Communications
in Computer and Information Science   323

Tianyuan Xiao   Lin Zhang
Minrui Fei (Eds.)

AsiaSim 2012

Asia Simulation Conference 2012
Shanghai, China, October 27-30, 2012
Proceedings, Part I

Springer

Volume Editors

Tianyuan Xiao
Tsinghua University
Department of Automation
National CIMS Engineering Research Center
Beijing 100084, China
E-mail: xty-dau@tsinghua.edu.cn

Lin Zhang
Beihang University
School of Automation Science and Electrical Engineering
Beijing 100191, China
E-mail: johnlin9999@163.com

Minrui Fei
Shanghai University
School of Mechatronics Engineering and Automation
Shanghai 200072, China
E-mail: mrfei@staff.shu.edu.cn

Preface

The Asia Simulation Conference and the International Conference on System Simulation and Scientific Computing 2012 (AsiaSim & ICSC 2012) was formed to bring together outstanding researchers and practitioners in the field of modeling and simulation and scientific computing areas from all over the world to share their expertise and experience.

AsiaSim & ICSC 2012 was held in Shanghai, China, during October 27–30, 2012. It was constituted by AsiaSim and ICSC. AsiaSim is an annual international conference organized by three Asia Simulation Societies: CASS, JSST, and KSS since 1999. It has now become a conference series of the Federation of Asia Simulation Societies (ASIASIM) that was established in 2011. ICSC is a prolongation of the Beijing International Conference on System Simulation and Scientific Computing (BICSC) sponsored by CASS since 1989. AsiaSim & ICSC 2012 was organized by the Chinese Association for System Simulation (CASS) and Shanghai University. In the AsiaSim & ICSC 2012 conference, technical exchanges between the research community were carried out in the forms of keynote speeches, panel discussions, as well as special sessions. In addition, participants were also treated to a series of social functions, receptions, and networking sessions, which served as a vital channel to establish new connections, foster everlasting friendships, and forge collaborations among fellow researchers.

AsiaSim & ICSC 2012 received 906 paper submissions from eight countries. All papers went through a rigorous peer-review procedure including pre-review and formal review. Based on the review reports, the Program Committee finally selected 298 good-quality papers for presentation at AsiaSim & ICSC 2012, from which 267 high-quality papers were then sub-selected for inclusion in five volumes published in the Springer *Communications in Computer and Information Science* (CCIS) series.

This proceedings volume includes 63 papers covering five relevant topics including modeling theory and technology, M&S technology on synthesized environments and virtual reality environments, pervasive computing and simulation technology, embedded computing and simulation technology, and verification/validation/accreditation technology. All of these offer us plenty of valuable information and would be of great benefit to the technical exchange among scientists and engineers in modeling and simulation fields.

The organizers of AsiaSim & ICSC 2012, including the Chinese Association for System Simulation and Shanghai University, made enormous efforts to ensure the success of AsiaSim & ICSC 2012. We hereby would like to thank all the members of the AsiaSim & ICSC 2012 Advisory Committee for their guidance and advice, the members of the Program Committee and Technical Committee and the referees for their effort in reviewing and soliciting the papers, and the members of the Publication Committee for their significant editorial work. In

particular, we would like to thank all the authors for preparing, contributing, and presenting their excellent research works. Without the high-quality submissions and presentations from the authors, the success of the conference would not have been possible.

Finally, we would like to express our gratitude to the National Natural Science Foundation of China, the Japanese Society for Simulation Technology, Korea Society for Simulation, the Society for Modeling and Simulation International, International Association for Mathematics and Computer in Simulation, Federation of European Simulation Societies, Science and Technology on Space System Simulation Laboratory, Beijing Electro-Mechanical Engineering Institute, Shanghai Electro-mechanical Engineering Institute, and Shanghai Dianji University for their support in making this conference a success.

July 2012   Bo Hu Li
                                                        Qinping Zhao

AisaSim & ICSC 2012 Organization

Honorary Chairs

Chuanyuan Wen, China	Robert M. Howe, USA	Osamu Ono, Japan
Sung-Joo Park, Korea	Myoung-Hee Kim, Korea	Mahammad Obaidat, USA
Sadao Takaba, Japan	Xingren Wang, China	Zongji Chen, China

General Chairs

Bo Hu Li, China
Qinping Zhao, China

General Co-chairs

Koyamada Koji, Japan	Jonghyun Kim, Korea	Axel Lehmann, Germany
Qidi Wu, China	Song Wu, China	Zicai Wang, China
Xianxiang Huang, China	Khalid Al-Begain, UK	

International Program Committee

Chairs

Tianyuan Xiao, China
Lin Zhang, China

Co-chairs

Bernard Zeigler, USA	Tuncer Ören, Canada	Ralph C. Huntsinger, USA
Xiaofeng Hu, China	Fengju Kang, China	Soo-Hyun Park, Korea
Satoshi Tanaka, Japan	Zaozhen Liu, China	H.J. Halin, Switzerland
Xudong Pan, China	Kaj Juslin, Finland	Roy E. Crosbie, USA
Ming Yang, China	Xiaogang Qiu, China	Satoshi Tanaka, Japan
Jin Liu, China	Min Zhao, China	Shiwei Ma, China

Technical Committee

Agostino Bruzzone, Italy	Anxiang Huang, China	Yoonbae Kim, Korea
Yu Yao, China	Fei Xie, USA	Toshiharu Kagawa, Japan

Giuseppe Iazeolla, Italy   Mhamed Itmi, France   Haixiang Lin,
                                                     The Netherlands
Henri Pierreval, France   Hugh HT Liu, Canada   Shengen Zhou, China
Wolfgang Borutzky,   Jong Sik Lee,Korea   Xiaolin Hu, USA
    Germany
Yifa Tang, China   Wenhui Fan, China   Mingduan Tang, China
Long Wang, China   Doo-Kwon Baik, Korea   Shinsuke Tamura, Japan
Pierre Borne, France   Ratan Guha, USA   Reinhold Meisinger,
                                                     Germany
Richard Fujimoto, USA   Ge Li, China   Jinhai Sun, China
Xinping Xiong, China   Gary S.H. Tan,   Francesco Longo, Italy
                             Singapore
Hong Zhou, China   Shin'ichi Oishi, Japan Zhenhao Zhou, China
Beike Zhang, China   Alain Cardon, France   Xukun Shen, China
Yangsheng Wang, China   Marzuki Khalid,   Sergio Junco, Argentina
                             Malaysia
Tieqiao Wen, China   Xingsheng Gu, China   Zhijian Song, China
Yue Yang, China   Yongsheng Ding, China Huimin Fan, China
Ming Chen, China

Secretaries

Ping Zhang, China
Li Jia, China

Publication Chairs

Huosheng Hu, UK
Fei Tao, China

Special Session Chair

Shiwei Ma, China

Organizing Committee

Chairs

Minrui Fei, China
Yunjie Wu, China

Co-chairs

Ping Zhang, China
Linxuan Zhang, China
Noriyuki Komine, Japan
Kang Sun Lee, Korea

Members

Shixuan Liu, China	Xiao Song, China	Ni Li, China
Baiwei Guo, China	Gang Zhao, China	Yanxia Gao, China
Yulin Xu, China	Tingzhang Liu, China	Shaohua Zhang, China
Xin Li, China	Li Jia, China	Xin Sun, China
Qun Niu, China	Min Zheng, China	Ling Wang, China
Shouwei Gao, China		

Awards Committee

Chair

Zongji Chen (China)

Co-chairs

Axel Lehmann (Germany)
Soo-Hyun Park (Korea)
Wakae Kozukue (Japan)

Members

Satoshi Tanaka (Japan)
Sung-Yong Jang (Korea)
Wenhui Fan (China)
Yifa Yang (China)
Xiao Song (China)

Table of Contents – Part I

The First Section: Modeling Theory and Technology

The Second Section: Modeling and Simulation Technology on Synthesized Environment and Virtual Reality Environment

The Third Section: Pervasive Computing and Simulation Technology

The Fourth Section: Embedded Computing and Simulation Technology

The Fifth Section: Verification, Validation and Accreditation Technology

Table of Contents – Part II

The First Section: Networked Modeling and Simulation Technology

The Second Section: Modeling and Simulation Technology of Continuous System, Discrete System, Hybrid System, and Intelligent System

The Third Section: High Performance Computing and Simulation Technology

The Fourth Section: Cloud Simulation Technology

Table of Contents – Part III

The First Section: Modeling and Simulation Technology of Complex System and Open, Complex, Huge System

The Second Section: Simulation Based Acquisition and Virtual Prototyping Engineering Technology

The Third Section: Simulator

The Fourth Section: Simulation Language and Intelligent Simulation System

The Fifth Section: Parallel and Distributed Software

The Sixth Section: CAD, CAE, CAM, CIMS, VP, VM, and VR

The Seventh Section: Visualization

Research on Behavior Model of Virtual Soldier

Qing Xue, Qing Deng, Jian Sun, and Huibo Gao

Academy of Armored Force Engineering, Beijing, China
154247597@qq.com

Abstract. In the Urbanized Terrain Simulation, it is an important task to build behavior model of virtual soldier for virtual city and to achieve independent actions. An object-oriented behavior model of virtual soldier, which has simple and comprehensive structure, is proposed in this paper. It includes basic reaction hierarchy, control hierarchy and object-oriented behavior hierarchy. And detailed action choice rules, hierarchy configuration and inner state model behavior of the behavior model are presented. The model is implemented based on the method of coalescent pattern to simulate veritably soldier behavior.

Keywords: behavior model, hierarchy structure, target oriented.

1   Introduction

Virtual soldier is an important part in the battle simulation system. It can consumedly reduce combat force and the cost of simulation training by utilizing virtual soldiers instead of training opponents

The paper mainly discussed how to establish the real soldier operational behavior model. Consequently, they apperceived information in the virtual environment, combined with their own goals and current intramural states. Then they chose exact tactical actions and called the corresponding behavior program for implementing appropriate tactical actions to complete the current mission.

2   Behavior Model Configure

The virtual soldier responded to their internal properties and external battle environmental information. According to Zoology and artificial life theory, behavior has different levels and is composed of basic behaviors [1]. A variety of behaviors can implement serial or parallel execution. Hierarchy Structure has prominent advantages in dealing with conflict and information share between behaviors [2].

(1) Hierarchy structure had a high intuitionism.

(2) High level behaviors can accommodate and obtain automatically access to low level ones.

