framework *Journal of North China Electric Power University* **2008**(6) 62-6 *(in Chinese)*

Authors

Song Qiang, born in Xintai city of Shandong province

Current position, grades: Associate professor of Anyang Institute of Technology.
University studies: Kunming University of science and technology in 2006.
Scientific interest: areas of intelligence control theory and rough set theory.

Wang Ai-min, born in 1957

Current position, grades: professors of Anyang normal college.
University studies: doctor's degree from Wuhan institute of technology.
Scientific interest: SVM and rough set of intelligent decision-making research.

Discussion on determination method of characteristic stress of Jinping marble under confining pressure condition

Jinglong Li[1]*, Bin Sui[2]

[1,2]*School of Civil Engineering Shandong University, Jinan, Shandong Provence, China*

[2]*JianBang Co. LTD, China*

Received 1 March 2014, www.tsi.lv

Abstract

The characteristic stress is coincident well with the internal crack propagation in brittle rock. The characteristic stress are separately called closure stress, cracking stress, damaging stress and peak stress according to the internal crack state in loading. The propagation and damage extent in brittle rock can be reflected. Limited by loading testing equipment, the characteristic stress in confining pressure condition cannot be determined in China. In order to confirm the stress, the strain curves under different confining pressure condition are used to analysis the problem. The results show that the closure stress, cracking stress and damaging stress can be accurately confirmed by this method. The characteristic stress relates to the confining pressure, and the relationship is approximately linear.

Keywords: brittle rock, characteristic stress, marble, confining pressure

1 Introduction

The deformation characteristics and failure mechanism of brittle rock has received widespread attention. After decades of development, in particular the progress of testing technology, gaining a new awareness of the cracking character in brittle rock [1-4]. According to different state of compaction, propagation, connection and perforation of crack rock under different stress levels, the stress-strain curve of rock can generally be divided into four stages: I Compaction stage; II Linear elastic stage; III The stable crack development stage; IV The unstable crack development stage Figure 1.

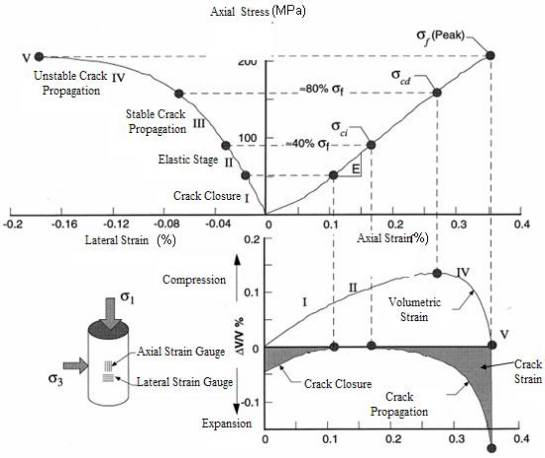


FIGURE 1 Schematic diagram of each stage in the process of uniaxial compression for granite [1]

(1) Compaction stage. Rock crack (including the crack of original and unloading) is compacted under axial compression. Stress-strain curve shows concave upward and nonlinear deformation. Axial pressure σcc corresponds to the minimum pressure of the crack with completely compaction.

(2) Elastic stage. The axial stress and axial strain is approximately linear. Deformation is mainly for the elastic deformation, but also contains a small amount of unrecoverable plastic deformation. The stress-strain relationship approximately obeys Hooke's law. In this stage, the diastrophism between micro fractures can be restrained by friction between closed fractures. Therefore, deformation is mainly elastic.

(3) The stable crack development stage. After The axial stress reaching the splitting strength σci, internal rock began to appear micro cracks, namely, began to appear rupture phenomenon. At this level of stress is about 40% of the peak intensity of rock.

(4) The unstable crack development stage. Axial pressure continues to increase to the intensity of damage σcd (about 80% of the uniaxial compressive strength of rock).Crack propagation way of rock starts into the unstable stage. The corresponding stress level is called damage strength. In this stage, the volume of crack formation and propagation formation is over the elastic deformation formed by compressive stress. Then the dilatancy began to appear.

Thus, it can be seen, the characteristic stress of brittle rock are closely linked with their internal crack propagation. In the whole process, volumetric strain, crack volumetric strain and sound emission divided the

* ***Corresponding author*** e-mail 215549020@qq.com

whole stress-strain curve into different stages. And these stages has a great significance in indicating the rupture process. However, due to the limit of current experimental technology and theory, currently China does not have the monitoring technology of sound emission under the condition of confining pressure, generally direct fixing the sound emission probe on the push rod of MTS. However, the effect is not good. Results of sound emission signals are rare. And it cannot correspond to the stress-strain curve well. Therefore, this paper tries to analyse the changing law of the rock stress-strain curve under different confining pressure to determine the characteristic stress under confining pressure

2 Closure stress and cracking stress

According to the existing research results, subtracting the elastic volumetric strain εev from the volumetric strain εv can get the crack volumetric strain εcv curve which could reflect the process of crack closure and crack opening in the process of loading. For testing of volumetric strain of rock specimen cannot be measured directly. However, the rock specimen is generally cylindrical, according to the assumption of small strain, volumetric strain can be calculated according to following formula:

$$\varepsilon_v = \frac{\Delta V}{V} = \varepsilon_1 + 2\varepsilon_2 \ , \qquad (1)$$

where ε1 is axial strain, ε_2 is lateral strain, ε_v is the total volumetric strain.

Using formula (2) calculates crack volumetric strain:

$$\varepsilon_{cv} = \varepsilon_v - \varepsilon_{ev} \, , \qquad (2)$$

where εev is elastic volumetric strain, $\varepsilon_{ev} = \frac{1-2\nu}{E}(\sigma_1 + 2\sigma_3)$.

In the first phase, the opening crack is gradually compacted and crack volumetric strain gradually increases. In the end of first phase, crack volumetric strain reaches the maximum, as the most crack is airtight closing. In this case, the corresponding axial stress is the closure stress. In elastic stage, as the compacted cracks have not yet appeared relative sliding and the new cracks have not yet been generated. Therefore, in this stage the total volume strain increment is equal to the elastic volume strain increment. Crack volumetric strain remains the same and the curve remains level standard. After the rock specimen entering into the stable crack development stage, because of the existing crack propagation and new crack generation, the total volume strain increment is less than the elastic volume strain increment. It leads the curve offsetting to the negative direction. Whereupon, there must be a point of inflection between the elastic stage and the crack propagation stage that the curve represents. This point of inflection corresponded to the axial stress is the cracking stress. Under no circumstances with sound emission, the research of changing law of closure stress and cracking stress under different confining pressure that can use the above method to analyse compacted testing results of triaxial test. Try to arrange triaxial test data using the above method, as shown in Figure 2.

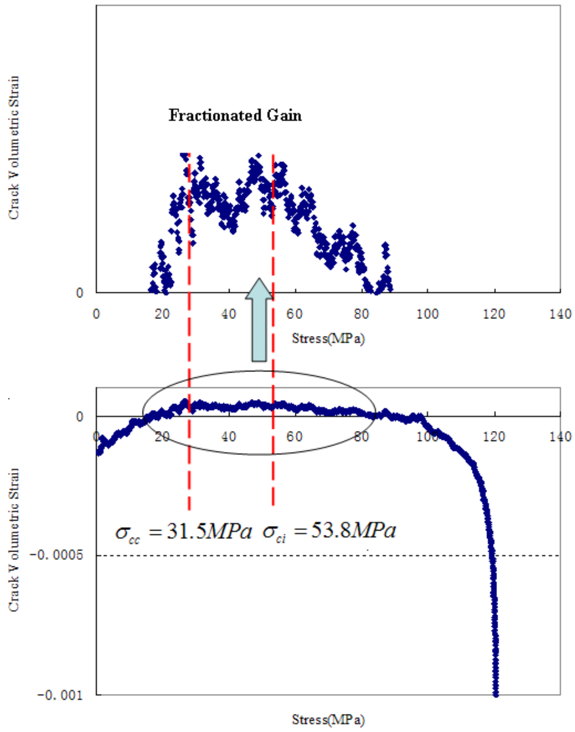


FIGURE 2 The closure stress and cracking stress under D12-28MPa confining pressure

3 Damaging stress

Damaging stress σcd corresponds to the stress that volume strain just began to recover. Therefore, the key to determine σcd is to accurately determine the recovering point of volumetric strain. Figure 5 shows the volumetric strain-axial strain curve. If this curve is used directly, the point is determined with greater randomness. In order to accurately determine the volume strain recovering point, Eberhard [5] proposed to use the relative volumetric strain stiffness method, in which the zero point corresponds to volume stiffness point is the volumetric strain recovering point. Volume stiffness is defined in Figure 3 and the specific steps are as follows: (1) The calculation shows that relative volume stiffness curve can accurately reflect the variation trend of volumetric strain curve when the number of data points between A1 and A2 are between 5% to 10% of the total number of data points.

(2) The least square method is used to fit the data (ε1, ε2) between A1 and A2 and the slope of the resulting line is the relative volume stiffness value corresponding to point A.

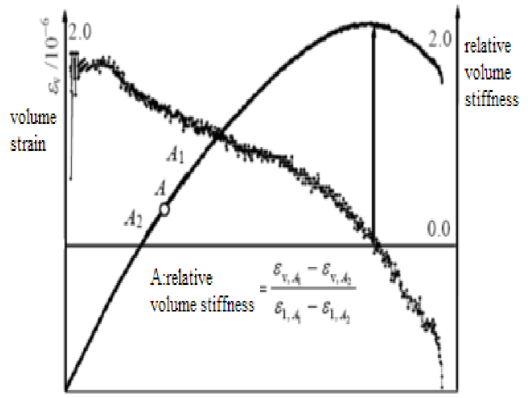

$\varepsilon_v/10^{-6}$
2.0
volume strain
2.0
relative volume stiffness
A_1
A
A_2
0.0
A:relative volume stiffness $=\dfrac{\varepsilon_{v,A_1}-\varepsilon_{v,A_2}}{\varepsilon_{1,A_1}-\varepsilon_{1,A_2}}$

FIGURE 3 Schematic diagram of determining the volume strain recovery points

(3) Relative volume stiffness can be gained by calculating each point based on step (2). When the relative volume stiffness is zero, that point is the volumetric strain recovering point.

Following the steps above, this method can be applied to determine the damaging stress under 8MPa confining pressure. The final result is shown in Figure 4.When the damaging stress reaches 102MPa, the peak strength is 120MPa. The ratio between damaging stress and peak strength is about 85%.

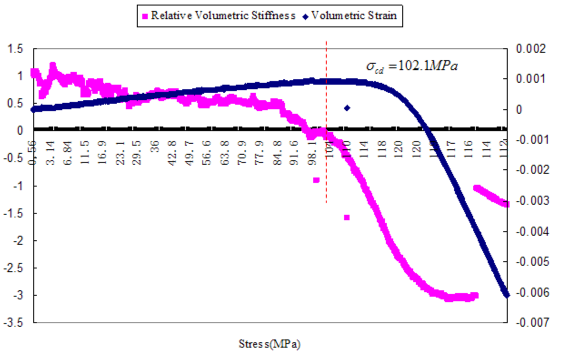


FIGURE 4 Damaging stress under the condition of 8MPa confining pressure

4 Characteristic stress under different confining pressure

Characteristic stress under different confining pressure shown in table 1 can be gained by the experiment curve combining with the method above. The result shows that σcc/σf under homotaxial condition is greater than σcc/σf under triaxial condition. Probably because the peak strength is affected by confining pressure and the crack is closed under confining pressure, the closure crack tends to be small. Significant linear relationship between each characteristic stress and confining pressure can be seen in Figure 5.

Figure 6 shows the fitting of each characteristic stress value based on Hoek-Brown criterion. The experimental data curve is moving closer to the Hoek-Brown curve with the transition from closure strength to cracking strength to damage strength to the peak strength. That is the Hoek-Brown failure criterion can be used to forecast the stress failure condition of rock. However, the development process of damage, which is the development stage of the crack, cannot be judged with Hoek-Brown criterion.

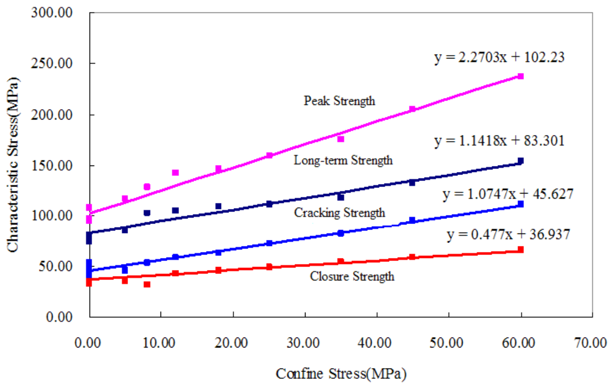
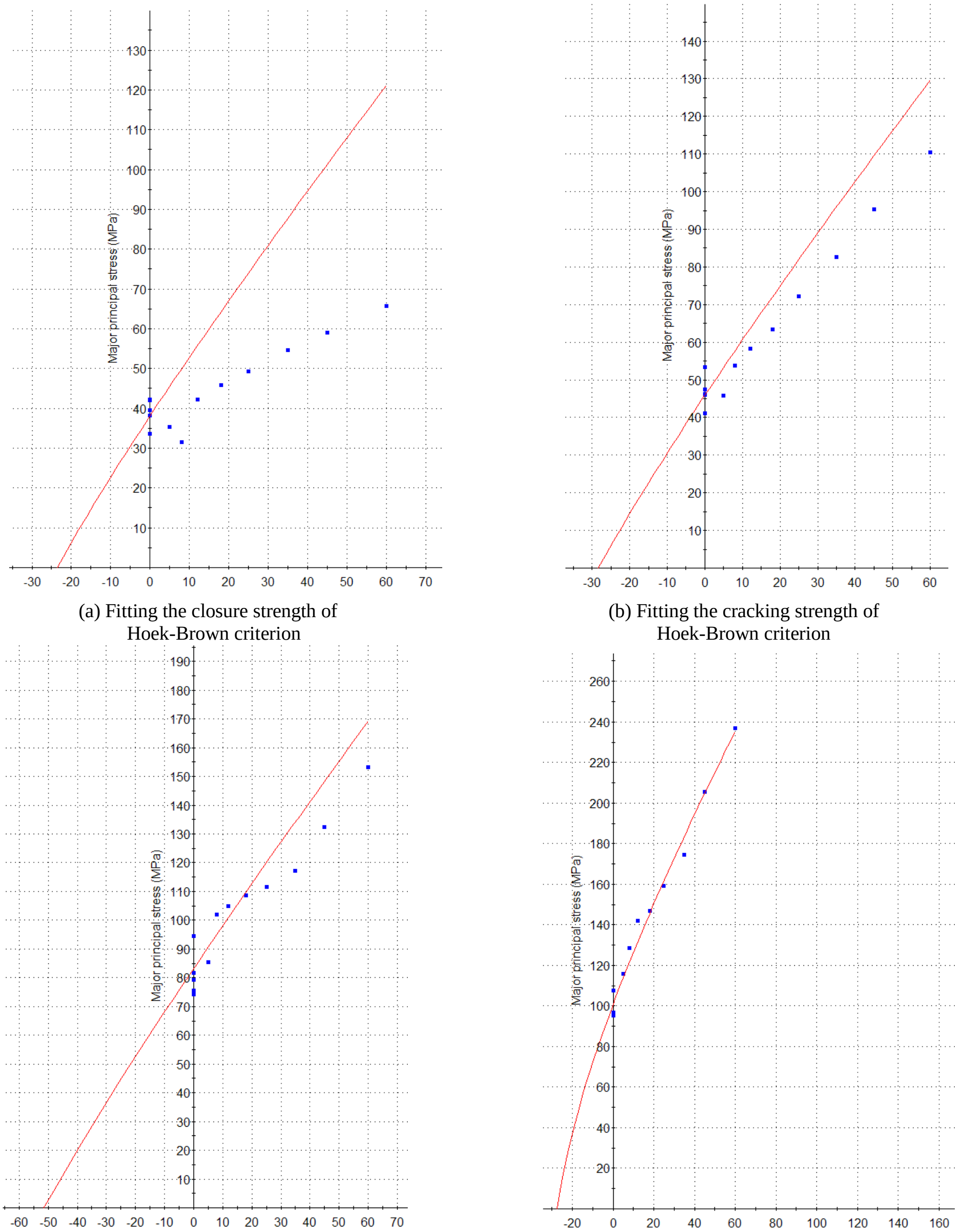