(3) Each behavior in the hierarchy structure focused on their goals which were determined by arbitration mechanism.
(4) Hierarchy structure contributed to problem decomposition.
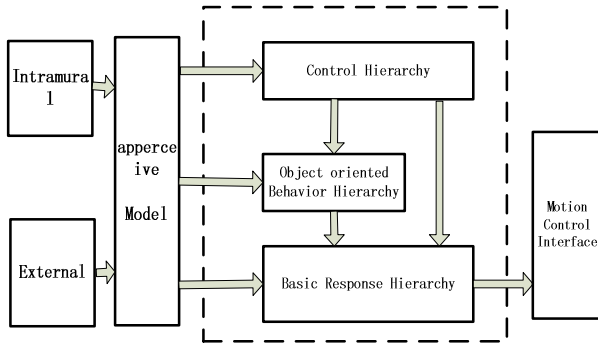
Fig. 1. Behavior model framework of virtual solider

The virtual soldier behavior model adopted hierarchy goal-oriented model. Shown in Figure1, it can be divided into three layers: basic response behavior layer, goal-oriented behavior layer and control layer. Basic reaction behavior layer responded with changes in the environment which belonged to - response rules. Object-oriented behavior layer was goal oriented complex behavior to environmental stimulus, internal state and current target. Object-oriented behavior layer and basic behavior layer were linked with container structure. That can achieve parallel execution between basic behaviors and goal-oriented behaviors. Control layer was the top level of behavior model, responsible for the behavior choice and termination. This model ensured the virtual soldiers response to real-time emergency, also exerted the behavior planning capacity to complete complex tasks.

3   Behavior Model Structure

The behavior model contained three important components: internal state model, behavior level structure and behavior selection mechanism. The virtual soldiers' behavior level structure was conducive to construct the whole behavior system. Behavior selection mechanism was to resolve how to choose the most appropriate behavior from competitive behaviors.

3.1   Intramural State Model

In the virtual soldier internal state model, an internal variable was designed: tiredness. Tiredness increasing can produce a rest behavior motivation. The size of tiredness affected velocity of virtual soldiers. Different values of fatigue can provide diverse maximum speed.

Virtual soldiers' tiredness model was defined as follows:

$$Tired_t = \begin{cases} Max\{Tired_{t-\Delta t} + k(V_t - V_{Normol}) \times \Delta t \ , \quad Tired_{Max}\} & (V_t \leq V_{Normol}) \\ Min\{Tired_{t-\Delta t} + k(V_t - V_{Normol}) \times \Delta t \ , \quad Tired_{Min}\} & (V_{Normol} \leq V_t \leq V_{HStrenth}) \\ Min\{Tired_{t-\Delta t} + k(V_t - V_{HStrenth})^2 \times \Delta t \ , \ Tired_{Min}\} & (V_{HStrenth} \leq V_t) \end{cases}$$

$$Tiredness_{Min} \leq Tiredness_t \leq Tiredness_{Max}$$

V_t represented the moment speed of the virtual soldier, the constant value $V_{Normorl}$ was speed upper limit in the rest state and $V_{HStrenth}$ lower limit .K represented change rate.

When the virtual solider Speed was less than the constant value the tiredness decreased over time. While the speed was greater than the constant value the tiredness increased.

A state change event was produced in the case that internal Property values exceeded the designed threshold . It implemented behavior trigger.

3.2   Behavior Hierarchy

The virtual soldiers conducted an inclusive bottom-up style structure. Behavior was divided into two layers: the bottom was basic behavior aggregate and the high level was goal-oriented behavior aggregate.

(1) Basic behavior layer

Basic behavior layer was consist of virtual soldiers' fundamental tactics behaviors supporting goal-oriented behaviors. It is the combat unit capability in the virtual battlefield environment [3]. Virtual soldiers can carry out multiple tactical activities which constituted the complete sequence of a target behavior. The basic behavior of the various tactics described public property and constituted the basic model of tactical actions to improve work efficiency.

Basic behavior layer not only determined the capacity of virtual soldiers' basic behavior but also affected the further expansion of combat simulation system. The appropriate basic behavior set choice should consider behavior properties and task demand. Mataric proposed a behavior selection standard: "We believe that, for each problem, there is a set of basic actions, the rest of the behavior can be derived from this collection. "[4] those chosen behaviors were necessary and met the corresponding task demand.

(2) Goal-oriented behavior layer

Behavior science research indicated that human always decided something according to current environmental conditions and their own needs, the thing that would be done was called a goal [5]. If this goal can be done one step, it was defined as an atomic target. Or it was defined as an abstract goal which was a combination of atomic targets. Human chose high abstract goals and then recursively decomposed into implemental action plans. [6] In this article, aims were tasks that the virtual soldier needed to complete.

Based on Goal-oriented behavior control mechanism, virtual soldiers had a series of high level goals, which may be abstract or atomic goals. Virtual soldiers used arbitration module to select the most appropriate target to execute. At this level, the handling of virtual soldiers was consistent.

3.3   Behavior Choose Mechanism

(1) goal arbitration
Virtual Soldiers in a complex and dynamic urban combat simulation environment tried to meet many combat goals. But they had limited resources. Consequently, it was necessary to require some kind of arbitration mechanism to arbitrate high level goals. Moreover, once virtual soldiers focused on achieving a goal. They should be absorbed in this goal continuously unless there was much more important goal implemented. Target arbitration was the core of the behavior model. It can arbitrate the current high level goals and choose the most suitable high level goal.

The expected goal value was calculated by the relevant internal and external factors. To avoid small changes of the perception variables leading to the current behavior oscillating between two acts, this paper presented a goal arbitration model based on inertia coefficient weight.

Each simulation cyce selected the target as following steps:

step 1: update the expected parameters of strategic goal value according to the input sensor;
step 2: Calculate each strategy goal value;
step 3: Calculate the activation value of the goal $Activation_{Gi}$;

step 4: Compare activation values of all the strategic goals;
step 5: treat the goal with the maximum activation value as the current goal.

Calculate the expected value of goal by using formula 1

$$Expectation_{Gi}^{t} = k \times comb(\sum_{j} r_{ij}^{t}, \sum_{j} e_{ij}^{t}) \tag{1}$$

Calculate the activation value of goal by using formula 2.

$$Activation_{Gi}^{t} = \beta_{i}^{t} \times Expectation_{Gi}^{t-1} + Expectation_{Gi}^{t} \tag{2}$$

Inertia coefficient β_{i}^{t} affected the goal durative. If it was lower the effect changing oscillations between two behaviors would not be obvious, or the behavior continuity would be enhanced. That made the low level behavior performs difficultly. To solve this problem, Ludtow proposed a tiredness coefficient associated with each behavior. [8] The tiredness coefficient value varied between 0 and 1. The value of the tiredness coefficient and behavior multiplied during the model calculations. When the behavior is activated the tiredness coefficient decreased. When the behavior is not active, its inhibition coefficient increased. This made the lower activation behavior have an opportunity to perform like time-sharing system.

(2) Behavior planning based on goal-oriented

The goal that arbitration outputted showed the current behavior goal of virtual soldiers. In the simulation, if the current target was a complex task, the virtual soldier would first plan to determine a series of sub-goals for transferring basic behaviors to achieve sub-goals. Behavior planning processes were established as follows: Planning goal. Through internal demand and external stimulation, the arbitration module generated current target goals;

Planning premise. Virtual battlefield environment information and its own internal properties;

Planning body. Virtual soldiers coordinated their behavior and identified the best course under certain constraints;

Planning result: the completion of the current target was successful or failed.

Take virtual soldiers' aggressive behavior for example, the paper explored the behavior planning algorithm. Firstly, the algorithm was definited as follows: n represented the enemy number virtual Soldiers perceived, d was the distance between the virtual soldier and the enemy.

Planning goal: attack; Planning premise: the virtual soldier was energetic and weapons were enough;

Planning body:

step1: Search the enemy.

step2: The detection of virtual soldiers perception model output, if n=0, turned to step1; If n=1, turned to step 3, else turned to step 8.

step3: Estimate whether there was a bunker within the perceived scope.

step4: Enter the bunker;

step5: Lie down;

step6: Shoot;

step7: Judge the living condition of the enemy, if it was dead, turn to step 9, else turn to step 1;

step8: Avoid;

step9: End.

Planning result: when the enemy was killed it was successful, else failed.

4   Realization of Behavior Model

The combination model of object-oriented program provided a method to solute the goal class of goal-oriented behavior model. The combination model defined an abstract component class, which was not only an atomic component also a combination component. The abstract class defined a common interface. The user utilized the interface interact with different component objects. Shown in the Figure 2, the atomic goals were achieved by class Goal and combined goals were implemented by class Goal_Composite. The difference between them was that class Goal acted both as abstract component class and as atomic component class. Class Goal_Composite acted as combination components which derived from class Goal and were converged by a number of Goals.
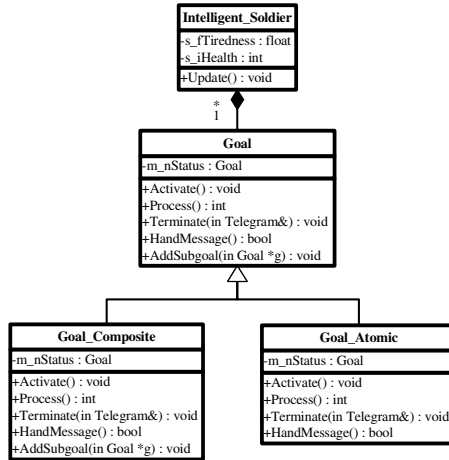
Fig. 2. Combination model of goal class

5   Conclusions

Based on behavior hierarchy theory and goal-oriented characteristics, a goal-oriented behavior model was built. The model can use the current task of virtual soldiers, the internal state and virtual battlefield information to plan behavior.

References

1. Song, Y.: The research of the agent technology and application in Computer generated forces. Harbin Engineering University (2006)
2. Ma, D., Sun, L., Tong, M.-A.: The collision detection algorithm in virtual scene in. Firepower Command and Control (2004)
3. Wu, X.: The research of artificial fish animation Based on behavior model technique. Zhejiang University (2004)
4. Wang, J.: Human perception model during Football match. Computer Simulation (2008)
5. Baerends: On drive, conflict and instinct, and the functional behavior. Perspective in Brain Research (1976)
6. Hu, P.: The concept model of Army units operations. Naval University of Engineering (2006)
7. Xue, Q.: Equipment combat simulation base. Armored Force Engineering Institute, Beijing (2007)
8. Pew, R.W., Mavor, A.S.: Representation Human Behavior in Military Simulations. Iterim Report (1997)

Analysis of Aperture Shape Changing Trend Base on the Shaped Charge Jet Penetration through the Steel Target

Junqing Huang, Yalong Ma, Kelei Huang, and Jianxun Zhao

Department of Equipment Command and Administration
Academy of Armored Force Engineering, 100072 Beijing, China

Abstract. Based on the characteristic of the linear shaped charge and the theory of jet penetration target, a method, which adopts explicit dynamic analyzing program AUTODYN to simulate the process of the linear shaped charge jet penetration through the target, is proposed in this paper. The trend and rule of aperture size changing with time during the process of the jet penetration the target, are analyzed. The physics phenomena and rule are basically same between emulational result and testing result. It shows that analytical model and emulational method are reasonable and practical.