FIGURE 5 Characteristic stress with changing curve under confining pressure

TABLE 1 Characteristic stress under different confining pressure

Rock specimen number	σ3	σf	σcc	σci	σcd	σcc / σf	σci / σf	σcd / σf
D25	0.00	95.20	33.60	41.20	75.50	0.35	0.43	0.79
D28	0.00	107.60	39.50	53.40	94.50	0.37	0.50	0.88
D30	0.00	96.50	42.30	47.50	74.20	0.44	0.49	0.77
D31	0.00	95.80	42.00	46.10	81.60	0.44	0.48	0.85
D35	0.00	96.90	38.20	46.30	79.20	0.39	0.48	0.82
D12-1	5.00	115.90	35.40	45.90	85.50	0.31	0.40	0.74
D12-2	8.00	128.50	31.50	53.80	102.10	0.25	0.42	0.79
D14-1	12.00	142.10	42.20	58.20	104.90	0.30	0.41	0.74
D18-1	18.00	146.70	45.80	63.50	108.70	0.31	0.43	0.74
D17-2	25.00	159.10	49.30	72.30	111.50	0.31	0.45	0.70
D18-3	35.00	174.60	54.70	82.60	117.10	0.31	0.47	0.67
D17-3	45.00	205.40	59.10	95.40	132.30	0.29	0.46	0.64
D18-2	60.00	236.90	65.80	110.50	153.30	0.28	0.47	0.65



(a) Fitting the closure strength of Hoek-Brown criterion

(b) Fitting the cracking strength of Hoek-Brown criterion

(c) Fitting the damaging strength of Hoek-Brown criterion

(d) Fitting the peak strength of Hoek-Brown criterion

FIGURE 6 Fitting the characteristic strength of Hoek-Brown criterion

5 Conclusion

Characteristic stress can reflect damage degree inside the rock. Based on the triaxial compression testing curve of Jinping marble, four characteristic stress that is closure stress, cracking stress, damaging stress and peak stress can be defined to reflect the extended state of rock internal cracks according to the internal cracks extending state of rock specimen in the loading process, where closure stress, cracking stress, damaging stress and peak stress can be obtained by stress-strain curve combining with regression technology. The results are well consistent with the existing test results.

Acknowledgments

This work was supported by Province Fund (ZR2013EEQ023),and the Open Project of State Key Laboratory Breeding Base for Mining Disaster Prevention and Control(Shandong University of Science and Technology) (NO.MDPC2012KF07)

References

[1] Martin C D, Chandler N A 1994 The progressive fracture of Lac du Bonnet granite *International Journal of Rock Mechanics and Mining Sciences* **31** 643–59

[2] Martin C D 1993 *The strength of massive Lac du Bonnet granite around underground opening* University of Manitoba, Winnipeg

[3] Lau J S O, Chandler N A 2004 Innovative laboratory testing *International Journal of Rock Mechanics and Mining Sciences* **41** 1427-45

[4] Bieniawski Z T 1967 Mechanism of brittle fracture of rock, Parts I, II and III *International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts* **4**(4) 395–430

[5] Eberhardt E, Stead D, Stimpson B 1996 Quantifying progressive pre-peak brittle fracture damage in rock during uniaxial compression *International Journal of Rock Mechanics & Mining Sciences* **39**(3) 361-80

[6] Jinglong Li 2009 Application of multi-factor fuzzy comprehensive evaluation in stability of surrounding rock of underground engineering *Sensors &Transducers* **16**(2) 269-76

[7] Most T, Knabe T 2010 Reliability analysis of the bearing failure problem considering uncertain stochastic parameters *Computers and Geotechnics* **37**(3) 299-310

[8] Budzier A 2011 The risk of risk registers managing risk is managing discourse not tools *Journal of Information Technology* **26**(4) 274-6

[9] LIU Ning, ZHANG Chunsheng, CHU Weijiang, Fracture characteristics and damage evolution law of jinping deep marble *Chinese Journal of Rock Mechanics and Engineering* **31**(8) 1606-13

[10] HOEK E 2005 *Rock engineering Amsterdam* Netherlands: A. A. Balkema Publishers, 161–202

Authors

	Jinglong Li, born in November, 1980, in Shandong Province, China **Current position:** Doctor of engineering, working in civil and hydraulic engineering, Shandong University **Experience:** underground engineering stability analysis and water inrush mechanism and Prevention
	Bin Sui, born in February, 1981, Shandong Province, China **Current position:** Doctor of engineering. Majored in mechanical engineering. Now working in School of Civil Engineering, Shandong University **Experience:** stability analysis, numerical simulation and test of underground engineering.

High resolution photoacoustic system based on acoustic lens and photoacoustic sensors array

Jianning Han[1]*, Tingdun Wen[2], Peng Yang[1], Lu Zhang[1]

[1]*School of Information and Communication Engineering, North University of China, 3 Xue Yuan Road, Taiyuan, China*

[2]*Key Laboratory of Electronic Testing Technology, North University of China, 3 Xue Yuan Road, Taiyuan, China*

Received 1 March 2014, www.tsi.lv

Abstract

Photoacoustic tomography is a nondestructive bio-photonic imaging method based on the differences of optical absorption within biological organization. An approach using the lens with negative refractive index and photoacoustic sensors array to make the evanescent wave involved in the imaging process was presented in this paper. A set of comparative experiments was demonstrated on the imaging effect between the ordinary lens and the lens designed in this work. The experiment showed that the imaging effect of photoacoustic tomography by the designed lens had greatly outperformed the ordinary lens. In order to illustrate the good results, according to the characteristics of ultrasonic waves produced in photoacoustic effect, the propagation properties of the acoustic waves in lens with different refractive index was discussed. On the basis of analysing evanescent decay of ordinary acoustic lens which results in the loss of high-frequency information with image details in current photo-acoustic tomography system, the diffraction limit of was broken through and the image resolution was greatly improved by the lens with negative refractive index in theory.

Keywords: Photoacoustic Tomography, Acoustic Lens, Negative Refraction, Image Resolution

1 Introduction

With more and more technology development in computers, material, and industries, a considerable progress has been achieved on the photoacoustic tomography in medical ultrasound imaging. Recently, an earthshaking change has been made on new methods, techniques, and materials in the photoacoustic tomography, which also paves a way for the commercial applications of the photoacoustic tomography. The photoacoustic effect explains how electromagnetic energy can be absorbed and converted into acoustic waves. The photoacoustic tomography benefits from the advantages of pure optical or ultrasound imaging, without the major disadvantages of each technique. Especially its importance of studying and popularization is shown in the diagnosis on the early cancer conducted by Wang's group [1]. Since Veselago from the former Soviet Union proposed the concept of left-handed material [2] (negative refraction material) in 1968, the enthusiasm of scientists in the study of negative refraction materials has not been reduced. Especially in 2000, J.B. Pendry published an article named Negative Refraction Makes a Perfect Lens [3] on the Phys. Rev. Lett, which induced a series of discussions on the topic of negative refraction materials. Negative refraction acoustic lens are just the one kind of the lens compounded of different mediums. These mediums are compounded according to a certain rule, which share the characteristics of the negative refraction [4]. Moreover, this kind of lens has some merits, e.g., focusing, band gap, direction propagation, and etc. In addition, the characteristics of this kind of lens, e.g., focusing, filtering and directional control on acoustic wave, are very suitable for the improvement of the inherent disadvantages of the photoacoustic tomography.

Moreover, there are some typical experimental applications using photoacoustic tomography technology. For example, Ronald E. Kumon et al from the Department of Biomedical Engineering, University of Michigan @ Ann Arbor Campus, used the photoacoustic tomography technology to make a frequency-domain analysis on the prostate cancer tumor in the body of white rat, and achieved some good findings [5]. Xueding Wang et al from the Department of Radiological Sciences, University of Michigan @ Ann Arbor Campus, used the commercial imaging equipment combining with the detector developed by themselves to make images on a double of human hair and the treelike vascular of rabbit's ear, and acquired some quite clear images [6]. Hui Wang et al, who worked in the Key Laboratory of Laser Life Science, Ministry of Education, South China Normal University, used Fresnel zone plate ultrasonic detector to make images, and realized photoacoustic tomography [7]. Beside the examples above, some correlative studies on the technology of photoacoustic tomography were conducted in other research institutions as well. Overall, they mainly conducted their researches in two aspects:

[*] ***Corresponding author*** e-mail: hanjn46@nuc.edu.cn

one is to use new materials, or to make the sensors array to ameliorate the receiving sensor, and the other is to use the new filtering and rebuilding algorithm of image to make a post-processing on the images. However, all of the studies did not solve the problems caused by the technology of photoacoustic tomography described above. Specifically, the spatial resolution of imaging is not high, and some tiny organizations still cannot be distinguished. In addition, when the imaging experiment is conducted on the living body, the quality of imaging is greatly affected by the complicated noise. Up to now, many researchers use acoustic lens to realize photoacoustic tomography [8-10], but few of them conduct on the application of negative refraction acoustic lens in photoacoustic tomography system.

In this work, with an analysis on the decay of the evanescent wave in the ordinary acoustic lens in current photoacoustic tomography system, an approach by using the lens with negative refractive index and photoacoustic sensors array to make the evanescent wave participate in the imaging process was proposed. In order to improve the resolution of images greatly, our research attempts to break through the diffraction limit of the ordinary acoustic lens. At the same time, the corresponding theory of acoustic information system was also investigated for better understanding of the performance.

2 The experimental Method

The acoustic lens photo-acoustic experiment system was built as shown in Figure 1. In order to distinguish different experiment effects of the lens, two kinds of refraction index lens to implement the experiment were chose. Four lasers YAG produced by the American company (Quanta-Ray PIV, Spectrum Physics) were used. The pulse repetition frequency is 25 KHz, pulse width is 10 ns, the wavelength is 1064 nm, and the incident light spot diameter is 1 mm. The sample used in the experiment was an artificial production with test panel of calibration. Panel was placed on the horsehair to show the different distances to display the maximum resolution of imaging system, which could be distinguished. The matrix surrounding sample box was liquid oil.

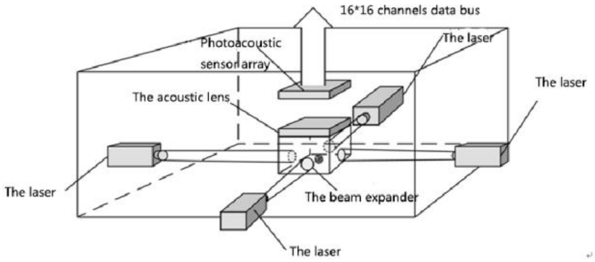


FIGURE 1 Experimental system of acoustic lens photoacoustic tomography

The acoustic lens was put on the sample box, and the acoustic sensor array was placed on the focus plane of the lens. To distinguish the contrast effects, two acoustic lens that one was the ordinary acoustic lens made of resin materials, and the other was negative refraction acoustic lens made of coat rubber layer embedded in epoxy resin were used [11-12]. Receiving array was the 16 × 16 micro-nano ultrasonic sensor, then connected to the data acquisition circuit after 16 × 16 channels data bus was connected.

Experimentation:

(1) Put the sample in the test platform of frosted glass vessels.

(2) Turn on the laser, through the beam expander and frosted glass beam expander, make the laser on the samples uniformly.

(3) The samples facing by laser irradiation will be heated up and expanded, then it will irradiate the acoustic wave. The acoustic wave through the acoustic lens (general lens and negative refraction lens) will focus on focal plane of photo-acoustic sensor array composed of 256 elements of 16×16.

(4) Photo-acoustic sensors make the acoustic signals of sample change into electrical signals. They get the analog signal through the filter amplifier signal conditioning, then the analog signals are converted to the digital signals. According to the principle of the lens, the two-dimensional image model of the hair.

(5) Transmit the characteristics of ultrasonic signal in the time and amplitude domain of each photo-acoustic sensor receiving into RAM, and combine with spatial information of acoustic lens imaging model, the final result of photo-acoustic tomography data can be got.