Keywords: the shaped charge, jet, numerical simulation, aperture.

1   Introduction

With the rapid development of computer technology, the numerical simulation methods could comprehensively reflect the changes of parameters and physical quantities in the process of Jet penetrate target, and could graphically display the whole process of the penetration, the calculation procedure could select different physical and geometrical parameters to calculate the process of the penetration and estimate the effect of various parameters on the results of the projectile penetration the target, which could make up for the lack of theoretical analysis, to broaden the experimental results. Usually in the shaped charge jet penetration target trial, perforation size of jet penetration was acquired by the measure of cutting the target plates. Temporality, the perforation size was no longer change. Due to limitations of the experimental conditions, it was difficult to record the process of the change of the perforation size with time, and investigate the law of change. The perforation size change course could be clearly showed by numerical methods. In this paper, adopted explicit dynamic analyzing program AUTODYN basing on finite element to simulate the process of the linear shaped charge jet penetration through the target and perforation being created, and the emulational and testing results were compared, emphatically analyzed the trend and law of change of the perforation size with time in the process of the jet penetration target.

2   Basic Theory

The governing equations were:

(1) The particle equation of motion

Took the initial moment of the particle coordinates $X_i (i= 1,2,3)$, at any time t, the particle coordinates was $x_i (i= 1,2,3)$, the equations of motion of the particle was:

$$x_i = x_i (X_i, t) \qquad i = 1,2,3 \tag{1}$$

At $t = 0$, the initial condition:

$$x_i (X_i, 0) = X_i \tag{2}$$

$$\dot{x}_i (X_i, 0) = V_i (X_i, 0) \tag{3}$$

Where, V_i——initial velocity.

(2) Conservation of momentum equation

$$\sigma_i + \rho f_i = \rho \ddot{x}_i \tag{4}$$

Where, σ_i ——Cauchy stress;

f_i——per unit mass of volume forces;

$\ddot{x}_i$——acceleration.

(3) Conservation of mass equation

$$\rho = J \rho_0 \tag{5}$$

Where; ρ——the current mass density;

ρ_0——Initial mass density.

3   Numerical Simulation Model

According to the linear shaped charge characteristics, the model was simplified to the plane symmetry, and the 1/2 model was taken to calculation. The single-layer entity grid was used to build 3D models, the calculation model shown in Figure1. The numerical model was adopted cm-g-us system of units. The Euler grid modeling was used to build the models of explosives, liner and air. The element used the multi-material ALE algorithm. The Euler grid modeling was used to build the steel target model, and the coupling algorithm was used to reflect the mutual effect of between the steel target and air and liner material. The detonating manner was along the top of the line of explosive detonation. The Charge, liner, steel target and air calculation grids were hexahedral solid elements. The computational model was divided into 18567 grids.
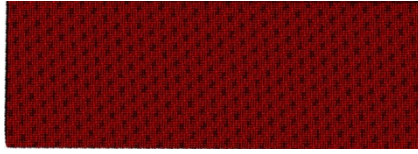
(a) the Warhead Model   (b) the Target Model

Fig. 1. The Jet Penetrating Target model

The material models include the material constitutive model, the equation of state, strength model. The liner, shaped charge, target and air materials were involved in the simulation of jet penetration target. The material models of the liner, charge, target and air shown in Table 1.

Table 1. Material Models

part name	material	the equation of state	strength model
liner	CU-OFHC	Grüneisen	Johnson-Cook
shaped charge	TNT	JWL	Hydro
target	RHA	Shock	Johnson-Cook
air	Air	Ideal-Gas	-
liner	CU-OFHC	Grüneisen	Johnson-Cook

The JWL equation of state was used to describe the shaped charge detonation products, formulation below, the material parameters of TNT shown in Table 2, the cm-g-us system of units was adopted.

$$P = A(1 - \frac{w}{R_1 V})e^{-R_1 V} + B(1 - \frac{w}{R_2 V})e^{-R_2 V} + \frac{wE}{V} \tag{6}$$

Table 2. Material Model Parameters of TNT

P_0	D	A	B	R_1	R_2	w	E_0
1.63	0.693	3.74	0.0323	4.15	0.95	0.3	2.7

The yield strength of copper was 120MPa, the shear modulus was 47.7GPa.

4   Conclusions Simulation Results and Analyzing

The calculation data of the numerical simulation showed that the perforation shape was continuously changing in the process of perforation form, as shown in Figure2. The radius and depth of the perforation were continuously changing with time in the process of the jet penetration target, as shown in Table 3. When the head of jet impacting the target, at the collision point, the high pressure and shock waves made

the target free interface crack, and the high temperature, high pressure, high strain rate region were formed in the target. The formation of small aperture hole depth accounted for only a small fraction of hole depth at this time, as shown in Figure 2 (a). In the process of penetration, aperture and depth of the hole were further increases. In the perforation expansion process, the central swelled rapidly, the expansion of both ends was slow, the shape of perforation at 658.2us was shown in Figure 2 (b). The cavity central drum began to form bulge, aperture and depth of the hole continued to increase. When the cavity of perforation expanded to a maximum, due to the elastic energy was released, the cavity of the perforation middle closing with rearward began to shrink, the contraction of tail radial was rapider and central radial was slower. At 5961.3us, the shape of perforating was no longer change. The stable perforation shape was shown in Figure2 (c).

(a) 658.2us   (b) 1035.0us

(c) 5961.3us
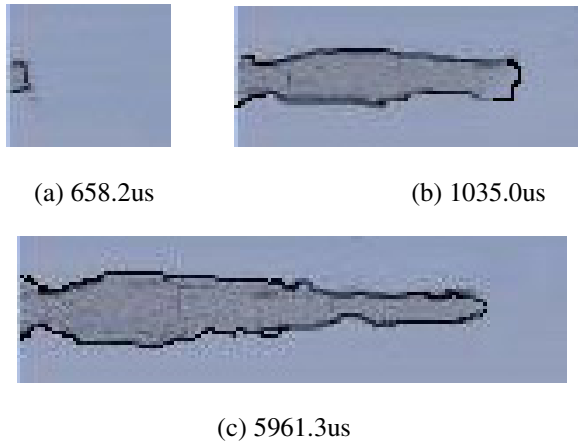
Fig. 2. The Perforation Shape in the Process of Penetration

Table 3. The Change of the Perforation Radius and Depth with Times

chart num	fig.2(a)	fig.2(b)	fig.2(c)
time/us	658.2	1035.0	5961.3
head radius/mm	27.9	32.1	46.0
middle radius/mm	24.7	45.8	67.9
tail radius/mm	24.9	42.8	29.2
depth/mm	21.0	264.0	406.0

The conditions of testing prepare were shown in Figure 3. The test conditions were as follows: The thickness of liner was 0.4cm, the shape of liner was cylindrical cone, the angle of liner was 36 °, the material was copper, dynamite was TNT and it's quality was 1.3kg. The material of target was adopted RHA, and the thickness of target was 800mm. The explosive manner adopted HEAT warhead vertical and immovable explosion. The height between the warhead and target was 600mm. The target was split to eight parts, the thickness of every partition was 100mm. Every part was accumulated together.

Fig. 3. The Prepare of Test
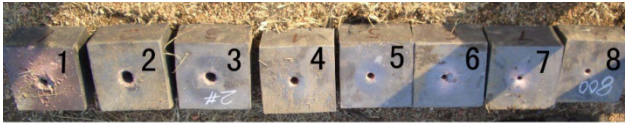
(a) the Test Result
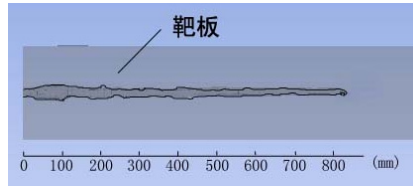
(b) the Simulation Result

Fig. 4. The testing result of the shaped charge jet penetrating target and simulation result

The testing result of the shaped charge jet penetrating target and simulation result were shown in Figure 4.

The results showed that the numerical results and the experimental results were basic coincident. The perforation depth was deep in the numerical results, had error in about 20% than the experimental results. The radius of perforation was small in the numerical results, had error in about 15% than the experimental results. In addition to the calculation error, the accuracy of the calculation results was mainly affected by the elastic- plastic model of describing material character of the target. In addition, the jet breakup and the effect of air were ignored. So, the difference should be considered in the analysis process.

5   Conclusions

The explicit dynamic analyzing program AUTODYN basing on finite element was adopted to simulate the process of the linear shaped charge jet penetration through the target. The numerical results showed that the central aperture of perforating the first increase to a maximum and then decreases to a certain value after the jet penetration target. The results are basic coincident by contrast between the numerical results and the experimental results, only have the error in about 20%. It indicates that the numerical simulation model used in this paper and the manner of numerical simulation were correct, can be used to the engineering design.

References

1. Anderson Jr., C.E., Bodner, S.R.: Ballistic impact: the status of analytical and numerical modeling. Int. J. Impact Engineering 7(1), 9–35 (1986)
2. Macek, R.W., Duffey, T.A.: Finite Cavity Expansion Method for near-surface Effects and Layering during Earth Penetration. International Journal of Impact Engineering 24, 239–258 (2000)
3. Ramesh, K.T.: Localization in Tungsten Heavy Alloys Subjected to Shearing Deformations Under Superimposed High Pressures. Metal Powder Industries Federation, 3–9 (1995)

A Commanding Model of Counterwork Simulation System Based on Value Driving Decision-Making

Weifeng Li[1], Shuli Wang[2], Xinjun Zhao[1], and Jiying Wang[1]

[1] Institute of Special Vehicle, No.6 Lab, No.3 Yuan, Fengtai District,
Beijing 100072, China
[2] Academy of Armored Force Engineering, Fengtai District, Beijing 100072, China
superfishmen@163.com

Abstract. A commanding model of counterwork simulation system based on value driving decision-making is proposed in this paper. The theoretical foundation of the method is Nash Equilibrium of game theory. This paper analyzes the value driving decision-making model which is based on entity's maximum benefit and gives examples of value driving decision-making in the Counterwork Simulation System.

Keywords: Value Driving Decision-making, Nash Equilibrium, Game Theory, Psychological Model.

1   Introduction

The Counterwork Simulation System is a system for researching many entities which confront with each other in the real war. It is a distributing experiment environment. Every experiment entity connects with each other by network .They are managed by uniform resource administer and experiment controller .This system is capable to resource sharing and experiment combining. This system is used for evaluating counterwork effects. It can be used for single equipment evaluation and system evaluation. And the experimenting and evaluation can analyze issues in any levels.

Decision-making is the cerebra of this system. And it is the most important and most difficult part of the system. The traditional model of decision-making in counterwork simulation system is the way of descending. Firstly evaluating all of the aims and making the aims line by the numerical value from evaluation. Secondly matching the most powerful weapon with the most important aim, and then matching the second powerful weapon, and so on. This way of decision-making is convenient and the result of simulation is logical. But the result of simulation is the best result of all capable results. The real result is different from the best result sometimes. And the difference between them is essential. It is proved the way of descending is not useful concerning on different force distributions in some interesting researches.

2   The Decision-Making Model of Counterwork Simulation System

The problem of counterwork system is the way of matching the thousands of arms with the thousands of aims. This arrangement is based on the distribution of enemy's aims and our mission. Besides the effects of the stratagem of friend forces, the action of any other forces can not be neglect in the counterwork simulation system. When a commander makes his decision, he must analyze the most probable decision-making of friend forces and the most probable decision-making of enemy. He wishes he can get the most benefit in this counterwork. .

The counterwork simulation system which manufactured by our team makes it comes true that several entities confront with several entities. Each decision-making entity gets information by inner sensor and communication link connecting with other entities. The data collected by this way pass to a inner perceptive model. It is made up of elementary senses which come from inner sensor and communication link and models which are made up in advance. In the course of simulation, the beginning value of decision-making is made up by the rule of decision-making which is controlled by value driving model. In a simulation, every decision-making entity uses its own sensor model to analyze the information of inner sensor and communication link connecting with other entities. Then the decision-making entity makes possible action aggregate. The decision-making entity uses its value evaluation model to evaluate any actions. Lastly, the high score action is the decision which is executed by simulation entity. This way of making decision can reflect the impact of command and do research of the system of command in real war.

3   A Commanding Model Based on Value Driving Decision-Making

3.1   Value Driving Decision-Making

Counterwork is a kind of Game. It is the theoretic basic of a Commanding Model Based on Value Driving Decision-making. Apparently, either the Games between enemy and our team or teams on the same side are dynamic no-cooperative Game. The members who take part in the Game want the maximum benefit and the minimum loss. So the psychological model of members in Game, namely value driving decision-making model aims to the maximum value. This model is entitled Nash Equilibrium in the Game—Theory which can bring maximum benefit.

Nash Equilibrium is called no-cooperative Game Equilibrium, too. The definition of Nash Equilibrium:

Supposing N members take part in a certain Game. In a precondition that every other body have made their strategy or decision, you choose the best one in the all possible decision-making (This decision-making maybe depend on the decision-making of other body, and maybe not.) to get the maximum benefit. All the decision-making of members in the Game constitutes a Strategy Profile. Nash Equilibrium is the one of the Strategy Profile. Everyone in the Nash Equilibrium thinks the decision-making of own

is the one which will bring the maximum benefit. Therefore, none of the members wants to break down the Nash Equilibrium.

Apparently, Nash Equilibrium is made up of single best decision-making, but the total benefit is not the maximum one sometimes. For example, the members in the Game have opposite benefit, but everyone wants his own maximum benefit. And then the benefit of different side will counteract each other. At last the Nash Equilibrium becomes an equilibrium make against everyone. The precondition of the Nash Equilibrium psychological model is that everyone in the battlefield is selfish. Without the superior compulsive order, everyone wants the maximum benefit from his decision-making and actions. But this equilibrium is always the adverse one in the real-life, especially in the war. So the Nash Equilibrium psychological model is a good model in the counterwork simulation system. Based on the Nash Equilibrium, value driving decision-making is a good application of Game—Theory in the domain of military affairs.

The value driving decision-making model of this system is:
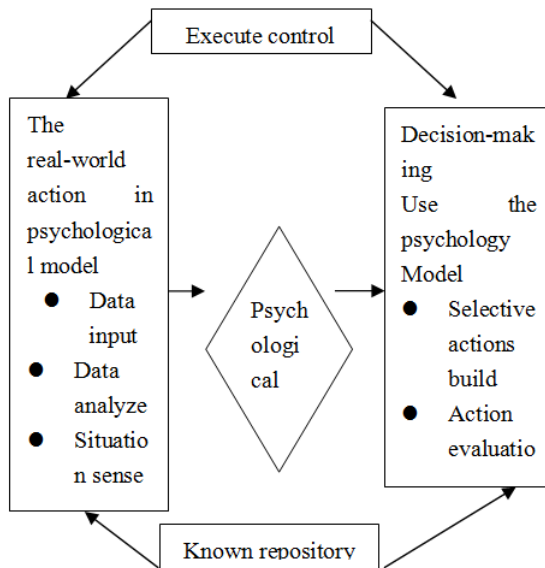
Fig. 1. The Value Driving Decision-making Model

The value driving decision-making model can make appropriate decision in real-world by using mature computer decision-making arithmetic. And it connects the numerical value with selective actions by using the theory of maximum benefit.

3.2   The Description of the Action of Decision-Making Model

The action of decision-making model in counterwork simulation system is:

The entities get information, order and the situation of battlefield from inner sensor and communication link connecting with other entity. Then classify them and analyze them.

Use the inner sensor model to classify the data. On the one hand executing all high-up's order without any condition, on the other hand matching the situation of battlefield and information with the project in command-database. This match is one information opposite to several selective actions by the cost of the action, how to carry out the action, the capability of the entity and advantage from the action. These projects are reasonable in tactics. But the entity can get different benefit from carrying out them.

Use the value model to evaluate every feasible action and get the value of this action. The primary index of the value model is: the maximum benefit for the entity itself, the maximum benefit for their team, the maximum benefit for friend forces and so on. All of the index, the maximum benefit for the entity itself is the most important one. The entity makes the final decision by the result of evaluation.

Make the decision into several idiographic steps. Ensure the action plan parameter. The plan is a structure, including aim, mission, the condition of touching off, the condition of success and the condition of invalidation. There are some successful cases in plan database. So the appropriate plan can be chosen from the database. For example, the action plan parameter of attack is:X battle unit, with X material, at X hour X minute, go across X spot ,the destination is X, the mission is X. The other entities can receive the action plan from the decision-making entity.

3.3   The Logical Structure of the Commanding Model

Based on the description of function and process of the commanding model of counterwork simulation system, the logical structure of the commanding model of value driving decision-making is:
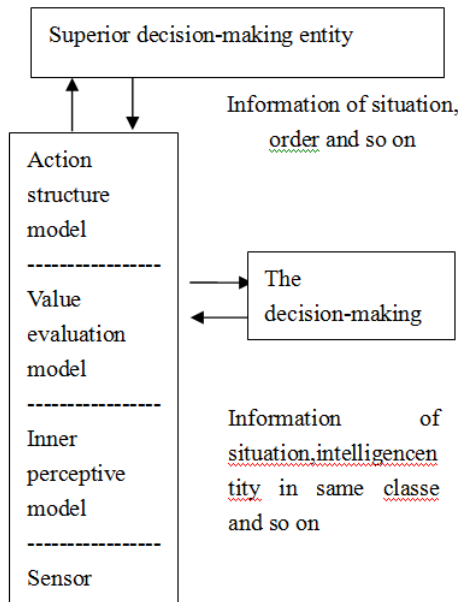
Fig. 2. The Drawing of the Commanding Model of Value Driving Decision-making

3.4   The Examples of the Commanding Model Based on Value Driving Decision-Making

The counterwork simulation system does research on several entities confront with several entities. A commanding entity should calculate the amount of enemy, the ability of enemy and the intention of enemy, and it also should calculate the cooperation of friend forces. It is a complex task. Next, it will give two typical situations of entity making its decision as examples .The two examples can show you how to use the commanding model of value driving decision-making.

Case 1: The commanding of a single entity vs. several entities. Condition:

The information given entity: the task is headed off the enemies. There is only one tank(C) in our side in covert. The rate of hit the target of first shooting is 100%. Once the tank shoots, the enemies can find it. Enemy has two tanks A and B. A and B will pass the covert in few minutes. The rate of hit the target of first shooting of A is 30%, B is 80%.

According the information, the decision-making entity get 2 conceivable project:

Project 1: Shoot A firstly. With the calculation of value evaluating model, the rate of shot by B is 80%.

Project 2: Shoot B firstly. With the calculation of value evaluating model, the rate of shot by A is 30%.

Conclusion: With the calculation of value evaluating model, the value of shooting B firstly is larger than shooting A firstly. So the decision-making entity will choose project 2 because it can finish the mission of heading off the enemies and C can get more survival chances.

Case 2: The commanding of cooperation with friend forces. Condition:

The information given entity: B asks A for cooperation. The booty of either A (the decision-making entity) or B attack enemy is 4. The booty of both A (the decision-making entity) and B attack enemy is 10. It is the booty of single attack is (4, 4), and the booty of both attack is (10, 10).

When A and B attack together, the distribution of booty is not always average.

Based on information above, the decision-making entity gets 4 conceivable projects:

Project 1: The distribution of booty is (10, 10). With the calculation of value evaluating model, it is an impartial distribution.

Project 2: Because of some reasons, the distribution is not the impartial one. But the distribution is like this (8, 12) or (6, 14). With the calculation of value evaluating model, the booty A can gain is more than that A attack enemy itself or do not take part in the battle. The value of attack together is larger than A attack itself. So the distribution of booty for A is still the maximum benefit.

Project 3: The booty of A is more than 10. With the calculation of value evaluating model, the distribution of booty for A is the maximum benefit.

Project 4: The booty of A is less than 4. With the calculation of value evaluating model, the value of attack together is less than A attack itself. So when Project 4 happens, A and B can not manage to cooperation.

Conclusion: With the calculation of value evaluating model, 4 is the critical spot of cooperation. When the booty of A is less than 4, A and B can not manage to cooperation.

For making clear the basic issue, we use some simple data and predigest the value evaluating model of the commanding model of counterwork simulation system in these 2 examples. In the real counterwork simulation system, he value evaluating model is more complicated. The result of counterwork simulation system is close to the real war.