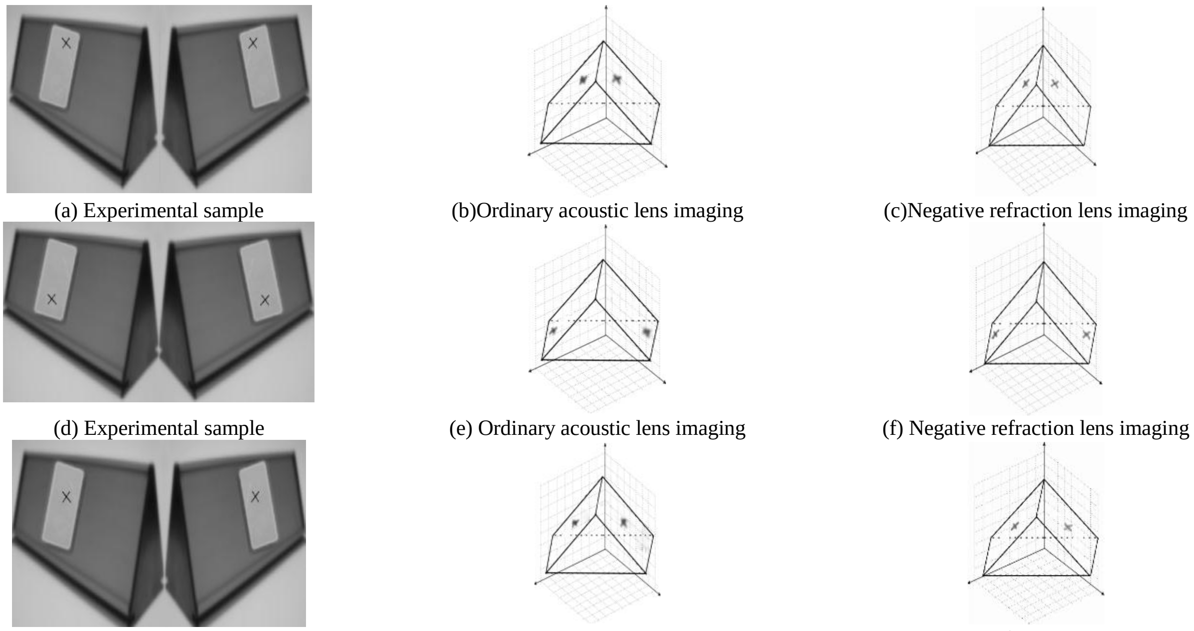
(6) After image post-processing, the final photo-acoustic tomography results can be obtained.

3 The experimental result and analysis

To examine the image resolution, a phantom sample with a cross of two-horse hair (65 μm diameter) in a block of transparent gelatine was made. This block of transparent gelatine is a rectangular thin plate with size of 4 mm × 1 mm. The block of transparent gelatine is placed on the tripod. Acoustic signals of the cross of two horse hair are generated by the laser irradiation, and the hair can be shown in the image after reconstruction. To understand the nature of the resolution improvement, photoacoustic imaging by the hair using different acoustic lens (ordinary lens and negative refractive index lens) was performed.

The experimental sample in Figure 2(a) shows that cross-shaped hair is on the top. The hair is irradiated by laser, and its imaging effect through ordinary lens is shown in Figure 2(b). The image is ambiguous, and the cross-shaped of the hair even cannot be identified. In order to distinguish imaging effect, negative refraction lens was used. If the hair is through laser's irradiating ultrasonic, its imaging effect through negative refraction lens is shown in Figure 2(c). The image is very clear and the cross-shaped of the hair is clearly visible. The experimental sample in Figure 2(d) shows that the cross-shaped hair is at the bottom. In Figure 2(e), the image through ordinary lens is fuzzy. From Figure 2(f), the image through negative refraction lens is very clear. In order to analyse the image resolution, we defined the resolution of an image as the width of the main lobe-crossing zero minus the width of the imaged hair. Here, we checked the image on the cross section of the hair. The photoacoustic images of the horsehair illuminated with the ordinary lens are shown in Figure 2(b) and 2(e). Here, the measured photoacoustic FWHM of the hair is found to be 235 μm on average.

Photoacoustic imaging with the negative refractive index lens provides both the highest SNR and the best resolution as shown in Figure 2(c) and 2(f). The measured FWHM of the horsehair decreases to 76 μm. In order to further verify the experimental results, we repeated the experiment in different positions. The experimental sample in Figure 2(g) shows that cross-shaped hair is in the centre position. Figure 2(h) and 2(i) show similar experimental results.


(a) Experimental sample (b)Ordinary acoustic lens imaging (c)Negative refraction lens imaging

(d) Experimental sample (e) Ordinary acoustic lens imaging (f) Negative refraction lens imaging

(g) Experimental sample (h)Ordinary acoustic lens imaging (i)Negative refraction lens imaging

FIGURE 2 Imaging results of Photoacoustic imaging system with different refractive index lens

According to the above experimental data, the measured photoacoustic FWHM of the hair the data was analysed in Table1. The resolution increases by three times compared to the acoustic resolution of the ordinary lens. Therefore, the photoacoustic tomography system with negative refractive index lens generally provides better spatial resolution, which is consistent with the theoretical analysis of the frequency dependence of spatial resolution.

TABLE 1 The measured photoacoustic FWHM of the hair

Different locations	**Ordinary acoustic lens (μm)**		**Negative refraction lens (μm)**	
	Left side	Right side	Left side	Right side
The hair on the top	234.81	234.61	76.14	76.18
The hair in the centre position	235.41	234.71	76.41	76.36
The hair at the bottom	234.95	234.82	76.48	76.42

Summarizing the above experiment images and combining with lens imaging knowledge, the limitation of the ordinary lens imaging system resolution leads to a virtual image of centring on the absorber and spreading outward. It makes the ambiguous intersection obviously. However, the image around the concentrated absorber (hair intersection) in negative refraction lens is very clear, it is easy to be identified and the image is consistent with the original sample. The result demonstrates perfectly that negative refraction lens is superior to the ordinary lens in terms of imaging resolution.

4 Discussion

4.1 ESTABLISHMENT OF THE DIFFERENT REFRACTIVE INDEX ACOUSTIC LENS MODEL

In order to describe the imaging resolution of different lens, the propagation of acoustic waves in different lens was analysed using COMSOL Multiphysics software. As shown in Figure 3, the nanosecond pulse laser irradiation is generally applied into the current photoacoustic tomography system to make an exposure on the detection area. When the detection is started, the detection area is expanded because of the heat of the laser. The coefficients of photoacoustic transition of the diseased and normal organization are different, which directly result in the difference between the thermal expansions of the media. Thus, different acoustic signals are generated in the two organizations. In addition, the signals can be received in the form of ultrasonic waves, and the distribution of the target can be shown in the image after reconstruction. Therefore, it can be concluded that the key to photoacoustic tomography is how to detect the ultrasonic signals. For the convenience, herein, the point radiating acoustic wave under the exposure of laser is assumed as point source reasonably in acoustic research, and the core of the study is how to detect the imaging information of the point source with high quality. Currently, some researchers have initiated the study on ordinary acoustic lens. To be more specific, as shown in Figure 4, the point source on the left side of the lens radiates waves under the exposure of laser, and the waves are gathered at the focal point on the right side of the lens using COMSOL Multiphysics software. In this work, the application of the negative refraction lens into the photoacoustic tomography system is put forward. The point source on the left side of the lens radiates waves under the exposure of laser, and the waves are gathered on the right side of the lens using COMSOL Multiphysics software, as shown in Figure 5.

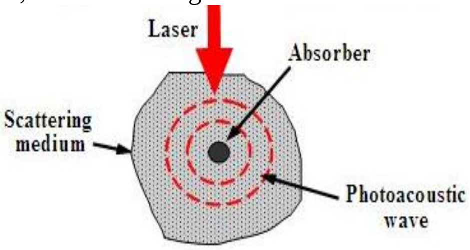


FIGURE 3 Thermal expansion of biological organization

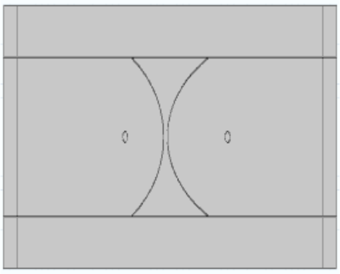


FIGURE 4 Diagrammatic sketch for the imaging of ordinary lens

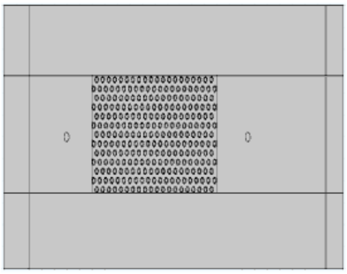


FIGURE 5 Diagrammatic sketch for the imaging of negative refraction lens

4.2 THEORETICAL SIMULATION ANALYSIS AND DISCUSSION

As a kind of ultrasonic waves excited by light, photoacoustic signal has all the characteristics of the waves substantially. Therefore, provided that, (1) a pulsed laser with uniform distribution (pulse width for the magnitude) is used to make exposure on the biological organization, and the dimension of photoacoustic signal generated by each point in the biological organization is proportional to the light absorption coefficient of the point in biological organization, and (2) each photoacoustic signal generated by the points in the biological organization is a point source with amplitude decayed by 1/r, and the function of the wave generated by the point source is $\varphi = Ue^{j\omega t}$, furthermore, in this function, U represents the complex amplitude of sound field [13-15]. When the acoustic wave is emitted onto the plane (x, y), there is a component for the wave propagated along the direction Z. According to the scalar wave equation of Helmholtz $\nabla^2 U + k^2 U = 0$, when the distance of propagation along the direction z of the wave is short enough, the solution to the wave equation can be expressed in the form of:

$$A(k_x, k_y, z_0) = A(k_x, k_y, z_0)e^{jk_z z_0}\ , \quad (1)$$

where $A(k_x, k_y, z_0)$ is the spatial spectrum distribution of the plane. Considering the sound velocity field v in front of the acoustic lens, the components of the field will be given by some 2D Fourier expansion

$$v(r,t) = \sum v(k_x, k_y)\exp(ik_x x + ik_y y + ik_z z - i\omega t)\ . \quad (2)$$

Considering that the wave is propagated in the three-dimensional space, according to the relationship among k_x, k_y, and k_z the expressions k_x, k_y, k_z can be used to express the wave vector of free space. Therefore, the equation $k = \pm\sqrt{k_0^2 - (k_x^2 + k_y^2)}$ can be obtained, where $k_0 = 2\pi/\lambda$.

In the same time, considering that the index of refraction of the lens is either positive or negative, the equation

$$k_z = \pm\sqrt{k_0{}^2 - (k_x{}^2 + k_y{}^2)} = \sqrt{n^2\omega^2/c^2 - (k_x{}^2 + k_y{}^2)}$$

can be deduced, where is positive in the positive refraction lens and negative in the negative refraction lens.

When $k_x{}^2 + k_y{}^2 > k_0{}^2$, so $k_z = j\sqrt{(k_x{}^2 + k_y{}^2) - k_0{}^2} = j\zeta$ (pure imaginary number). At present, the is a pure imaginary number, so

$A(k_x, k_y, z_0) = A(k_x, k_y, 0)e^{-\varsigma z_0}$, it leads to the resolution's limit of the lens $\Delta = 2\pi/k_{\max} = 2\pi c/\omega = \lambda$.

Obviously, when an acoustic wave propagates in the positive index lenses, wave vector k_z is changed into an imaginary number. The resolution limit is a wavelength. Although the radiation wave field phase of body surface carrying more fine structure information remains unchanged in the direction of propagation, the amplitude will have an exponential attenuation. This part carries high surface fine structure information of evanescent attenuation called evanescent wave. In Figure 6, it can be observed that the attenuation wave in positive refractive index lens. Therefore, the information carried fine structure cannot be propagated to the far field, and they are limited in the near-field area closed to the lens. It leads to the low resolution of ordinary lens, and it cannot meet the needs of photo-acoustic tomography. The decay of the evanescent wave in the ordinary acoustic lens results in the loss of the high-frequency information containing image details, so the effect of the imaging is not ideal, as shown in Figure 7.

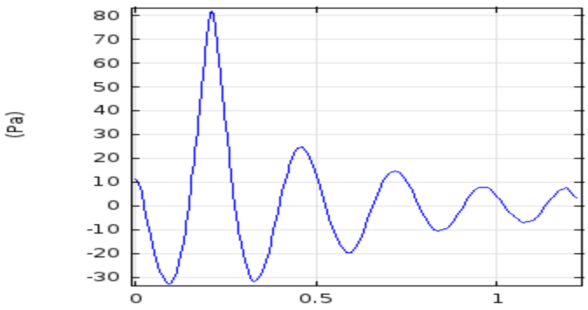


FIGURE 6 The evanescent wave of ordinary lens

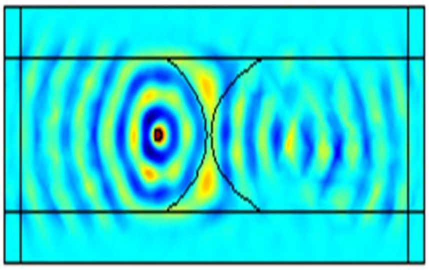


FIGURE 7 Diagrammatic Sketch for the Imaging of Ordinary Lens

If the transmission medium in the opposite direction was considered, when n<0, then $k_x{}^2 + k_y{}^2 > k_0{}^2$, so $k_z = -j\sqrt{(k_x{}^2 + k_y{}^2) - k_0{}^2} = -j\zeta$ (pure imaginary number), the Formula (1) can be expressed as:

$$A(k_x, k_y, z_0) = A(k_x, k_y, 0)e^{-j^* j\xi_0} = A(k_x, k_y, 0)e^{\xi_0}. \quad (3)$$

Although the propagation constants are changed into the imaginary, but the amplitude will be changed e^{ξ_0} as exponential form, acoustic amplitude along the Z direction is not decayed as exponential form. Calculations confirm that all of the energy is perfectly transmitted into the medium but in a strange manner. Overall, the transmission coefficient of the medium is

$$T = tt' = \exp\left(ik_2^i d\right) = \exp\left(-i\sqrt{\omega^2 c^{-2} - k_x^2 - k_y^2 d}\right), \quad (4)$$

where *d* is the slab thickness. It is this phase reversal that enables the medium to refocus sound by cancelling the phase acquired by the sound wave as it moves away from its source. To calculate the transmission through both surfaces of the slab, the multiple scattering events must be summed,

$$T = tt' \exp(ik_z' d) + tt' r'^2 \exp(3ik_z' d) + tt' r'^4 \exp(5ik_z' d) + \ldots = \frac{tt' \exp(ik_z' d)}{1 - r'^2 \exp(2ik_z' d)}, \quad (5)$$

$$R = r + tt'r'e^{jk_z' d} + tt'r'^3 e^{j3k_z' d} + \cdots = r + \frac{tt'r'e^{j2k_z' d}}{1 - r'^2 e^{j2k_z' d}}, \quad (6)$$

where *t* and *r* are the transmission coefficient and reflection coefficient at the interface within the medium respectively, *t'* and *r'* are the transmission coefficient and reflection coefficient at the interface between the medium and the vacuum, respectively. When the refractive index of the lens is close to -1, we can get the formula

$$\lim_{n \to -1} T = \lim_{n \to -1} \frac{tt' \exp(ik_z' d)}{1 - r'^2 \exp(2ik_z' d)} = \exp(ik_z d), \quad (7)$$

$$\lim_{n \to -1} R = \lim_{n \to -1} r + \frac{tt'r'e^{j2k_z' d}}{1 - r'^2 e^{j2k_z' d}} = 0 \quad (8)$$

Therefore, the evanescent wave carrying surface fine structure information can retain well, it cannot be decayed just like the positive refractive index medium. We can use that transmission medium to break the diffraction limit when the refractive index is negative, so we can consider it to make the evanescent wave involve in the imaging. Then, the imaging is conducted by using the advantages of negative refraction lens, such as, focusing, filtering and directional control on acoustic wave.