4   The Merit of Value Driving Decision-Making

The traditional decision-making rule is related with the methods of decision-making table. It is a simple way to deal with the decision-making in the counterwork simulation system. In a very war situation, the commander of each side of the war must do something to control the situation of the war like attack, defense or drawing back. There are a series of threshold numbers which are considered as critical numbers of the conversion of situation. The proportion of some factors of the both sides of the counterwork is connected with the critical numbers. The simple example is the numbers of arms. 3:1 is the critical number of attack, and 1:8 is the critical number of defense. It is used the proportion of current forces as the critical number of the traditional decision-making rule. So the traditional decision-making rule is not the dynamic decision-making.

The value driving decision-making is more flexible than the traditional decision-making .The merit of value driving decision-making is:

The value driving decision-making makes the decision and layouts the steps of actions dynamically. It can make the counterwork simulation system more factually than the traditional decision-making.

There are not explanation of data and sensors of situation in traditional decision-making rule.

The value driving decision-making makes the command by many factors. The counterwork simulation system based on the value driving decision-making is closer to the real war.

5   Conclusions

From the description of the commanding model, we can make sure that the value driving decision-making based on Nash Equilibrium is a good method in counterwork simulation system. Using the value driving decision-making, the counterwork simulation system is closer the real war. This paper has discussed several commanding models of the value driving decision-making and offered the value evaluate model

based on the entity's maximum benefit. We have given 2 examples of the value driving decision-making. In the counterwork simulation experimentation, the value driving decision-making still needs to be developed.

References

1. Bestavos, A.: Advances in Real-time Database Systems Research. SIGMOD Record 25(1), 221–223 (1996)
2. Pew, R.W., Mavor, A.S.: Modelling human and organization behavior: Application to Military Simulations, pp. 335–341. National Academic Press, Washington, D.C. (1998)
3. Nash, J.: No-cooperative Game, pp. 10–11 (1952)

Research on Assembly Line Modeling and Simulation Optimization

Fuli Huang, Wenhua Zhu, Bin Bai, and Baorui Li

CIMS and Robot Center, Shanghai University, Shanghai 200072, CHN

Abstract. In this paper, assembly line is discussed with plant simulation. Firstly, the method of assembly line modeling is proposed. Secondly, assembly line simulation optimization is researched. Finally, assembly line modeling and simulation optimization are both described by an instance.

Keywords: Assembly line, Plant Simulation, Modeling, Simulation optimization.

1   Introduction

Assembly line is the effective combination of Man and machine, reflecting the flexibility of the equipment fully, integrating conveying system, accompanying jig and testing equipment organically in order to meet the requirements of product assembly. According to the type of object assembly, assembly line can be divided into Single-Model Assembly Line, Multi-Model Assembly Line and Mixed-Model Assembly Line. No matter which type of assembly line, generally has the following basic characteristics [1]: 1) High level of professionalization. In the assembly line, one or several products are produced fixedly; in every workstation, one or several processes are completed fixedly.2) The production with obvious rhythmic. The time of process on each workstation should meet the beat of requirements for ease to avoid the products packing and waste time in workstation, since the driving devices move with a certain speed. 3) The workstations are put in order according to the process sequence , assembly object does unidirectional movement between the workstations.

Assembly line is discrete event system, which widely used in the manufacturing industry. The traditional assembly line simulation is based on mathematical model, which is time-consuming. Generally, it can't meet the demand of product's fast response to the market. In this situation, the computer simulation arises at the historic moment, which can quickly and accurately finish simulation, as a result of shortening the production cycle, reducing manufacturing costs. There are many computer simulation software. Especially, Plant Simulation is object oriented, graphical, integration of modeling, simulation tools, which used to call eM-Plant until the takeover of Siemens. Plant Simulation has the following characteristics [2]: 1) The object-oriented modeling. As a modeling object, properties and methods are encapsulated as a separate entity in Plant Simulation, which can be accessed by other

objects. 2) Support layer type structure. Layered modeling can be realized and other hierarchical structure can be added in the course of modeling. 3) The WYSIWYG graphics work environment. Modeling, simulation and animation can be intuitive displayed in the interface of Plant Simulation.4) Easy to control. SimTalk language embedded in Plant Simulation can achieve fine, flexible control. 5) The data processing is simple. Plant Simulation provides various forms of interfaces, 3D display and HTML report. Because of the above features, Plant Simulation is applied to research assembly line in this paper.

2   Assembly Line Modeling

2.1   Assembly Line Modeling Basis

Assembly line modeling needs to plan organization structure, which is very important for modeling, especially when assembly line is more complex, as well as the situation of requiring different person to complete different parts. If the structure of the organization is well organized, it's easy to find the object in modeling process. It is not only easy to manage, but also can improve the efficiency of the modeling effectively with the good organization structure. The organizational structure of assembly line usually have a feeding workstation, assembly workstation, grinding-in workstation, detection workstation, laying-off workstation, mobile equipment, transportation equipment, lifting equipment, rotating equipment and testing equipment, etc. Object library must be corresponded to simulation software library after planning organization.

Plant Simulation provides objects library of the basic modeling, including logistics object, information flow object, the user interface object and moving object. Logistics object have the ability to change the moving objects parameters of the object, including workstation, buffer, feeding workstation and laying-off workstation, etc; Information flow object can control and record the data of simulation, and most of the information flow objects are presented in the form of table, including variables, table, storage file, trigger and circulation, etc; The user interface object is the method of the communication between simulation model and the user, which not only can provide system simulation of relevant information for users, but also can be a control simulation tool for users; The moving object refers to the object that physical location is not limited to one place, such as the assembly parts of the assembly line, the vehicle and tray of transport parts. The system model can be built according to actual production with inheriting and duplicating objects. Practical application is also convenient. The corresponding object model can be generated with the mouse click and the drag of corresponding object icon.

2.2   The Design Method of Assembly Line Modeling

Assembly line modeling needs to base on the actual resources. Firstly, the information of assembly line should be collected into the resource database of assembly line, then

the information will be classified in the aspect of organization structure according to function, the classification of this treatment is easy for corresponding actual assembly line in modeling.

Modeling is the process of constantly extracting Plant Simulation object library according to the information resource of assembly line. In the course of modeling, constraint model will be needed, the constraint conditions is different as the different products, this is because regardless of which kind of assembly lines are for the products, and different products have different processes, namely the assembly process library restricts the arrangement and order of modeling in assembly line. Figure 1 is the specific design flow chart of assembly line.
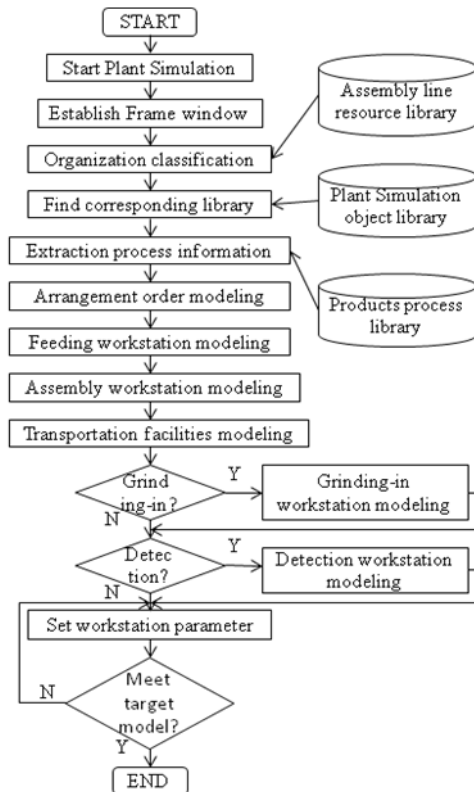
Fig. 1. Flow chart of assembly line modeling design method

3   Assembly Line Simulation Optimization

The simulation operation of the simulation model can only provide feasible scheme in certain conditions, not giving the optimal solution of the problem, and simulation optimization technique is general embed in the process of simulation, which is the

process finding the best input variable value from feasible domain, making the output be the optimal solution or satisfied solution, and the goal is to consume the least resource while get the most information in simulation experiment , and let the user decide easily [3].

Plant Simulation has many assembly line optimization tools, such as genetic algorithm, bottleneck analysis, which can find the bottleneck of restricting production and optimize the system allocation parameters with genetic algorithm. The optimization of assembly line is realized through the followings methods.

3.1   Equipment Layout of Assembly Line

Production equipment layout and material handling system design impact the production costs and profits of manufacturing enterprise , Tompkins research thinks a manufacturing enterprise materials handling cost takes up 10% ~ 30% of the total cost in production, from production efficiency   it can be highest improved three times[4]. This shows, the reasonable design equipment layout is effective way to reduce material flow rate to optimize the assembly line. Equipment layout is typical problem of quadratic assignment, it is the design considered from the process principle, and the following is its design method.

1) The known n equipments and between handling quantity size is W_{ij} (i=1,2,…,n, j=1,2,…,n), and n fixed positions, and the distance between the working place is D_{ij} , rating the total material flow to a minimum, namely.

$$\min \sum_{i=2}^{n} \sum_{j=2}^{n} W_{ij} \times D_{ij} \tag{1}$$

When homework unit number n is smaller, heuristic method is more convenient, it includes Construction and Improvement. Construction starts thoroughly according to the determined rules, every time addressing a work unit to a certain work place until all of the candidates are arranged with the corresponding work units garrisoned in homework. While Improvement is to determine an initial layout, and then improves decorate Settings through the mutual exchange or change workplace of some homework unit, the most common method of exchange technology.

When homework unit number n is bigger, this problem turn to be NP-complete problem, along with the increase in the number of operating unit layout, the solution space exponential expands, the combination explosion phenomena appears, the solver is very difficult. At this time, the use of genetic algorithm for solving this kind of problem is more convenient, and at the same time Plant Simulation provides general genetic algorithm to solve these problems, it is GA wizard, which can be search in the solution space extensive and get the global optimal solution.

2) The known n quantities of material handling between operation departments are W_{ij} (i=1,2,…,n, j=1,2,…,n),and work units and the area, reasonable arrangement of the

relationship between the operation departments make all the material flow rate minimal or close to the minimum [5]. It is convenient to use SLP theory under this situation, using SLP method for the equipment layout is to make analysis for the relationship between the unit works, including the mutual relationship of the logistics and the non-logistics, after a comprehensive, obtains the comprehensive relation tables of each department. Then, according to each homework unit's position displayed in the comprehensive relation tables, draw the related graphs of operation department position, combining the operating unit area with the actual locations, concluding relevant figures of the work unit area, through the correction, the feasible layout can be obtained, and finally by evaluation, scheme will be decided.

3.2   Balancing of Assembly Line

In the industrialized developed country, according to relevant data statistics, the industrial assembly production will have wasted 5% to 10% of the production time in balancing delay [6]. This is because after the differentiation of homework, the operation time of each process could not be completely equal in theory and fact , each working procedure is not imbalance.

In order to solve each the imbalance of the working procedure, the working time of each procedure must be averaged, and the assignments must be standard at the same time, in order to make assembly line flow smoothly. Assembly line balancing is the equalize of all the process, adjust the work load, in order to make a similar operation time as far as possible.   The purpose of assembly line is to make the optimized of the assembly line in time, generally speaking, it is will reduce free time of each station to a minimum. through finding, analyzing and improving the problem, making the assembly line to a new higher balance, and then in the new conditions, again through the continuous circulation of finding, analyzing and improving problems, to improve the production efficiency continuously. Plant Simulation provides general genetic algorithm which is GA wizard to solve this problems

3.3   Distribution Scheduling and the Optimization of AGV Car

After the model building, certain methods of control also need to provide to make that the simulation process really reflect the actual production. Plant Simulation provides object-oriented simulation control language of SimTalk, and it provides interpretation and debugging tools for editors, defining the control of material flow object and information flow object. Taking the SimTalk language of the control program of material flow object and information flow object as the trigger condition of the control strategy and the trigger of operation after execution can let the simulation process execute according to actual production, being the basis of the further analysis and optimization [7]. The following is the scheduling strategy of the assembly line of AGV car, it can control the distribution of material.
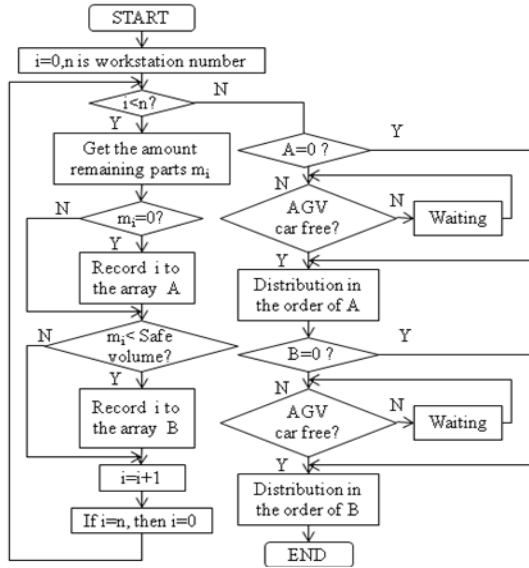
Fig. 2. Flow chart of AGV car scheduling strategy

Assembly line materials are not only large number of handling, but also has the characteristics of the strong relevance around the working procedure of handling, handling short distance, and low carrying mechanization, so the study of AGV distribution optimization is imperative. By changing the different distribution strategies, the time distribution, distribution path and the distribution priority order information, AGV distribution optimization evaluate distribution capacity, finding out the bottleneck and improve it.

4   Application Instance

This application instance is the assembly line of vacuum circuit breaker, this line has fifteen assembly workstations, one grinding-in workstation, five inspection workstations. Using the method of this paper to make this assembly line modeling and simulation optimization, the following is the specific steps.

1) Modeling assembly line by Plant Simulation.

Different products have different process, modeling the assembly line according to the process. Under the condition of meeting the requirement of the process models according to the actual situation of the assembly lines.

2) The equipment layout design of the assembly line.

Genetic algorithm is applied for equipment layout, taking the coding of work location as chromosomes, and equipment can be thought of as a chromosome genes, it turns to the most common scheduling problems in genetic algorithm, which is to code the series machine. Using genetic calculate to find the optimal solution, calculating the

minimum material flow rate by the optimal solution equipment layout, so as to realize the optimization of the assembly line, shorten the work hour.

3) Assembly line balancing.

Assembly line balancing realize optimization by genetic algorithm of Plant Simulation, the first is the problem statement and coding, expressing the coding method by assembly sequence, and the length of the chromosome is equal to the task number, expressing each task by a digital, and the deposit sequence is the order of the assembly sequence, and then work will be assigned to workstation according to the assembly sequence, making the sum of the assembly time of each workstation is not greater than the beat of production; GA Optimization is applied to create initial coding and Evaluate is applied to inspect the rationality of code; The reasonable coding is stored in GA Optimization as father generation; GA Sequence is applied to cross and generate the descendant, and the rationality of coding will be inspected by the Evaluate; The reasonable code is applied on workstation distribution and fitness evaluation by Function; GA Optimization call termination method to record the optimal several groups of results in Best Solution list.

4) Design the distribution of the scheduling and optimization of AGV car.

This instance   adopt the distribution of the scheduling with the center of workstation , and the scheduling is cater to workstation and the active distribution, predicting location and time demand according to production plan and Material list, actively delivering materials to workstation on time. Scheduling process adjusts operation of the system according to the changing logistics information, Tools and the type and quantity of assembly parts, distribution path and time distribution are decided by the assembly relations, demand and assembly product. Because the time of each workstation follows a triangle assembly distribution, it will lead to the assembly parts requirement inconsistency, needing to adjust dispatching information according to the change of assembly parts.
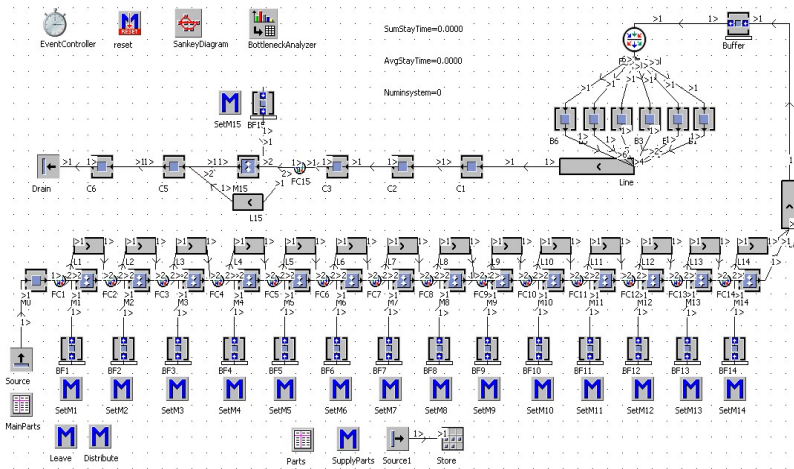
Fig. 3. The assembly line of vacuum circuit breaker

5) Handle assembly line information.

After assembly line modeling and simulation optimization, Plant Simulation also provides statistical analysis tools, experimental tools, various statistical analysis of charts can display the utilization rate of buffer, equipment, personnel dynamically, and it will help to analysis the data on the assembly line effectively

The modeling and simulation optimization of Vacuum breaker assembly line is shown in figure 3.

5   Conclusions

Research on assembly line with Plant Simulation software gets the design method of general assembly line modeling. Modeling method and genetic algorithm is applied for the modeling and simulation optimization of vacuum circuit breaker assembly line, and research on scheduling of AGV car in the assembly line concludes the general strategy of AGV car scheduling and optimization, verifying the correctness of the method through the instance in this paper. At the same time, research on optimization and improvement of the assembly line can satisfy the enterprise's requirement of improving and optimizing assembly line based on the existing assembly line, and it can greatly improve the efficiency of the enterprise.

References

1. Ciobanu, G.: A branch-and-bound algorithm to solve an assembly line balancing problem. Economic Computation and Economic Cybernetics and Research 4, 69–79 (1977)
2. Shi, Y., Deng, Y., Jiang, W.: eM_plant simulation technology tutorial. Science Press, Beijing (2008)
3. Yang, X., Wang, F., Feng, Y.: The Review of Simulation Optimization Theory and Techniques. The Computer Simulation (5), 20–23 (2000)
4. Tompkins, et al.: Facilities planning. In: Yi, J., Yuan, H., et al. (eds.) The Original Book Version 3, vol. (7). Mechanical Industry Press, Beijing (2007)
5. Zhou, J.: Production system Simulation-Plant Simulation application tutorial. Electronic Industry Press, Beijing (2011)
6. Jia, D., Han, Y., An, J.: Intermittent mobile production line analysis and organizational design. Industrial Engineering, 42–46 (July 2004)
7. Bangsow, S.: Manufacturing Simulation with Plant Simulation and SimTalk. Springer, Heidelberg (2010)

Research and Simulation on Template Surface Construction and Orthogonal Curvature Extraction[*]

Hesheng Zhang[1,2], Xiaoping Qiao[1], Ping'an Ding[1], and Xiaojin Zhu[1,**]

[1] School of Mechatronics Engineering and Automation, Shanghai University, Shanghai, 200072, P.R. China
[2] Shanghai Institute of Aerospace Electronic Technology, China Aerospace Science and Technology Corporation, Shanghai, 201109, P.R. China
mgzhuxj@shu.edu.cn

Abstract. In this paper, face to surface fitting algorithm based on orthogonal curvature is proposed, and then with the help of related theories of differential geometry, the template surface can be constructed based on arc tangent function. Then, the method of extracting curvature on equal arc meshes and the method for numerical solution of nonlinear equations are analyzed. Besides, with the help of computer software technology and computer graphics library technology, a software simulation platform is firstly built, and then the construction of arc tangent rotating surface, the extraction of orthogonal curvature and the 3D rendering of template surface and fitting surface are successfully realized. The operation effect of software simulation platform shows that the construction algorithm of template surface, the extraction algorithm of orthogonal curvatures and the method for numerical approximation can effectively construct template surface and extract orthogonal curvature accurately. These algorithms can provide basic technical support for the simulation and experimental verification of surface fitting algorithm based on orthogonal curvature.

Keywords: Orthogonal curvature, equal arc meshes, template surface, arc tangent function, surface construction.

1 Introduction

Non-visual form perception method is generally known that the discrete fiber Bragg grating sensors are inserted into the surface of the research object and real-time dynamic detection of structural deformation is conducted by utilizing detected strain information[1~3]. This method transforms strain information to curvature information at first, then makes curvature information continuous by adopting linear interpolation、bilinear interpolation and so on, and then fits the structural

---

deformation to conduct visualization of dynamic reconfiguration by utilizing continuous curvature information[4,5]. Because non-visual form perception has the features that the sensors have high accuracy of data collection and the amount of data transmission is small[6], one important application direction of this method is the platy structure form reconstruction of aerospace vehicles, such as aviation aircraft wing、 the spacecraft solar panels and so on[7,8]. These platy structures can be simplified as one end fixed cantilever structure, and they can be reconstructed by the strain information which orthogonally distribute in its surface and the given boundary conditions, and the key of its realization is curved surface fitting algorithm based on orthogonal curvature information.