Figure 8 shows that the evanescent wave in the flat lens is amplified exponentially, more and more high frequency component is able to transport, which can effectively compensate the exponential decay of evanescent wave in the water. So that it can be successfully transferred to the image plane and involved in the imaging, and the negative refraction lens in the photo-acoustic tomography was used. Figure 9 is a simplified model for the photo-acoustic tomography, and it is consistent with that in Figure 7, which shows that the model is reasonable.

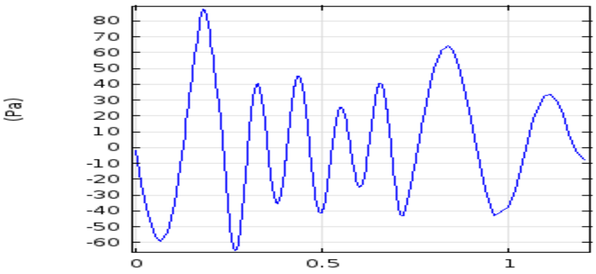


FIGURE 8 The enhanced evanescent wave of negative refraction lens

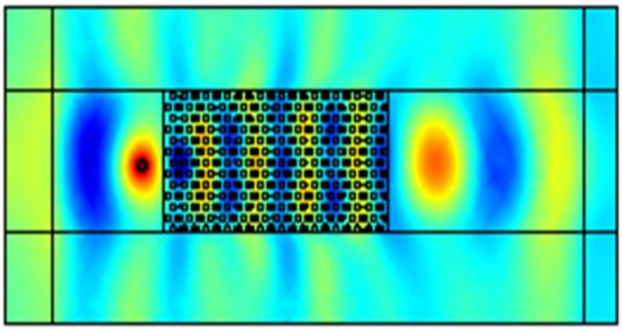


FIGURE 9 Amplified analog effect for the negative refractive imaging

In Figure 10 (a) and 10(b), according to the experimental device in this paper, the simulation environment was adjusted. The equivalent of experimental equipment illustrated the negative refractive index lens imaging effect. From the result of simulation, after the original point source through negative refraction lens, it converges into a clear image point on the right side of the lens. From the point in the waveform as shown in Figure 10(c), the resolution of the lens reaches 0.3 wavelength. So the scheme of the lens is a better design for the high resolution photo-acoustic tomography both theoretically and practically.

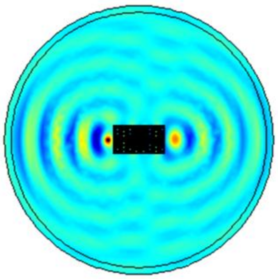


FIGURE 10 (a) Simulation effect of the negative refractive imaging

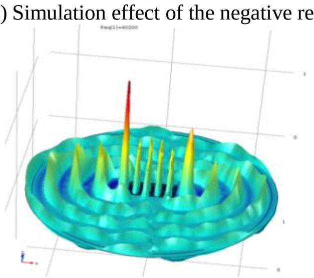


FIGURE 10 (b)3D simulation effect of the negative refractive imaging

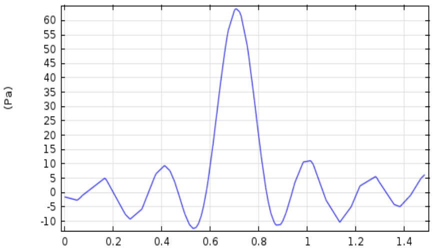


FIGURE 10 (c) Sound pressure distribution of the focus image point

5 Conclusions

In this paper, through lens imaging model, the imaging effects of the cross-shaped hair was analysed, and the best imaging effect through comparisons of many experiment samples was obtained. It proved that the imaging effect of negative refraction lens is better than that of normal acoustic lens. Then the phenomenon that the photo-acoustic tomography of negative refraction index lens can enhance the evanescent wave and the evanescent decay of ordinary acoustic lens with image details makes the resolution sharply were discussed . The enhanced evanescent wave makes the evanescent wave take part in the imaging, which greatly increases the resolution ratio of lens imaging. It also plays an important role in enhancing photo-acoustic tomography in the field of medical research and clinical examination. As for the weaknesses of our approach, the unit of sensor array was small, which limited the power of higher resolution ratio. In addition, normal image processing in image post-processing was used. If more advanced image processing techniques can be employed, better image resolution even can be achieved.

Acknowledgements

This project was sponsored in part by the National Natural Science Foundation of China (No.11304289 and 61302159).

References

[1] Pramanik M, Ku G, Li C, Wang L V 2008 Design and evaluation of a novel breast cancer detection system combining both thermoacoustic (TA) and photoacoustic (PA) tomography *Medical physics* **35**(6) 2218-23

[2] Veselago V G 1968 The electrodynamics of substances with simultaneously negative values of img align= ABSMIDDLE alt= eps/IMG AND μ *Physics-Uspekhi* **10**(4) 509-14

[3] Pendry J B 2000 Negative refraction makes a perfect lens *Physical review letters* **85**(18) 3966

[4] Torres-Silva H, Vivas-Hernández A, Guerrero-Moreno I J 2011 Chiral waves in a metamaterial medium *International Journal of Pure and Applied Sciences and Technology* **2**(2) 54-65

[5] Achilefu S, Dorshow R B, Bugaj J E, Rajagopalan R 2000 Novel receptor-targeted fluorescent contrast agents for in vivo tumor imaging *Investigative radiology* **35**(8) 479-85

[6] Wang X 2004 *Functional photoacoustic tomography of animal brains* (Doctoral dissertation, Texas A&M University)

[7] Wang H, Xing D, Xiang L 2008 Photoacoustic imaging using an ultrasonic Fresnel zone plate transducer *Journal of Physics D: Applied Physics* **41**(9) 095111

[8] Maslov K, Zhang H F, Hu S, Wang L V 2008 Optical-resolution photoacoustic microscopy for< i> in vivo</i> imaging of single capillaries *Optics letters* **33**(9) 929-31

[9] Wang L V, Hu S 2012 Photoacoustic tomography: in vivo imaging from organelles to organs *Science* **335**(6075) 1458-62

[10] Hu S, Maslov K, Tsytsarev V, Wang L V 2009 Functional transcranial brain imaging by optical-resolution photoacoustic microscopy *Journal of biomedical optics* **14**(4) 040503

[11] Liu Z, Zhang X, Mao Y, Zhu Y Y, Yang Z, Chan C T, Sheng P 2000 Locally resonant sonic materials *Science* **289**(5485) 1734-6

[12] Sheng P, Zhang X X, Liu Z, Chan C T 2003 Locally resonant sonic materials *Physica B: Condensed Matter* **338**(1) 201-5

[13] Pendry J 2003 Perfect cylindrical lenses *Optics Express* **11**(7) 755-60

[14] Hooft G W 2001 Comment on ”Negative Refraction Makes a Perfect Lens” *Physical review letters* **87**(24) 249701

[15] Pendry J B 2004 Negative refraction *Contemporary Physics* **45**(3) 191-202

Authors

JianNing Han, born in December, 1980, China County, Shanxi Province, P.R. China

Current position, grades: the Lecturer of School of Information and Communication Engineering, North University of China, China.
University studies: received his B.Sc. in Electronic information engineering from North University of China in China. He received his M.Sc. from North University of China in China.
Scientific interest: His research interest fields include Acoustic signal processing, Graphics processing
Publications: more than 7 papers published in various journals.
Experience: He has teaching experience of 11 years, has completed five scientific research projects.

TingDun Wen, born in November, 1957, China County, Shanxi Province, P.R. China

Current position, grades: the Professor School of Science, North University of China, China.
University studies: received his B.Sc. in Optoelectronic professional from Shanxi University in China. He received his M.Sc. from The National Technical University of Athens.
Scientific interest: His research interest fields include mesoscopic physics, sensor design
Publications: more than 40 papers published in various journals.
Experience: He has teaching experience of 30 years, has completed fifteen scientific research projects.

Peng Yang, born in July, 1988, China County, Shanxi Province, P.R. China

Current position, grades: the student of School of Information and Communication Engineering, North University of China, China.
University studies: received his B.Sc. in Electronic information engineering from North University of China. He received his M.Sc. from North University of China in China.
Scientific interest: His research interest fields include Acoustic signal processing,
Experience: He has researching experience of 2 years, has completed one scientific research projects.

Lu Zhang, born in July, 1988, China County, Shanxi Province, P.R. China

Current position, grades: the student of School of Information and Communication Engineering, North University of China, China.
University studies: received his B.Sc. in Electronic information engineering from North University of China in China. He received his M.Sc. from North University of China in China.
Scientific interest: His research interest fields include Acoustic signal processing,
Experience: She has researching experience of 1 years, has completed one scientific research projects.

Frequency domain minimum error probability medical CT image blind equalization algorithm

Yunshan Sun[1, 2], Liyi Zhang[1, 2*], Haiyan Zhang[1]

[1] *School of Electric Information Engineering, Tianjin University, Tianjin 300072, China*

[2] *School of Information Engineering, Tianjin University of Commerce, Tianjin 300134, China*

Received 1 March 2014, www.tsi.lv

Abstract

A frequency domain minimum error probability medical CT image blind equalization algorithm was proposed. Blind image equalization is implemented by minimizing a cost function consisting of estimated image and blur. The steepest descent method was adopted to solve the proposed cost function. Computer simulation experiments show that the new algorithm reduces mean square error and improves restoration effect, peak signal to noise ratio and improving signal to noise ratio.

Keywords: Blind equalization algorithm, minimum error probability, medical CT image

1 Introduction

Transform domain or complex value transform is an important research spot of blind image restoration. Such methods obtain image smooth domain, border characteristics and other information mainly through the transform domain or orthogonal equivalent conversion to be conducive to signal analysis and be helpful to simplify the analysis. The common transform include Fourier transform, Z transform, wavelet transform, Curvelet and Contourlet sparse decomposition, etc. A frequency domain iterative blind restoration algorithm was proposed in [1]. The approach involves processing the projection data with Bellini's method for attenuation compensation followed by an iterative deconvolution technique which uses the frequency distance principle (FDP) to model the distance-dependent camera blur. Modelling of the camera blur with the FDP allows an efficient implementation using fast Fourier transform (FFT) methods. Reference [2] proposed to improve convergence toward global minima by single-site updating on the wavelet domain. For this purpose, a new restricted DWT space is introduced and a theoretically sound updating mechanism is constructed on this subspace.

An efficient algorithm is proposed for blind image restoration based on the discrete periodic Radon transform (DPRT) in [3]. The discrete periodic Radon transform is utilized to transform two-dimensional image into one dimensional signal sequence, then image restoration can reduce storage space and improve work efficiency. Partial differential equation is a method commonly used in image restoration. In [4], the partial differential equation method combined with a sampling wavelet transform to improve the ability of image edge preservation. [5] utilized a sparse representation transformation to realize image fusion and recovery. The implement of blind image restoration is equal to the solution and update of the sparse coefficient. Compare with the traditional wavelet transform, discrete wavelet transform and curvelet transform, the method has better feasibility [6] deals with nonconvex nonsmooth minimization methods for image restoration and reconstruction. The main goal of this method is to develop fast minimization algorithms to solve the nonconvex nonsmooth minimization problem. This method proposed the largest characteristic that completes image regularization in the frequency domain. Fast Fourier Transform, Fast Fourier Transform and FFT, improve the working efficiency, and improve the recovery effects.

In this paper, we utilized frequency domain transformation to convert two-dimensional medical CT image into one dimensional signal sequence, and found a corresponding complex sequence cost function. The cost function was solved by optimization strategy and the optimal estimation of complex sequence was obtained. We made use of the corresponding inverse transformation to complete image restoration. This paper is organized as follows. A description of the proposed algorithm is given in Section 2, together with implementation details. The experimental section, Section 3, empirically validates the proposed method, and Section 4 concludes this paper.

2 Principle of Frequency Domain Minimum Error Probability Medical CT Image Blind Equalization Algorithm

In order to describe the dimension reduction medical CT blind equalization algorithm in the frequency domain. The signal model is shown as

* ***Corresponding author*** e-mail: sunyunshan@tjcu.edu.cn

$$G(k) = \mathrm{DFT}\left[g(n)\right] = \begin{cases} \sum_{n=0}^{N-1} g(n)\omega_N^{kn} & 0 \le k \le N\text{ -}1 \\ 0, & otherwise \end{cases}, \quad (1)$$

where, N is pixel point of image to be processed. According to the principle of the convolution of two signal sequences in time domain being equal to the product of their respective conversion in the frequency domain. Two sides of the image degradation equation was realized by FFT simultaneously. We obtain

$$G(k) = H(k)F(k) + N(k), \quad k = 0,1,\cdots,N-1, \quad (2)$$

where, $F(k)$ and $N(k)$ are the discrete Fourier transform of $f(n)$ and $n(n)$, respectively, $f(n)$ and $n(n)$ respectively represent real image sequence and noise sequence of the equivalent dimension reduction. $H(k)$ is the frequency response function of the point spread function (PSF) of the equivalent dimension reduction. In the other word, $H(k)$ is the discrete Fourier transform of $h(n)$.

Similarly, the eliminate form of dimension reduction point spread function can be represented as in the frequency domain

$$\tilde{F}(k) = G(\omega)W(\omega), \quad (3)$$

where, $W(\omega)$ denotes a frequency response function of blind equalizer, and it is the discrete Fourier transform of $w(n)$, namely $W(\omega) = \mathrm{DFT}\left[w(n)\right]$.