Building a simulation platform and conducting a simulation experiment is a necessary method to verify the correctness of the basic theory about fitting algorithm and the operation effect of the algorithm. Building a suitable template surface and extracting its orthogonal curvature information is the key step of achieving the simulation platform. This paper makes use of the differential geometry theory to build a template surface, and extracts orthogonal curvature information from the template surface. In order to verify the fitting algorithm of space surface based on the orthogonal curvature, two conditions must be meeted when the template curved surface is constructed, one is the curvature of the two orthogonal directions are non-zero, another is one end of the curve surface must be fixed . On the one hand, the arc tangent function has asymptotic lines, and the surface are obtained by rotating central axis, it meets the boundary conditions that one end is fixed; On the other hand, any point on the surface has the non-zero curvature of the orthogonal. So arc tangent rotating surface is an ideal template surface.

By utilizing the surface-related theory, the template surface based on the arc tangent function should be constructed at first. Then orthogonal curvatures of equal arc meshes can be extracted by using the arc tangent rotating surface; the numerical treatment approach to obtain approximate solutions of nonlinear equations could be used during the process of the software implementation. The construction of the arc tangent rotating surface and the extraction of orthogonal curvature are accomplished. The running results of the simulation platform show that the construction of curved surface and the extraction of orthogonal curvature are achieved accurately by using the method described in this paper. The technical support is provided for the simulation of the surface fitting algorithm based on the orthogonal curvature.

2   Construction of Arctangent Rotating Surface

As the basic knowledge of differential geometry about curved surface, a curved surface can be considered as a continuous mapping of planar region. If (u, v) and (x, y, z) are used to denote the dot of the planar region E^2 and the spatial region E^3 respectively, there are the following relationships:

$$\begin{cases} x = f_x(u,v) \\ y = f_y(u,v) \\ z = f_z(u,v) \end{cases} \tag{1}$$

The (u,v) is usually called the curved stria coordinates of the space curved surface, the curved surface which is expressed by formula 1 is called the parametric surface, f_x、 f_y、 f_z are the mapping relationships. If the orthogonal equal arc meshes are drawn in the planar region E^2, there must be a unique equal arc meshes corresponding with certain position of the space curved surface. The normal curvature which goes along the two orthogonal direction of the mesh is corresponding to the orthogonal curvatures which the space curved surface fitting algorithm based on （that is the strain of plate-type structure orthogonal detection）. If u、 v can be expressed by the arc length of direction along u and along v respectively, normal curvature can be obtained according to the theory of differential geometry curved surface. For arc tangent rotating surface, it is harder to get the arc length parameter equation of curved surface, so the method may not be used to solve it and the analytic equation of rotating surface could be used directly to obtain numerical solution. Arc-tangent function is defined in the XOY surface, the arc tangent rotating surface is obtained by making arc-tangent function rotate with rotation axis that parallel to Y axis. As is shown in figure 1:
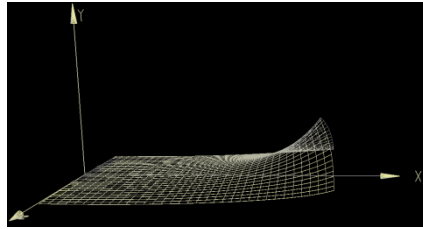
Fig. 1. Arc tangent rotating surface and equal arc meshes

Set arc tangent function as:

$$y = a(arctg(b(x-c)) + \frac{\pi}{2}). \tag{2}$$

One certain point $P(x, y, z)$ in the curved surface is known, so the corresponding point P' of the point before rotating in the XOY surface meets Eq.2:

Set P' as $(x', y, 0)$, then:

$$y = a(arctg(b(x'-c)) + \frac{\pi}{2}). \tag{3}$$

It can be deduced from Eq.3:

$$x'-c = \frac{tg\dfrac{y-\dfrac{\pi}{2}}{a}}{b}.$$

(4)

Since the point P and P' has the same distance from $P_o(c, y, 0)$ in the rotation axis of curved surface, then:

$$(x-c)^2 + z^2 = (x'-c)^2.$$

(5)

By substitution of Eq.4 to Eq.5, it can be deduced:

$$(x-c)^2 + z^2 = (\frac{tg\dfrac{y-\dfrac{\pi}{2}}{a}}{b})^2.$$

(6)

Eq.6 is the analytic expression of arc tangent rotating surface. The coordinate of the point can be easily obtained in the boundary $z = 0$ or $x = 0$; because the arc length in each intersection point of meshes with equal arc length ,which is corresponding in the direction of u and v, is known, then its coordinate can be obtained by calculating the sum of the two arc length and determined the coordinate of the point in the direction u and v, as is shown in figure 2:
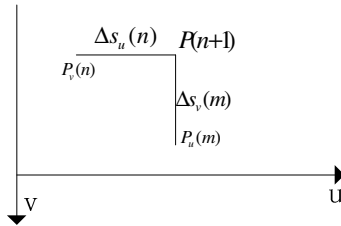
Fig. 2. Solving the coordinate of the point in arc tangent rotating surface

If micro-arc length $\Delta S_u^2(n)$, $\Delta S_v^2(m)$ is enough small, the arc length can be thought that it is equal to the corresponding length of chord, so:

$$\begin{cases} | P(n+1)P_v(n) |^2 = \Delta S_u^2(n) \\ | P(n+1)P_u(m) |^2 = \Delta S_v^2(m) \end{cases}.$$

(7)

Combine Eq.6, the coordinate of the point $P(n+1)$ can be deduced.

3   Extraction of the Orthogonal Curvatures

It is harder to solve the curvature of the points on curved surface accurately in the direction of u and v by utilizing Eq.6, in order to facilitate the implementation of computer algorithms, some methods must be used to approximate the solution .By the construction of curved surface, the coordinate of each point on arc tangent rotating surface is known, during the construction of curved surface, if the selected micro-arc length $\Delta S_u^2(n)$ and $\Delta S_v^2(m)$ is enough small, arc can be considered as circular arc. Any adjacent two points on the surface can be used as examples, as is shown in figure 3:
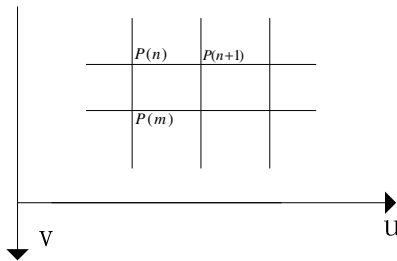
Fig. 3. The adjacent points on arc tangent rotating surface

The coordinate of the three points $p(m)$, $P(n)$, $P(n+1)$ shown in figure 3 are known, for example the micro-arc is solved in the v direction of $p(m)$ and $P(n)$. The normal N of the triangle $\Delta P(m)P(n)P(n+1)$ is solved at first, then translate N to the midpoint P'' of the chord $P(m)P(n)$. Set R as one point of the normal after the translation, as is shown in figure 4:
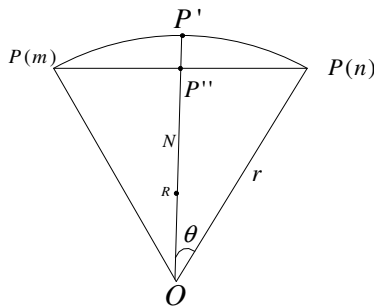
Fig. 4. The schematic diagram of micro-arc on the chord $P(m)P(n)$

As is shown in figure 4,the point P' is the point of intersection of normal N and the curved surface, the radius of micro-arc is r, it is easy to know the following relationship:

$$\begin{cases} \| P''P(n) \|= r\sin\theta \\ | P''O |= r\cos\theta \\ | P''O |= r-| P'P'' | \end{cases} \quad . \tag{8}$$

It can be deduced from Eq.8:

$$r^2 = (r-| P'P'' |)^2+| P''P(n) |^2 . \tag{9}$$

So:

$$k = \frac{1}{r} = \frac{2| P'P'' |}{| P'P'' |^2 +| P''P(n) |^2} . \tag{10}$$

In Eq.10, the coordinate of the point $P(n)$ is known, the coordinate of the point P'' is solved by using the point $P(n)$ and $P(m)$, it is the key to solve the coordinate of the point P' when the curvature needs to be obtained. $P.x$ is used to mean the x Coordinates component of the point P. Due to the point P' is intersection point of the straight line and arc tangent rotating surface, the point P' meets the following relationship:

$$\begin{cases} (P'.x - c)^2 + P'.z^2 = (\dfrac{tg\dfrac{P'.y - \dfrac{\pi}{2}}{a}}{b})^2 \\ \dfrac{P'.x - P''.x}{P'.y - P''.y} = \dfrac{P''.x - R.x}{P''.y - R.y} \\ \dfrac{P'.x - P''.x}{P'.z - P''.z} = \dfrac{P''.x - R.x}{P''.z - R.z} \end{cases} \quad . \tag{12}$$

Through solving the Eq.11, the coordinate of the point P' and the curvature of micro-arc $\Delta s_{P(m)P(n)}$ in direction of v are obtained by using Eq.10, by the same token, the curvature of micro-arc $\Delta s_{P(n)P(n+1)}$ in direction of u could be solved, so the orthogonal curvature are obtained. By the same token, the orthogonal curvature of each point on arc tangent rotating surface can be obtained.