In order to establish a suitable cost function of blind equalization algorithm for medical CT image, here, medical CT images were simply pre – whitened and the transform of dimension reduction medical CT image is a smooth white sequence. Noise sample sequence is independent of the source signal and is a white noise sequence, thus we can get

$$\mathrm{E}\left[F(k)\right] = \mathrm{E}\left[N(k)\right] = 0, \quad (4)$$

$$\mathrm{E}\left[F_n(\omega)F_m(\omega)\right] = 0, n \ne m, \quad (5)$$

$$\mathrm{E}\left[F_n(\omega)N_n(\omega)\right] = 0, \quad (6)$$

$$\mathrm{E}\left[N_n(\omega)N_m(\omega)\right] = N\sigma_n^2\delta(n-m), \quad (7)$$

$$\mathrm{E}\left[F_n(\omega)F_m^*(\omega)\right] = 2A\delta(n-m). \quad (8)$$

According to image acquisition and restoration process, we know that

$$G(\omega) = H(\omega)F(\omega) + N(\omega). \quad (9)$$

By equation (3) and (9), we can derive the total characteristics $H_{\sum}(\omega)$ the equalization system

$$H_{\sum}(\omega) = \frac{\tilde{F}(\omega)}{F(\omega)} = \frac{1}{2A}W(\omega)\mathrm{E}\left[G(\omega)F^*(\omega)\right]. \quad (10)$$

In order to get the best estimate of the equalizer, namely $\hat{F}(\omega) = F(\omega)$, the frequency characteristics of the equalizer must achieve $W_{opt}(\omega)$, then

$$\begin{aligned}\tilde{F}(\omega) &= G(\omega)W_{opt}(\omega) \\ &= F(\omega)H_{\sum}(\omega)\end{aligned}. \quad (11)$$

Blind equalization cost function is defined as

$$\begin{aligned} &J(W^{(n)}(\omega), p^{(n)}) \\ &= \left|\frac{1}{2A}W^{(n)}(\omega)\mathrm{E}\left[G(\omega)\hat{F}^*(\omega)\right] - H_{\sum}(\omega)\right|^2 \\ &+ \mathrm{E}\left[\left|G(\omega)W^{(n)}(\omega) - \hat{F}(\omega)H_{\sum}(\omega)\right|\right] \end{aligned}. \quad (12)$$

Among them, (n) is the number of iterations; p is the error probability of the blind equalizer. In the whole process of iterative calculation, error probability changes slowly, namely $p^{(n)} \approx p^{(n-1)}$. In algorithm operation process, we must ensure that the decision device output $\hat{F}(\omega)$ and the source signal $F(\omega)$ propose the same statistical properties, and it can be expressed as

$$\mathrm{E}\left[\left|F(\omega)\right|^2\right] = \mathrm{E}\left[\left|\hat{F}(\omega)\right|^2\right], \quad (13)$$

$$\mathrm{E}\left[\hat{F}(\omega)G^*(\omega)\right] = \left(1 - p^{(n)}\right)\mathrm{E}\left[F(\omega)G^*(\omega)\right]. \quad (14)$$

We set

$$\mathrm{R_a} = \mathrm{E}\left[\left|G(\omega)\right|^2\right], \quad (15)$$

$$R_b = \frac{1}{4A}\mathrm{E}\left[F(\omega)G^*(\omega)\right]\mathrm{E}\left[F^*(\omega)G(\omega)\right], \quad (16)$$

$$z = \left(1 + \frac{1}{2A}\right)H_{\sum}(\omega)\mathrm{E}\left[F(\omega)G^*(\omega)\right]. \quad (17)$$

By equation (15) and (16), $\mathrm{R_a}$, R_b are real number, and $\mathrm{R_a} > 0$, $R_b > 0$; z is a complex value number. In view of Equation (13) - (17), we have

$$\begin{aligned} &J(W^{(n)}(\omega), p^{(n)}) \\ &= \left|W^{(n)}(\omega)\right|^2\left[R_a + \left(1 - p^{(n)}\right)^2 R_b\right] \\ &- W^{*(n)}(\omega)\left(1 - p^{(n)}\right)z \\ &- W^{(n)}(\omega)\left(1 - p^{(n)}\right)z^* + \left|H_{\sum}(\omega)\right|^2\left(1 + 2A\right) \end{aligned}. \quad (18)$$

Minimum frequency error probability of medical CT image blind equalization algorithm is that the equalization problem is transformed into solving the optimal solution of equation (18).

The derivative of the cost function equation (18) with respect to equalizer frequency characteristic $W(\omega)$ can be evaluated. Under the condition of minimum error probability, we can get the best value of equalizer

$$W_{opt}(\omega) = \frac{(1-p_{\min})z}{R_a + (1-p_{\min})^2 R_b} \, . \tag{19}$$

We adjust the parameters of the equalizer by the steepest descent method, and the iterative update formula of frequency response function can be obtained as

$$\begin{aligned} & W^{(n+1)}(\omega) \\ & = W^{(n)}(\omega) - \frac{1}{2}\mu\nabla J(W^{(n)}(\omega), p^{(n)}) \\ & = W^{(n)}(\omega) - \mu\left[R_a + (1-p_{\min})^2 R_b\right] \\ & \times \left[W^{(n)}(\omega) - W_{opt}(\omega)\right] \\ & + \mu\left[(1-p_{\min})^2 - (1-p^{(n)})^2\right] R_b W^{(n)}(\omega) \\ & + \left[p_{\min} - p^{(n)}\right] z \end{aligned} \quad , \tag{20}$$

where, μ is step-size.

Using equation (20), we can obtain the equalization effect of frequency adaptive blind equalization algorithm.

3 Simulation results

We illustrate the performance of the proposed method on handling noisy degraded medical CT images in the experiment. The sectional CT image of chronic inflammation of nasopharyngeal is utilized to illustrate the effectiveness of the proposed algorithm. The original images of size 256 × 256 shown in Figure 5(a) were degraded by a 25×25 Gaussian blur with a standard deviation of 0.002, followed by an additive white Gaussian noise with 0 mean and deviation of 0.05 to form the noisy blurred images shown in Figure 1(b). In the simulation, $\mu = 5\times10^{-6}$, $p = 0.05$. We compare the results of the proposed algorithm with iterative blind deconvolution algorithm, dispersion minimization algorithm, and reduction dimension constant module algorithm. Figure1(c) is the restoration effect of the maximum likelihood method. The number of iterations of IBD is 100 and the restored image is shown in Figure 1(d).The restored images of dispersion minimization algorithm [7] is depicted in Figure 1(e). Figure 1(f) shows the result of the proposed algorithm.

As shown in Figure 1, compare with divergence minimization blind restoration algorithm, maximum likelihood algorithm and IBD algorithm, the minimum frequency domain error probability medical CT image blind equalization algorithm obtain better recovery results.

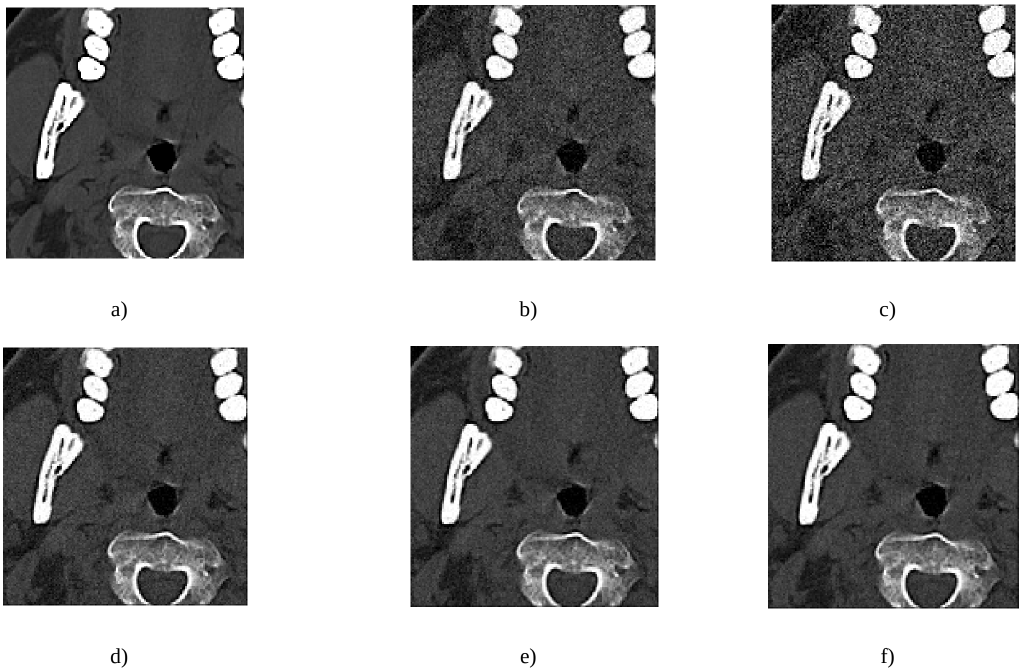


FIGURE 1 Experimental images and restoration results a) CT image (chronic recovery inflammation of nasopharyngeal), b) degraded images, c) maximum likelihood image d) IBD method image, e) divergence minimization image, f) the proposed method

Table 1 shows the PSNR [8], MSE [9], and ISNR [10] values achieved by the four algorithms mentioned above. The minimum frequency domain error probability image blind equalization algorithm improve the peak signal to noise ratio and the signal to noise ratio comparing with divergence minimize blind restoration algorithm. The new algorithm completes image blind equalization in the frequency domain. This method is essentially that a two-dimensional image signal is transform into a form of complex signal sequence. Image restoration is equivalent to minimizing a complex cost function. Compared with divergence minimization restoration algorithm, the new method is no longer competitive in the aspect of calculation performance, but it improves the signal to noise ratio PSNR improvement and reduce the minimum mean square error. IBD blind restoration algorithm and maximum likelihood methods are sensitive to noise. In particular, maximum likelihood algorithm can reduce calculate complex, but is all susceptible to noise. Computer simulation show that maximum likelihood method may yield a worse restored image than the degraded image, and even yields a negative ISNR value in low SNR.

TABLE 1 All kinds of algorithm performance

	The proposed method	Dispersion minimization algorithm	IBD algorithm	Maximum likelihood
PSNR	23.9758	23.7871	23.4624	18.7743
MSE	38.2247	40.6308	43.8088	69.9812
ISNR	1.1622	0.9736	0.6488	----

4 Conclusion

In this paper by frequency domain transformation, minimum medical CT images in frequency domain error probability of blind equalization algorithm, the algorithm of iterative formula is deduced, analysed the iteration step length selection principles and convergence performance, the computer simulation, the simulation shows that the new algorithm improves the peak signal-to-noise ratio and improve the signal-to-noise ratio, reducing the minimum mean square error. Line drawings should be drawn in black ink on drawing or tracing paper or should be glossy prints of the same, if they have not been prepared on your computer facility. All illustrations should be clearly displayed by leaving at least a single line of spacing above and below them.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No.61340034), China Postdoctoral Science Foundation(No.2013M530873), Tianjin Research Program of Application Foundation and Advanced Technology (No.13JCYBJC15600) and Tianjin City High School Science and Technology Fund Planning Project (No.20110709).

References

[1] Glick S J, Weishi Xia 1997 Iterative Restoration of SPECT Projection Images *IEEE Trans. Nuclear Science* **44**(2) 204-11

[2] Robini M C, Magnin I E 2003 Stochastic Nonlinear Image Restoration using the Wavelet Transform *IEEE Trans. Image Processing* **12**(8) 890-905

[3] Lun D P K, Chan T C L, Tai-Chiu Hsung, Feng D D, Yuk-Hee Chan 2004 Efficient Blind Image Restoration using Discrete Periodic Radon Transform *IEEE Trans. Image Processing* **13**(2) 88-200

[4] Junmei Zhong, Huifang Sun 2008 Wavelet-Based Multiscale Anisotropic Diffusion With Adaptive Statistical Analysis for Image Restoration *IEEE Transactions on Circuits and Systems-I* **55**(9) 2716-25

[5] Bin Yang, Shutao Li 2010 Multifocus Image Fusion and Restoration with Sparse Representation *IEEE Transactions on Instrumentation and Measurement* **59**(4) 884-92

[6] Nikolova M, Ng M K, Chi-Pan Tam 2010 Fast Nonconvex Nonsmooth Minimization Methods for Image Restoration and Reconstruction *IEEE Trans .Image Processing* **19**(12) 3073-88

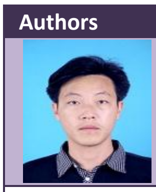
[7] Vural C, Sethares W A 2006 Blind Image Deconvolution via Dispersion Minimization *Digital Signal Processing* **26** 137-48

[8] Dalong Li, Mersereau R M, Simske S 2010 Blind Image Deconvolution Through Support Vector Regression *IEEE Trans. Neural Networks* **18**(3) 931-5

[9] Wu Y D, Sun Y, Zhang H Y, Sun S X 2007 Variational PDE based image restoration using neural network *IET Image Process* **1**(1) 85–93

[10] Almeida M S C, Almeida L B 2010 Blind and Semi-Blind Deblurring of Natural Images *IEEE Trans .Image Processing* **19**(1) 36-52

Authors



Yunshan Sun, born in 1980, in Shanxi, China

Current position: He works in the School of Information Engineering, Tianjin University of Commerce, Tianjin, China
University studies: B.Sc. degree in informatics and telecommunications from Taiyuan University of Science, Taiyuan, in 2003, and the M.Sc. degree from Taiyuan University of Technology, Taiyuan, in 2006, and the Ph.D. degree from Tianjin University, Tianjin, in 2012.
Scientific interest: blind equalization, image restoration, and medical image processing

Liyi Zhang, born in 1963, in Shanxi, China

Current position: works in the School of Information Engineering, Tianjin University of Commerce, Tianjin, China
University studies: Diploma of Electrical Engineering from the Taiyuan University of Technology, Taiyuan, in 1984, and the Ph.D. degrees from the School of optoelectronics, Beijing Institute of Technology, in 2003
Scientific interest: blind signal processing and intelligent signal processing and communications applications

Haiyan Zhang, born in 1984, in Liaocheng, China

Current position: study for the Ph.D. degree in School of Electronic Information Engineering, Tianjin University, Tianjin, China
University studies: M.Eng. degree in communication and information system from Taiyuan University of Technology, Taiyuan, China, in 2011
Scientific interest: CT image reconstruction

Authors' index

Cumulative Index

Mathematical and Computer Modelling

Xiang Zhang, Huaixiang Zhang, Yifan Wu A SoPC design of a real-time high-definition stereo matching algorithm based on SAD

Computer Modelling & New Technologies 2014 ***18****(4) 7-14*

The System-on-Programmable-Chip (SoPC) architecture to implement a stereo matching algorithm based on the sum of absolute differences (SAD) in a FPGA chip is proposed. The hardware implementation involves a 32-bit Nios II microprocessor, memory interfaces and stereo matching algorithm circuit module. The Nios II microprocessor is a configurable soft IP core in charge of managing the buffer of the stereo images and users' configuration data. The system can process any different sizes of stereo pair images through a configuration interface. The maximum horizon resolution of stereo images is 2048. When the algorithm core works under 60MHz, the 1396×1110 disparity map can be achieved at 30 fps speed.

Keywords: Stereo matching, System-on-programmable-chip, FPGA, Disparity map, SAD

Song Fei, Cui Zhe Study on improved LLE algorithm based on a sample set of well-distributed and weights matrix

Computer Modelling & New Technologies 2014 ***18****(4) 15-18*

There are large amounts of data has accumulated along with technology of computer, information and network developed. How can we using these data and mining out the valuable information are hot topics in information processing field. There are some distress and difficulties caused by the high-dimensional data on data modelling and data analysis. In this paper, a local linear embedding algorithm based on the improved uniform sample set and the weight value matrix is proposed. The test shows that the improved dimensionality reduction algorithm accuracy is significantly higher than the original LLE algorithm.

Keywords: LLE, Well-distributed, Weights Matrix

Gongfa Li, Jia Liu, Guozhang Jiang, Jianyi Kong, Liangxi Xie, Wentao Xiao, Yikun Zhang, Fuwei Cheng The nonlinear vibration analysis of the fluid conveying pipe based on finite element method

Computer Modelling & New Technologies 2014 ***18****(4) 19-24*

A Coupling between the fluid and the structure existed almost in all industrial areas the vibration of fluid solid coupling for fluid conveying pipe was called the "dynamics of typical"[1], Because of the physical model and mathematical description for the fluid conveying pipe was simple, especially it was easy to design and manufacturing, according to the characteristics of fluid conveying pipe, transformed the transverse vibration of the fluid conveying pipe to the beam element model of two nodes. Using Lagrangian interpolation function, the first order Hermite interpolation function and the Ritz method to obtain the element standard equation, and then integrated a global matrix equation. Used the mode decomposition method, obtained the vibration modal of the fluid conveying pipe with Matlab programming. The vibration modal of the fluid conveying pipe in four kinds of boundary conditions was analysed. The characteristics of pipes conveying fluid was obtained which the pipeline system parameters under different boundary constraints. To provide the theoretical support for the research of vibration attenuation of fluid conveying pipes.

Keywords: Fluid solid coupling, Nonlinear vibration, Modal analysis, Interpolation, The finite element algorithm

Jianhua Zhang Dynamic coupling analysis of rocket propelled sled using multibody-finite element method

Computer Modelling & New Technologies 2014 ***18****(4) 25-30*

Rocket propelled sled is a most important testing tool in aerospace and aviation industries flying along the rails on the ground. It is very difficult to simulate the operating conditions in the computer using numerical analysis method. In consideration of this fact, the dynamics analysis and simulation of the rocket propelled sled were done based on

Multibody System Dynamics and Finite Element Analysis theory in this paper. The most difficult work during the analysis was establishing the boundary conditions of the rocket propelled sled. This paper made this kind of attempt. Then the relevant post processing figures and data were obtained, thereby providing the designer and manufacturer with detailed and reliable data. The conclusion is the combination of finite element analysis and multisystem theory is more effective than those before and the boundary conditions are correct and acceptable. The results of it can be important references of structure designers.

Keywords: Rocket propelled sled, Finite Element Analysis, Multibody Dynamic Analysis (MDA), Multibody-Finite-Element Method, Rail Irregularity

Gaochao Xu, Yunmeng Dong, Xiaodong Fu, Yan Ding, Peng Liu, Jia Zhao A heuristic task deployment approach for load balancing

*Computer Modelling & New Technologies 2014 **18**(4) 31-36*

The load balancing strategy, which is based on the mission deployment, has become a hot topic of green cloud data centre. For the question that currently the overloaded physical hosts in the cloud data centre causes the load imbalance of the whole cloud data centre, the proposed makes an intensive study which is about the select location question of the deployment tasks on the physical host and then this proposed a new heuristic method which is called LBC. Its main idea consists of two parts: First, based on the function, which denotes the performance fitness of physical hosts, it conducts a constraint limit to all physical hosts in cloud data centre. So a task deployment strategy with global search capability is achieved. Secondly, using clustering methods can further optimize and improve the final clustering results. Thus, the whole way achieves the long-term load balancing of the cloud data centre. The results show that compared with the conventional approach, LBC significantly reduces the number of failure of the deployment tasks, improves the throughput rate of the cloud data centre, optimizes the performance of external services of the data centre, and performs well in terms of load balancing. Besides, it makes the operation of cloud data centres be more green and efficient.

Keywords: load balancing strategy, cloud data centre, task deployment, LBC, clustering

Gongfa Li, Wentao Xiao, Guozhang Jiang, Jia Liu Finite element analysis of fluid conveying pipeline of nonlinear vibration response

*Computer Modelling & New Technologies 2014 **18**(4) 37-41*

Fluid filled pipe system was widely used in the city water supply and drainage, water power, chemical machinery, aerospace, marine engineering and the nuclear industry and other fields, it was play an important role for improving the living standards of the nation and the national economic strength. Pipe conveying fluid was easy to design and manufacture, according to the characteristics of fluid conveying pipe, transformed the axial vibration mathematical model of the fluid conveying pipe, which considerate the fluid solid coupling to the beam element model for two nodes. Using Lagrangian interpolation function, the first order Hermit interpolation function and the Ritz method to obtain the element standard equation, and then integrated a global matrix equation, obtained the response of conveying fluid pipe with the Newmark method and Matlab. With the Matlab to simulate the axial motion equation of the conveying fluid pipe, study the response of the system in two aspect of fluid pressure disturbance and the fluid velocity disturbance, and the simulation results are analysed, which provides theoretical support for the work of fluid conveying pipes.

Keywords: response, MATLAB, Numerical simulation, nonlinear vibration

Yi Wang Birkhoff normal forms for the wave equations with nonlinear terms depending on the time and space variables

Computer Modelling & New Technologies 2014 ***18****(4) 42-49*

The one-dimensional (1D) quasi-periodically forced nonlinear wave equation with periodic boundary conditions is considered. It is proved that there is a real analytic and symplectic change of coordinates, which can transform the Hamiltonian to the Birkhoff normal form.

Keywords: Infinite dimensional Hamiltonian system, quasi-periodically forced nonlinear wave equation, quasi-periodic solution, periodic boundary condition, Birkhoff normal form

Sun Zeyu, Yang Tao, Shu Yunxing Wireless sensor networks optimization covering algorithms based on genetic algorithms

Computer Modelling & New Technologies 2014 ***18****(4) 50-56*

This paper starts with two methods applied widely of computational intelligence; Evolutionary computing and swarm intelligence. It makes the Genetic Algorithms (GA) that is classic in evolutionary computing and genetic algorithm that is representative in swarm intelligence as its study foundation. It presents theory and characteristic of the two methods to seek the application of intelligent optimization in engineering practice. In application, in view of the feature that wireless sensor network (WSN) must possess auto-organization, auto-adaptation and robustness, especially, energy of WSN is very limited, this paper fully utilizes the advantages of computational intelligence, marries together both the research focuses. It proposes some methods and ideas for applying computational intelligence to solve optimization problems of WSN. This paper depicts coverage problem of WSN, for the feature that this problem is the problem of multi objective optimization, under the topology control of GA, it applies GA based on sorting to solve the problem, then improves this algorithm to maintain population diversity and obtain high-quality, well distributed solutions. The algorithm it proposes realizes the aim that using the least number of sensor nodes to achieve the best coverage, which is able to save energy of the network, decrease the interference between signals and prolong the network life-time.

Keywords: Wireless Sensor Network (WSN), Coverage rate, Sensor node, Genetic Algorithms

Na Wang, Yanxia Pang An improved light-weight trust model in WSN

Computer Modelling & New Technologies 2014 ***18****(4) 57-61*

WSN is often deployed in unattended or even hostile environments. Therefore, providing security in WSN is a major requirement for acceptance and deployment of WSN. Furthermore, establishing trust in a clustered environment can provide numerous advantages. We proposed a light-weight trust model which considers data aggregation and communication failure due to wireless channels. It computes retransmission rate to get success, failed and uncertain value, and details the data in parameters to depend against attacks. With comparing our model with LDTS and Model using Trust Matrix, we conclude that our model has implemented a trade-off between detection rate and communication consumption.

Keywords: direct trust, light-weight, trust matrix, retransmission rate, indirect trust

Information and Computer Technologies

Gaochao Xu, Yushuang Dong, Bingyi Sun, Xiaodong Fu, Jia Zhao An approach of VM image customized through Linux from scratch on cloud platform

Computer Modelling New Technologies 2014 ***18****(4) 62-67*

The cloud platform provides abundant resources and services for users. More and more mobile users began to use the cloud services. They have higher real-time demands on service. The size of traditional virtual machine (VM) operating system is basically large. It will take many resources in deployment and communication process, and always affect the real-time performance of system. To reduce communication overhead and improve deployment speed of VMs, this paper proposes an approach of customized VM image with LFS. LFS can reduce the size of VM image efficiently and enable flexible customization of the VM image by incremental customization. The experimental results show us that the size of VM image generated by the proposed method is smaller than the one generated by kernel tailoring technology in system overhead. Meanwhile it is also faster in running speed.

Keywords: Cloud computing, Linux from scratch, Customized virtual machine

Fengqing Qin Blind multi-image super resolution reconstruction with Gaussian blur and Gaussian noise
*Computer Modelling & New Technologies 2014 **18**(4) 68-99*

A framework of blind multi-image super resolution reconstruction method is proposed to improve the resolution of low resolution images with Gaussian blur and noise. In the low resolution imaging model, the shift motion, Gaussian blur, down-sampling, as well as Gaussian noise are all considered. Firstly, the Gaussian noise in the low resolution image is reduced through Wiener filtering method. Secondly, the Gaussian blur of the de-noised image is estimated through error-parameter analysis method. Thirdly, the motion parameters are estimated. Finally, super resolution reconstruction is performed through iterative back projection algorithm. Experimental results show that the Gaussian blur and motion parameters are estimated with high precision, and that the Gaussian noise is restrained effectively. The visual effect and peak signal to noise ratio (PSNR) of the super resolution reconstructed image are enhanced. The importance of Gaussian blur estimation and effect of Gaussian de-noising in multi-image super resolution reconstruction are tested in an experimental way.

Keywords: blind, multi-image super resolution, Gaussian blur, Gaussian noise, iterative back projection

Xu Gaochao, Ding Yan, Ou Shumao, Hu Liang, Zhao Jia A 'Follow-Me' computing scheme based on virtual machine movement for QoS improvement in mobile cloud computing environments
*Computer Modelling & New Technologies 2014 **18**(4) 74-81*

Mobile cloud computing utilizes virtualized cloud computing technologies in the mobile Internet. To improve Quality of Service (QoS) and execution efficiency of mobile cloud applications, we propose a novel computing scheme called "Follow-Me" (FM), which is based on live wide-area virtual machine (VM) migration. In a virtualized mobile cloud environment based on the VMs of cloud side and mobile devices of user side, the purpose of the proposed FM scheme is to migrate the corresponding VM in real-time when a mobile device moves from one service area to another. FM obtains the current positions of mobile devices, estimates the next servicing areas, and finally migrates the VMs along with the mobile users' movement. The proposed FM scheme has been tested in an experimental environment by using the CloudSim platform. The experimental results demonstrate that FM evidently improves the QoS of mobile cloud computing compared with the existing approaches. FM achieves a better average service response time, a clearly smaller error rate and consumes less energy.

Keywords: Mobile Cloud Computing, Mobile Device, Virtual Machine, Area Localization, Live Wide-Area Migration

Xiao Xiaohong, Wu Yonggang A cloud-removal method based on image fusion using local indexes
*Computer Modelling & New Technologies 2014 **18**(4) 82-88*

For optical images, cloud and cloud shadow is always a problem during image processing and interpretation. Landsat ETM+ images, as a kind of optical images, are affected by cloud too. On the other hand, microwave images such as ALOS PALSAR images, which depend on microwave, is not affected by cloud, thus they are cloud-free. The aim of this study is to develop a semi-automatic method for removing cloud and cloud shadow in Landsat ETM+ images based on fusion of Landsat ETM+ image and ALOS PALSAR image. The key point of this method is to develop a cloud and cloud shadow mask based on which Landsat ETM+ and ALOS PALSAR images can be fused. To accurately define cloud and cloud shadow area, we first approximately draw the area of interest containing cloud and cloud shadow manually, and the resulted AOI image greatly reduce the number of ground objects and the confusion between objects as well. By analysing the spectral and the grey value of the AOI image, we then define LCI (local cloud index), LSI (local shadow index), and LGI (local ground index) to accurately identify cloud and cloud shadow area in Landsat ETM+ images. Finally, a combination mask of cloud and cloud shadow is developed. Based on this mask, Landsat ETM+ image and ALOS PALSAR image are merged. The fused image is cloud free, at the same time;

it keeps the spectral feature and the integrity of Landsat ETM+ image.

Keywords: Cloud-removal, AOI, Landsat ETM+, ALOS PALSAR, LCI, LSI, LGI

Song Xiaoxia An efficient method for acquiring and processing signals based on compressed sensing
Computer Modelling & New Technologies 2014 ***18****(4) 89-94*

Compressed sensing (CS) theory provides a novel sensing/sampling and processing paradigm that breaks through the limitation of Nyquist rate to some applications. However, it is usually happened to the instability and redundancy of the acquired CS measurements. In view of this, we propose an efficient method to achieve adaptive minimal measurements with fewer measurements and good reconstruction performance by adding the pre-processing block into CS data acquiring and processing paradigm. In the proposed method, we firstly obtain the measurements to perfectly reconstruct the signal, and then design the optimization method to obtain adaptive minimal measurements by eliminating the redundant measurements. Experimental results show that the proposed method can obtain fewer measurements to perfectly reconstruct the signal than that of classical CS and sequential compressed sensing frameworks.

Keywords: compressed sensing, sequential compressed sensing, signal reconstruction, homotopy method

Xiaoping Wang, Jin Yao, Wangang Wang Detection of WCDMA uplink signal with combination between sliding match and power spectrum
Computer Modelling & New Technologies 2014 ***18****(4) 95-101*

Aiming at problem of WCDMA uplink signal being difficult to be detected under low SNR, this paper proposes a type of algorithm in which sliding match combines with power spectrum to detect WCDMA signal. Firstly, this algorithm estimates desynchronizing point of signal using Frobenius norm. According to desynchronizing point, a whole cycle of information sequence is intercepted. Correlation of OVSF code sequence is utilized in which residual carrier or DC component of signal would come into being while the received OVSF code sequence completely matches with local OVSF code sequence. Then its power spectrum is calculated and sharp spectral peak would appear in the frequency position of power spectrum. Through detecting amplitude and position of spectral peak, frequencies of OVSF code sequence and residual carrier utilized in WCDMA signal could be accurately estimated. Simulation results show that this algorithm rapidly realizes the estimation on OVSF code sequence and desynchronizing point keeps good detection effect and may effectively overcome the influences of residual frequency offset on it.

Keywords: WCDMA signal, OVSF code sequence, Frobenius norm, sliding match, power spectrum

Operation research and decision making

Lei Li, Jianrun Zhang Multidisciplinary design optimization of complex products based on data fusion and agent model
Computer Modelling & New Technologies 2014 ***18****(4) 102-107*

Multidisciplinary design optimization (MDO) of complex products is discussed in this article. For the characteristics of higher order, high-dimensional, multi-input and multi-output in design of complex products, application of MDO in design and optimization of complex products is difficult. An effective MDO framework combined with the method of data fusion and agent model is proposed. Firstly, date fusion is applied to deal with the process with a large number of incomplete, vague and uncertain in complex product's evaluation and optimization; secondly, agent model is used to reduce the complexity of the MDO model; and finally, MDO is applied to complex products design and optimization according to the collaborative design and optimization method. In order to identify the feasibility of this method, the design of diesel engine motion mechanism is discussed and shown a good result. The current study provides a powerful tool for complex products designing and optimization and owns great theory and practical values.

Keywords: Complex Products, MDO, Date Fusion, Agent model

Si Chen, Gan Mi, Dingqi Shuai Research on the influence of the different logistics demand structures of the city in regional logistics planning
Computer Modelling & New Technologies 2014 ***18****(4) 108-112*

This research analysis the different parts of regional logistics demand at first. There are three parts of the logistics demand considered in this paper; they are logistics demand in the city, logistics demand between cities and the logistics demand from or outside the area. The relationships have been studied by the Grey Theory, and a numerical example has been made to show the way how to analysis the logistics demand structure in the regional logistics planning. In the regional logistics, planning the difference of logistics demand structures of the cities should be fully considered. Then the logistics planning with different regional logistics planning purposes have been programmed. Based on the numerical example, different plans and different influence scopes have been got at last.

Keywords: Logistics Demand, Grey Theory, Logistics planning, Demand Structure

Yongqin Jin, Dongsheng Liu Integrating TTF and TAM perspectives to explain mobile knowledge work adoption

Computer Modelling & New Technologies 2014 ***18(****4) 113-120*

It is an advanced research subject to information technology as well as a great influence to the development of mobile work that how mobile work service is adopted and how it provides effectiveness. This article refers to 1) Analyse principles of TAM & TTF models, clarify the precondition and strength & weakness of the model, and propose a new mobile work service model combined with TAM & TTF; 2) Practical study to the new model. The conclusion for this model is as follow: a) two basic characteristics of mobile work and support from up-level managers in a firm are the preconditions whether service will be adopted. b) task-technology matching is the significant factor on service acceptation. c) It could improve employees' efficiency within the practical use of task-technology matching mobile work service.

Keywords: mobile work, TTF, TAM, Embedded model

Jie Liu, Qiusi Dai, Huifeng Yan Application of magnified BP algorithm in forecasting the physical changes of ancient wooden buildings

Computer Modelling & New Technologies 2014 ***18****(4) 121-125*

CA using neural network model of ancient buildings to predict changes in the physical properties of Applied X-ray detector collection of ancient buildings grey wood elements, so that the ancient wooden building components of each pixel grayscale and Neural Network CA model correspond to each cell, using the CA model "grey" concept learning through the improved BP Algorithm to calculate the grey value of each cell changes, so as to arrive ancient architectural wood elements over time the case of damage by example through the projections obtained wood over time damage to the picture.

Keywords: ancient building, BP neural network, cellular Automata (CA)

Cunjie Guo, Wei Liang, Laibin Zhang The Semi-quantitative evaluation method and application of the risks of geological disaster of the Shaan–Jing pipeline

Computer Modelling & New Technologies 2014 ***18****(4) 126-130*

Based on the index scoring method, the semi-quantitative method for assessment of pipeline geological disaster risks calculates the relative risk of a disaster by investing and assessing the objective existence state of index in accordance with pre-determined scores and weights, and meets the requirements of risk prioritizing and ranking at the geological disaster investigation stage so as to guide the development of risk control planning. A geological disaster risk semi-quantitative assessment system and risk grading standards both of which are applicable to oil and gas pipelines have been established. What has been developed also includes the pipeline geological disaster risk management system software, which integrates the risk semi-quantitative assessment technique based on the index-scoring-method, and other techniques such as information management and risk management, and thus provides a platform of information, technology and management for the management of pipeline geological disaster risks. This method has been used for a unified risk assessment of more than 3300 disaster points along the oil and gas pipeline, and satisfactory evaluation results are obtained, thus providing an important basis for the development of planning of pipeline geological disaster risk remediation.

Keywords: oil and gas pipeline, geological disasters, index scoring method, semi-quantitative assessment, risk grading

Dasheng Wu Estimation of forest volume based on LM-BP neural network model

*Computer Modelling & New Technologies 2014 **18**(4) 131-137*

Since cost factors are of primary importance, continuously searching for more efficient and reliable estimation models that could integrate or, in some cases, substitute the traditional and expensive measuring techniques for forest investigation is necessary. The evaluation indexes set, which included 10 factors: elevation, slope, aspect, surface curvature, solar radiation index, topographic humidity index, tree ages, the depth of soil layer, the depth of soil A layer, and coarseness, was established. Then, using the integration data of the administrative map, Digital Elevation Model (DEM), and forest resource planning investigation data of the key forestry city of Longquan, Zhejiang Province, PRC, the membership of each factor was empirically fitted by polynomials, and the forest volume was estimated via an improved back propagation (BP) neural network(NN) model with Levenberg-Marquardt(LM) optimization algorithm(LM-BP). The results show that the individual average relative errors (IARE) were from 23.29% to 47.87% with an average value of 33.06%; The groups relative errors (GRE) were from 0.38% to 9.31% with an average value of 3.65%, this meant that groups estimation precision was more than 90% which is the highest standard of overall sampling accuracy about volume of forest resource inventory in china.

Keywords: LM-BP; Forest Volume; Estimation

Kun Zhang, Yanxi Chen, Minfei Qiang Required screw length measurement in distal tibia based on three-dimensional simulated screw insertion

*Computer Modelling & New Technologies 2014 **18(**4) 138-144*

The objective of the study was to provide morphological data of the distal tibia to offer guidance on the required screw length. Computed tomography scans of the ankle in 225 patients were reviewed. Then parameters in the three-dimensional reconstruction images were measured by three independent, qualified observers on 2 separate occasions. The anteroposterior length increases from medial to lateral margin at the level of the base of the tibiofibular syndesmosis. On both proximal and distal planes of tibiofibular syndesmosis, the medial-lateral width increases from posterior to anterior margin. Significant differences were observed in all parameters between male and female and in the minimum width at the level of the roof of the syndesmosis between left and right limbs ($P<0.05$). All of the parameters exhibited moderate to excellent intra-class correlation coefficient. The anteroposterior screws would probably penetrate the far cortex and injure the structures surrounding the distal tibia if longer than 35.35 mm and 32.53 mm in male and female. The screws should not longer than the maximum diagonals which are 51.29 mm and 46.58 mm on distal plane and 43.64 mm and 38.24 mm on proximal plane in male and female respectively, or inadvertent distal tibiofibular syndesmosis penetration may occur.

Keywords: Tibia, Tibiofibular syndesmosis, Tomography, X-ray computed, Imaging, three-dimensional

Weiming Chen, Xiaoyang Ni, Hailin Guo Disruption management for resource-constrained project scheduling based on differential evolution algorithm

*Computer Modelling & New Technologies 2014 **18**(4) 145-153*

In this paper, we study the problem of how to react when an ongoing project is disrupted. The focus is on the resource-constrained project scheduling problem with finish–start precedence constraints and the recovery strategies based on disruption management for the different types of disruptions are proposed. The goal is to get back on track as soon as possible at minimum cost, where cost is now a function of the deviation from the original schedule. The problem is solved with a differential evolution (DE) algorithm that can be solved more perfectly on the objective function. The new model is significantly different from the original one due to the fact that a different set of feasibility conditions and performance requirements must be considered during the recovery process. Project scheduling problem library (PSPLIB) has been taken into account so as to test the effect of novel hybrid method. Simulation results and comparisons determine the effects of different factors related to the recovery process and show that the differential evolution algorithm is competitive and stable in performance.

Keywords: disruption management, scheduling, resource-constrained, differential evolution

Xing Yu The optimal dynamic robust portfolio model
Computer Modelling & New Technologies 2014 ***18****(4) 154-157*

This paper is concerned with the optimal dynamic multi-stage portfolio of mean- dynamic var based on high frequency exchange data with the constraint of transaction costs transaction volume. The proposed solution approach is based on robust optimization, which allows us to obtain a worst best but exact and explicit problem formulation in terms of a convex quadratic program. In contrast to the mainstream stochastic programming approach to multi-period optimization, which has the drawback of being computationally intractable, the proposed setup leads to optimization problems that can be solved efficiently.

Keywords: dynamic portfolio, mean-var, robust, high frequency exchange

Baohui Jin Travel route choice model based on regret theory
Computer Modelling & New Technologies 2014 ***18****(4) 158-163*

Travel route choice behaviour research is a hot issue in the field of urban traffic planning, and it mainly researches the traveller's route choice decisions under uncertainty conditions, which theory includes such as expected utility theory, prospect theory, and regret theory. Based on the analysis of expected utility theory and prospect theory's applicable condition and the insufficiency, this paper establishes a travel route choice model according to regret theory. Study shows that people always try to avoid occur that other options is better than that selected option, and the properties of selected option cannot be replaced each other, which fits regret minimization of regret theory. The travel route choice model based on regret theory is simpler than others, and it is suitable for describing traveller's route choice behaviour under uncertainty conditions.

Keywords: urban traffic, travel route choice, regret theory, Bayesian updating

Heping Zhong Optimal contracts of production personnel's innovation based on slack resources
Computer Modelling & New Technologies 2014 ***18****(4) 164-171*

Based on the analysis frames of the multi-task principal-agent model, this paper establishes a principal-agent model of production personnel's innovation based on slack resources and obtains the optimal incentive contracts for production personnel while they are engaged in "production task" and "slack innovation" through the analysis of the model. In order to improve the performance of production personnel's "slack innovation", on one hand, the firm can reward their "slack innovation" according to the optimal incentive contracts; on the other hand, the firm can optimize the incentive contracts for their "production task" according to the interdependence of the cost functions of "production task" and "slack innovation" to promote indirectly the performance of "slack innovation". The originality of this paper is not only examining the multi-task problems of the compensation incentives for production personnel's "slack innovation" but also considering the impacts of the firm's active actions to support the production personnel's "slack innovation" on incentive contracts.

Keywords: contract, incentive, optimization, multi-task agent model, innovation, slack resources, production personnel

Zhou Huijuan, Zhao Huan, Liu Baoxun, Fan Qinglan A facilities state-based evaluation method on level of service in subway station
Computer Modelling & New Technologies 2014 ***18****(4) 172-177*

Subway station is the key node in the urban rail transit system. Its level of service affects directly the subway's operation efficiency and traveller's choice of track traffic way to travel. Considering the facility characteristics in subway station and pedestrians perspective, on the basis of a large number of survey data, this paper identifies the facilities which impact the level of service in subway station mainly, takes safety, comfort and smoothness as evaluation index, and evaluates respectively from the entrance, channel and platform area. The judgment matrix of facilities condition influence in each region on pedestrians is constructed and the evaluation model of level of service in subway station has been built based on the facility state. Finally taking PingGuoYuan subway station as instance for analysis, the result verifies that the evaluation method is effective.

Keywords: urban transportation, facilities state, level of service, judgment matrix, pedestrian experience

Hailing Li, Kejian Liu Resource management modelling and simulating of construction project based on Petri net

Computer Modelling & New Technologies 2014 ***18****(4) 178-183*

This paper establishes a model to exactly express the resource configuration, task duration and information transmission during the project execution phase. Based on the resources' properties in the projection execution phase and the hierarchical timed coloured Petri net (HTCPN), this hierarchical model exactly express the information required for project resource management, such as the task dependencies, resource demands and the task durations by defining a non-empty colour set as coloured tokens to represent the classes and combinations of the resources. This model is then simulated and analysed on the model structure, resource conflicts and run time using CPN Tools to verify the correctness and effectiveness of the HTCPN modelling of the project resources in the project execution phase.

Keywords: Construction Project, Petri Net, Resource, Modelling, Simulation

Lifei Yao, Ruimin Ma, Maozhu Jin, Peiyu Ren Study on distribution centre's location selection of internal supply chain for large group manufacturing companies

Computer Modelling & New Technologies 2014 ***18****(4) 184-190*

The purpose of this paper is to study what distribution centre's location selection can bring to the internal supply main management for large group manufacturing companies. This paper chooses the analytic hierarchy process to select an optional location for internal distribution centre', and evaluate it through the simulation method. Internal distribution centre construction can effectively shorten the delivery time, reduce the logistics intensity, and improve utilization rate of transport equipment. Therefore, the distribution centre's location selection is necessary and reasonable. This paper simplified some information when running the simulation and it is not all the same as the actual situation. This paper provides a good internal supply main management method for large group manufacturing companies. This paper put forward the importance of internal supply chain for a large group manufacturing company and studied the internal distribution centre's location selection.

Keywords: manufacturing, internal supply chain, distribution centre, simulation

Ouyang Fang, Chih-hung Hsu Exploring Dynamic Performance improvement in Service SCM: the Lean Six Sigma's perspective

Computer Modelling & New Technologies 2014 ***18****(4) 191-196*

This paper defines the performance evaluation system of Service SCM. As service is intangible and heterogeneous, the paper is to develop a model that illustrates under which conditions Lean Six Sigma is deemed most appropriate according to the type of service delivered. It investigate Lean Six Sigma practice in service supply chain and show how the Lean Six Sigma improve the performance of Service SCM from the statistics perspective. Furthermore, it stresses the CTQ (critical to quality) to the customer and clarifying their demands in terms of value-added requirements.

Keywords: Service Supply Chain Management, Intangible, Lean Six Sigma

Shimei Wu, Yulong Pei, Guozhu Cheng The study of urban traffic modal splitting method based on MD model under the low-carbon mode

Computer Modelling & New Technologies 2014 ***18****(4) 197-202*

Aimed at the problem of overly-simplify in the factors of travel cost in the traffic modal splitting method, built an Urban Traffic Modal Splitting Method Based on MD Model Under the Low-carbon Mode, to predict transportation share rate; Put forward four considerations such as the travel time, cost, safety and low carbon to describe the travel cost on the basis of the application of MD model; Gave the forecasting process of the prediction model and key variables algorithm, applied the model by the examples of DONGGUAN city. The results show that the urban structure of the transportation changed in DONGGUAN with rapid construction, development in traffic and implementation of transport policy, on the one hand, the travel occupies proportion of public transport (including

conventional bus and rail transit) will increase significantly in the future, expected to reach more than 25% by 2020; on the other hand, motorcycle travel will gradually fade away.

Keywords: Transportation Planning, Low-carbon Transportation, MD model, traffic modal split

Guangbin Wang, Honglei Liu, Lei Zhang Research on dynamic evolution of innovative virtual prototyping technology diffusion based on cellular automata
*Computer Modelling & New Technologies 2014 **18**(4) 203-209*

The construction industry plays a very important role in the national economy; it is widely criticized because of its slow technical progress and long-term inefficiency all over the world. Building information modelling (BIM) is a transformative virtual prototyping technology for construction industry. VP (Virtual Prototyping Technology) based on BIM as the core technology has been widely regarded as a tool to solve this problem, but was questioned by both academia and industry due to its delayed diffusion. To solve this problem, this paper is based on the characteristics and the evolution rules of cellular automata, built on the CA model of the BIM proliferation process in construction projects, simulating this process, then analysing the impact of important factors such as diffusion willing, decision-making preferences, national and industry support and other factors to the BIM technology diffusion, studying the changes in the proportion of BIM recipients and the importance of the distribution position of the initial to the BIM proliferation process. Finally, it analyses the randomness of BIM technology diffusion.

Keywords: Building Information Model Diffusion, CA Evolution Model, Diffusion Willingness, Decision-Making Preferences, National and Industry Support

Chen Kai, Xie Yi Research on well-formed business process modelling mechanism
*Computer Modelling & New Technologies 2014 **18**(4) 210-214*

It is very important to ensure that the logic structure of business process model is correct before the model is implemented. Because traditional graphical process modelling methods lack efficiency mechanisms or rules to ensure correctness of the logical structure during business process modelling, they need additional methods to verify its correctness of the logic structure after the business process model is established. Therefore, the well-formed business process modelling mechanism is researched. The business process logic structure model is built firstly. Then the semantic and syntactic rules are presents for the correctness of business process logic structure model, and the algorithm is proposed to detect whether the model meets the rules. The modelling mechanism has been applied in our business process scheduling optimization system with integration of modelling and simulation, which shows its feasibility and effectiveness.

Keywords: business process modelling, modelling mechanism, well-formed model, model verification

Luya Wang, Liang Xiao Research on wisdom urban public security management system integrated into the situation of urban safety
*Computer Modelling & New Technologies 2014 **18**(4) 215-221*

With the expanding of the sizes of the cities, the urban population and property space distribution becomes more concentrated, urban public safety incidents into the increasingly frequent stage. How to intelligent and efficient manage the urban public safety is imminently. On the basis of defining the urban security situation management model systematic, this article will establish the urban safety stratified hierarchical data acquisition of internet of things which is based on urban monomer-group region, study the tracking-summarized-warning-optimization handling mechanism which support the city security complex event, construct the wisdom urban public security management system which is integrated into urban security situation and provide an effective means to realize the wisdom management of the city public security.

Keywords: public security, safety situation, wisdom city, internet of things

NATURE PHENOMENA AND INNOVATIVE TECHNOLOGIES

Wenxia Qiu, Huixi Xu, Zhengwei He Study on the difference of urban heat island defined by brightness temperature and land surface temperature retrieved by RS technology
*Computer Modelling & New Technologies 2014 **18**(4) 222-225*

At present, the Remote Sensing is the most advantage method of studying on the Urban Heat Island (UHI) from the space. In general, the method uses remote sensing images to inverse the brightness temperature or land surface temperature to define the UHI. But have any differences of UHI defined by the two kinds of temperature? And what are the differences? This problem is rarely being studied now. Based on this, the brightness temperature (BT) and the land surface temperature (LST) of the Chengdu City were retrieved using Landsat ETM+ image obtained on July 30, 2006. And then, the differences of UHI defined by the BT and the LST were studied from three aspects, which were temperature value, temperature classification and heat island intensity respectively. Research result are the following: (1) There were some differences between BT and LST, and the variation level of LST was higher than BT. (2)There was a slight difference only on the area covered by the low temperature and the secondary low temperature, and the area covered by the others was equal. Therefore, there was no difference on the area of UHI defined by BT and LST. (3) The UHI intensity defined by LST was slightly higher than that was defined by BT, and the intensity value was determined by the method used.

Keywords: Urban Heat Island (UHI), Brightness Temperature (BT), Land Surface Temperature (LST), Remote Sensing Technology (RS)

Xuechao Liao, Zhenxing Liu Multi-level dosing and preact self-adaption correcting automatic batch control model
*Computer Modelling & New Technologies 2014 **18**(4) 226-231*

The process flow and system structure of automatic batch weighing system are presented. In order to increase production speed and dosing accuracy, the multi-level dosing control model (high/low speed dosing + inching dosing) is designed. Besides, the inching dosing mode is adopted to accurately compensate the weight deviation. In order to solve the problem that the fall of materials in-air cannot be easily controlled and out of tolerance. The multi-level dosing control model and preact will correct after each dosing dynamically with iteration method, moreover, the target value is predicted with second-order estimator, so as to increase the dosing speed with high weighing accuracy. The successful application proves that the control model can realize the rapid and accurate control of batch weighing process and has quite favourable control and reliability.

Keywords: Automatic batch, Multi-levels dosing, Fall of dosing, Self-adaption correcting, preact

Wu Zhang, Wei Guo, Chuanwei Zhang, Yizhi Yang, Yu Zhang Experimental research on transmission efficiency of metal belt continuously variable transmission
*Computer Modelling & New Technologies 2014 **18**(4) 232-237*

Transmission efficiency is one of the main limiting factors on metal belt CVT large-scale assembly car. Metal belt CVT transmission efficiency has been invested in this paper, and, test-bed has been established by L13A3 engine, MB-CVT, brake, input sensor, output sensor, coupling and half shaft. Efficiency test results show that, with the decrease of transmission ratio, CVT efficiency first increases and then decreases. The range of efficiency is nearly 45%-89% in increases part (i>1), the range of efficiency is nearly 85%-89% in decrease part (i<1), the efficiency reaches the highest when transmission ratio is 1. The conclusions are in consistent with others conclusion, whereby demonstrating that the established transmission efficiency test-bed is rational and that the experiment results are reliable.

Keywords: metal belt, CVT, pulley, strain

Zhang Feng, Xue-Hui Feng A method of reliability modelling based on characteristic model for performance digital mock-up of hypersonic vehicle
*Computer Modelling & New Technologies 2014 **18**(4) 238-242*

In order to grasp the complexity of the hypersonic vehicle dynamic characteristics, create its reliability control model, for the mathematical model of hypersonic vehicles is highly nonlinear and strong coupling, introduced the object-oriented modelling method, design a neural network, Petri control algorithm based on characteristics model, the mathematical model of the nonlinear is transformed into the equivalent linear model with control design requirements. And through the appropriate transform, design of the hypersonic vehicle dynamic inversion control system, building performance prototype reliability model based on Petri net, can stabilize the system, get decoupled affect purposes. Simulation results show that the performance digital mock-up reliability model is high accuracy, robustness, anti-jamming capability, has a good dynamic and steady-state performance.

Keywords: hypersonic vehicle, performance digital mock-up, characteristic model, reliability, flight control

Jiangtao Liu, Jinggang Yi Research of key technology on self-propelled farmland levelling machine and hydraulic servo system simulation

*Computer Modelling & New Technologies 2014 **18**(4) 243-248*

According to the present situation of the farmland levelling, the equipment cost is high, maintenance is complex and its cost is high. The paper carries a research on the key technology of self-propelled farmland levelling machine. The key technology includes the levelling knife, the levelling part, the sundry separating device and the measurement and control system of the laser and inclination sensor. At the same time, the paper establishes the hydraulic servo system mathematical modal and utilizes MATLAB to analyse, revise and simulate for the system mathematical modal.

Keywords: farmland levelling machine, levelling knife, levelling part, inclination sensor, simulation

Manli Dou, Chun Shi, Gang Wu, Xiaoguang Liu Comfort and energy-saving control of electric vehicle based on nonlinear model predictive algorithm

*Computer Modelling & New Technologies 2014 **18**(4) 249-254*

This paper develops a control-oriented drivability model for an electric vehicle and a nonlinear model predictive optimization algorithm for an electric vehicle. A cost function is developed that considers the tracking error of setting value and the variation of control volume. Longitudinal ride comfort and energy-saving is also considered. Simulations show that the developed control system provides significant benefits in terms of fuel economy, vehicle safety and tracking capability while at the same time also satisfying driver desired car following characteristics.

Keywords: Nonlinear Model Predictive, Comfort, Energy-saving, Electric Vehicle

Ying Zhang, Zhaohui Yuan Force-fight problem in control of aileron’s plane

*Computer Modelling & New Technologies 2014 **18**(4) 255-261*

In order to reduce or eliminate the force-fight phenomenon of single feedback loop of redundant-channels, modelling the whole control system on the basis of analysing the structure of aileron, the correctness of the model is verified by experiments. Simulation results show that set the dead band of the valve which control the feedback loop smaller is conducive to the decrease of system’s fighting-force; for every reduce in the difference of the two valves’ overlap of 0.01mm, the fighting-force decreases one time; when the driving speed is more than 50mm/s, system abstains smaller fighting-force. Therefore, the optimization of structure parameters can reduce fighting-force effectively. When the parameters of valves and driving speed is restricted, another method of using a bypass orifice to connect the two cavities of the cylinder is proposed to solve the problem, simulation results shows that fighting-force reduce 2000N for every increase in the orifice’s diameter of 0.1mm when using the fixed orifice, and using the variable orifice can abstain a small fighting-force and meanwhile reduce the wastage of hydraulic oil.

Keywords: single feedback loop, force-fight phenomenon, difference of the two valves’ overlap, driving speed of motor, bypass orifice

Xiaobo Li Gene selection for cancer classification using the combination of SVM-RFE and GA

*Computer Modelling & New Technologies 2014 **18**(4) 262-267*

Gene selection is a key research issue in molecular cancer classification and identification of cancer biomarkers using

microarray data. Support vector machine recursive feature elimination (SVM-RFE) is a well known algorithm for this purpose. In this study, a novel gene selection algorithm is proposed to enhance the SVM-RFE method. The proposed approach is designed to use the combination of SVM-RFE and genetic algorithm (GA). The performance of the proposed model is validated on a binary and a multi-category microarray gene expression datasets. The results show that the proposed gene selection method is able to elevate the performance of SVM-RFE, which extracts much less number of informative genes and achieves highest classification accuracy.

Keywords: cancer classification, gene selection, support vector machine recursive feature elimination (SVM-RFE), genetic algorithm (GA), microarray data

Shigang Mu Dynamic analysis of ball-screw with rotating nut driven
*Computer Modelling & New Technologies 2014 **18**(4) 268-272*

There is a certain degree difference between the static and operation condition for the high-speed Ball-screw with Rotating Nut. Therefore, this paper establishes a dynamic model of a preload-adjustable ball-screw with rotating nut by means of lumped-parameter and analyses the effects of changeable table position and work piece mass on the first three axial modes of the free vibration. A high-speed feeding system is modelled and its nature characteristics when the feeding system is in static, low and high rotate state. The results show that, at low speed state, the dynamics of the feeding system is the same as stationary state, and in high-speed conditions, the dynamics is quite different with the static state. The natural frequencies are notably changed with the position change of the table movement. The research lays an important theoretical foundation for developing this novel feed drive system.

Keywords: Ball screw, Dynamic analysis, Modal analysis, Frequency response

Ming Lv, Haiqiang Liu Thermodynamic analysis of hydrogen production via zinc hydrolysis process
*Computer Modelling & New Technologies 2014 **18**(4) 273-277*

The thermodynamic studies were carried out for the hydrogen production via zinc hydrolysis. It is shows that it is reasonable to keep the temperature of zinc hydrolysis under 900 ºC. The system pressure has no notable thermodynamic influences on the hydrolysis reaction. The initial H_2O/Zn molar ratio should be controlled in a reasonable range. The concentration of steam in carrying gas in experiments should better be kept above 50%.

Keywords: hydrogen, hydrolysis, thermodynamics, zinc

Qiang Song, Ai-min Wang Study on prediction of sintering drum strength under small sample lacking information
*Computer Modelling & New Technologies 2014 **18**(4) 278-283*

The paper provides a grey model and support vector machine algorithm and method for prediction of sinter drum strength based on the characteristics of large time delay, strong coupling, nonlinear, sintering process, put forward a kind of Combination forecasting model of drum strength based on grey model and support vector machine, the drum strength of sinter ore Laboratory values as output variables, the variables associated with the drum strength of sinter as input variables, using support vector machine powerful machine learning method and strong nonlinear fitting ability, so as to establish a stable, high precision of drum strength, the drum strength stronger generalization ability of the forecasting model, the method of the method has the high prediction accuracy, fast and convenient, and has great popularization and application value, and lay a good foundation for the green sintering technology of sintering.

Keywords: GM(1,1), LS-SVM, drum strength, Prediction

Jinglong Li, Bin Sui Discussion on determination method of characteristic stress of Jinping marble under confining pressure condition
*Computer Modelling & New Technologies 2014 **18**(4) 284-288*

The characteristic stress is coincident well with the internal crack propagation in brittle rock. The characteristic stress are separately called closure stress, cracking stress, damaging stress and peak stress according to the internal crack state in loading. The propagation and damage extent in brittle rock can be reflected. Limited by loading testing equipment, the characteristic stress in confining pressure condition cannot be determined in China. In order to confirm the stress, the strain curves under different confining pressure condition are used to analysis the problem. The results show that the closure stress, cracking stress and damaging stress can be accurately confirmed by this method. The characteristic stress relates to the confining pressure, and the relationship is approximately linear.

Keywords: brittle rock, characteristic stress, marble, confining pressure

Jianning Han, Tingdun Wen, Peng Yang, Lu Zhang High resolution photoacoustic system based on acoustic lens and photoacoustic sensors array

*Computer Modelling & New Technologies 2014 **18**(4) 289-295*

Photoacoustic tomography is a nondestructive bio-photonic imaging method based on the differences of optical absorption within biological organization. An approach using the lens with negative refractive index and photoacoustic sensors array to make the evanescent wave involved in the imaging process was presented in this paper. A set of comparative experiments was demonstrated on the imaging effect between the ordinary lens and the lens designed in this work. The experiment showed that the imaging effect of photoacoustic tomography by the designed lens had greatly outperformed the ordinary lens. In order to illustrate the good results, according to the characteristics of ultrasonic waves produced in photoacoustic effect, the propagation properties of the acoustic waves in lens with different refractive index was discussed. On the basis of analysing evanescent decay of ordinary acoustic lens which results in the loss of high-frequency information with image details in current photo-acoustic tomography system, the diffraction limit of was broken through and the image resolution was greatly improved by the lens with negative refractive index in theory.

Keywords: Photoacoustic Tomography, Acoustic Lens, Negative Refraction, Image Resolution

Yunshan Sun, Liyi Zhang, Haiyan Zhang Frequency domain minimum error probability medical CT image blind equalization algorithm

*Computer Modelling & New Technologies 2014 **18**(4) 296-299*

A frequency domain minimum error probability medical CT image blind equalization algorithm was proposed. Blind image equalization is implemented by minimizing a cost function consisting of estimated image and blur. The steepest descent method was adopted to solve the proposed cost function. Computer simulation experiments show that the new algorithm reduces mean square error and improves restoration effect, peak signal to noise ratio and improving signal to noise ratio.

Keywords: Blind equalization algorithm, minimum error probability, medical CT image