4   Simulation Analysis and Implementation

The simulation platform for surface fitting algorithm based on orthogonal curvature must realize the construction of arctangent rotating surface and the extraction of orthogonal curvature with the help of computer software technology, and implement the rendering of template surface and fitting surface on the base of computer graphics library technology. For the scalability and maintainability, C/S architecture, composed of server software and client software, is selected in software design. The two softwares communicated with each other by TCP protocol. Server software receives human-computer interaction instructions, generates surface coordinates or orthogonal curvatures, and then sends these datas out. Client software implements the fitting of surface or directly displays surface based on the data type after receiving datas that server software sent, and subsequently compare template surface and fitting surface to analyse error. Software development is realized by VC++, where 3D rendering is implemented by OpenGL open graphics library technology. The system structure of the software is shown in Fig. 5.
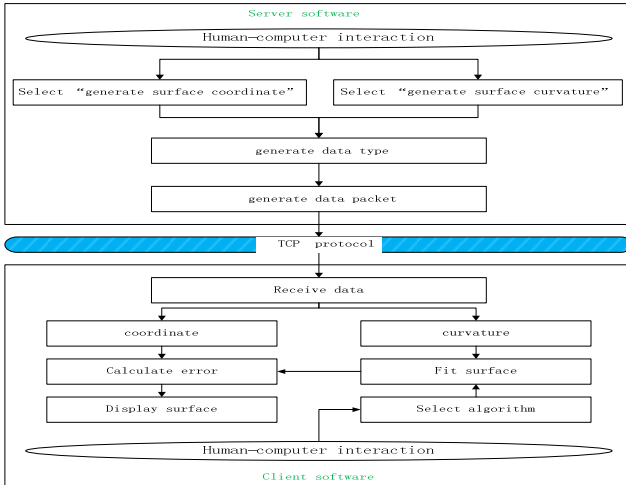
Fig. 5. The structure of simulation software

There are quite a lot of surface fitting algorithms based on orthogonal curvature and among them surface fitting algorithm based on space curve array is selected to implement the experimental verification of simulation platform in this paper. The concrete steps can be expressed as follows:

1.   To run server software and client software and implement essential initialization settings. To set the parameters of arc tangent function as follows: $a = 50, b = 0.05, c = 150$, to set the number of orthogonal curvature points on surface as $150*80$, and to set the unit arc length of meshes with equal arc length as 1 cm.

2. To connect client software with server software and make them stay in blinding state.
3. To select "generate surface coordinates" in server software and then client software begins to realize the 3D rendering of template surface after receiving the data about surface coordinates.
4. To select "surface fitting algorithm based on space curve array" in algorithm type of client software and then select "generate surface curvature" in server software. Client software receives orthogonal curvature and implements surface fitting, after which the 3D rendering of fitting surface will be enforced.
5. To implement the error analysis for template surface and fitting surface and obtain the maximum error and mean square error.

The 3D rendering of template surface and fitting surface is shown in Fig. 6.

(a) The display of template surface   (b) The display of fitting surface
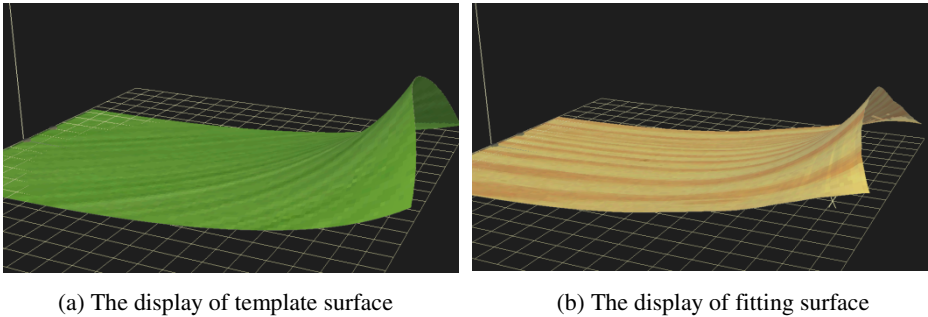
Fig. 6. The rendering results of surfaces

As Fig. 6 shows that (a) is the construction architecture of template surface while (b) is the display result of fitting surface. Error analysis for two surfaces can be enforced and then the maximum error and mean square error can be obtained, which are 5.7 cm and 1.5 cm respectively. According to the template surface and fitting surface shown in Fig. 6 and the result of error analysis, the construction method of template surface and the extraction method of orthogonal curvature introduced in this paper can efficiently build template surface and obtain orthogonal curvature. Besides, the built software simulation platform, playing the role as simulation experiment platform for surface fitting algorithm based on orthogonal curvature, can provide effect support for simulation experiment and data analysis of algorithms about surface.

5   Conclusion

Experimental environment of software simulation has important significance in simulation experiment analysis for algorithms and even the realization of subsequent project. By taking arctangent rotating surface as the research object, this paper

constructed template surface based on arc tangent function and then extracted the orthogonal curvature of equal arc meshes. In the software realization, a certain numerical approximation method was adopted and the accurate solution of nonlinear equations was obtained, thus effective realizing the construction of template surface and the extraction of orthogonal curvature. Based on the computer software technology and computer graphics library technology, software simulation platform for surface fitting algorithm based on orthogonal curvature was constructed, and then the 3D rendering of template surface and fitting surface was accurately realized, thus providing basic technical support for the simulation and experimental verification of surface fitting algorithm based on orthogonal curvature. Because of the usage of the numerical approximation method in the process of software realization, error will exist unavoidably. Thus the future work is how to construct the parameter equations of surface with curve coordinate and orthogonal curvilinear grid and consequently obtain more precise construction result.

References

1. Fan, H., Qian, J., Zhang, Y., Shen, L.-Y.: Sensor network design for flexible surface shape measurement. Optics and Precision Engineering 16(6), 1087–1092 (2008)
2. Fan, Y., Kahrizi, M.: Characterization of a FBG strain gage array embedded in composite structure. Sensors and Actuators A-Physical 121(2), 297–305 (2005)
3. Zhu, X., Lu, M., Fan, H.: Experimental research on intelligent reconstruction based structure vibration shape perception and on fiber grating network. Chinese Journal of Scientific Instrument 30(1), 65–70 (2009)
4. Di, H., Fu, Y.: Three dimensional reconstruction of curved shape based on curvature fiber optic sensor. Optics and Precision Engineering 18(5), 1092–1098 (2010)
5. Zhu, X., Xie, C., Zhang, H., Fan, H.: Visual Reconstruction of Vibration Shape for Space Solar-Sailboard Using OpenGL. Vibration, Measurement and Diagnosis 27(4), 278–283 (2007)
6. Wang, Y., Zhou, Z., Hao, Q., Ou, J.: Research on Fabrication Technique and Sensing Properties of Smart FRP-FBG Composite Laminates. Optoelectronics Laser 18(8), 900–902 (2007)
7. Du, S., Zhang, B.: Status and Developments of Intelligentized Aircraft Structures. Journal of Astronautics 28(4), 773–778 (2007)
8. Chen, Y., Xiong, K., Wang, X.: Progress and Challenges in Aeronautical Smart Structure Systems. Acta Aeronautica et Astronautica Sinica 25(1), 21–25 (2004)

BOM Ontology-Based Composite Modeling Approach for Simulation Model

Jianchun Zhang[1,2], Fengju Kang[1,2], Huaxing Wu[1,3], and Wei Huang[3]

[1] Marine College of Northwestern Polytechnical University, Xi'an 710072, China,
[2] National Key Laboratory of Underwater Information Process and Control, Xi'an, China
[3] Aeronautics & Astronautics Engineering College, Air Force Engineering University,
Xi'an 710038, China
`511604709@qq.com`

Abstract. It is an efficient way for developing models by composition of reusable components. Successful composition of models means correct in both syntactic and semantic level. Base Object Model (BOM) facilitates and improves the semantic information of simulation model, and its purpose is to improve reusability and composition. However, there is no sufficient information for BOM matching in semantic level because that BOM has no rich and clear semantic information. In this paper BOM ontology is built to enhance BOM semantic information leaving the BOM unaltered by using ontology and an iterative approach is proposed to reduce the complexity of composition. The approach mainly consists of three phases: transformation from conceptual model to event classification model; model search; model matching and composition. Finally, we demonstrate this through a simple simulation system. The result shows that this approach is effective and can simplify the composition of ontologies.

Keywords: BOM ontology, event classification model, search, match, composition.

1   Introduction

Creating simulation models by composition of predefined and reusable components is a way to reduce the costs and time associated with the simulation model development process. Composability has been defined as "the capability to select and assemble reusable simulation components in various combinations into simulation systems to meet user requirements [1] [2]". The BOM is a concept created by Simulation Interoperability Standards Organization (SISO) to provides a mean to facilitate simulation interoperability, reusability and composability. However, a challenge of BOM methodology is how to find the BOM to meet or close the simulation requirement. Though "Model Identification", as a part of BOM, contains the metadata, it lacks of semantic information for component description. Therefore, there is a need to find a way to enrich the BOM with more semantics than it presently carries. While several approaches have been offered, among others the proposal of

BOM++ extended the BOM specification with new elements[3], thus altering the integrity and structure of the original BOM.

In this paper, at first we specified BOM using Web Ontology Language (OWL) [4] to improve semantic information while keeping the BOM specification unaltered for three major reasons: ⅰ) to maintain the integrity of the original BOM, ⅱ) to preserve the compliance with the methods and concepts that are aligned with the original BOM specification, ⅲ) to explore in detail the capabilities of BOM expressed using OWL, before considering any extensions to the specification. Then we proposed an iterative search, matching and composition approach to reduce complexity of search and develop simulation system more effectively.

The paper is organized as follows: Section 2 gives the design of the BOM ontology. In Section 3, we detail the proposed approach in its implementation. In Section 4, we present a case study and conclusion and future work are discussed in the last section.

2   BOM as an Ontology

2.1   BOMs Overview

The BOM concept is based on the assumption that piece-parts of simulations and federations can be extracted and reused as modeling building blocks or components in different contexts and simulations. BOMs contain the necessary elements both to create the conceptual model and to specify and document the interface for a simulation model. BOMs are structured into four major parts [5], which are Model Identification, Conceptual Model, Model Mapping and HLA Object Model. The first part is the Model Identification, which includes descriptive information about BOMs. The main goal of this part is to help developers to find and reuse BOMs. The second part is the Conceptual Model, which contains information about how the component actually works. It is divided in four different sub models that are the pattern of interplay(POI), the state machine, the entity types and event types. The third part is the Model Mapping, which describes the mapping between entity and event elements from BOMs to their counterparts in HLA's object model. The last part of a BOM is the HLA Object Model, which defines the structure of object and interaction classes, and their associated attributes and parameters.

2.2   BOM Ontology

The current BOM standard lacks the required semantic information to avoid ambiguity. One way to overcome the lacks is to specify the meaning and the intended context explicitly for the terminology used using an ontology. Ontologies are known to promote easy understanding, shared consensus view as well as semantic interoperability. The BOM ontology describes a BOM component's structure. It is similar to a schema or meta-model that sets constraints regarding what is allowed and what is not allowed when describing BOMs. While creating the BOM ontology, it is essential to review the BOM Template Specification. This defines the format and

syntax for describing the elements of a template for representing BOMs and also specifies the semantics of the elements of a BOM and the syntax for constructing BOMs from these elements. In addition, it provides a data interchange format (DIF) for the representation of BOMs using XML. This BOM DIF should enable tools to exchange and reason about BOMs. However, in reviewing the BOM DIF, it was noted that certain outcomes desired by the implementation of the BOM DIF could be more effectively carried out if it were expressed as an ontology using OWL, rather than as pure XML.
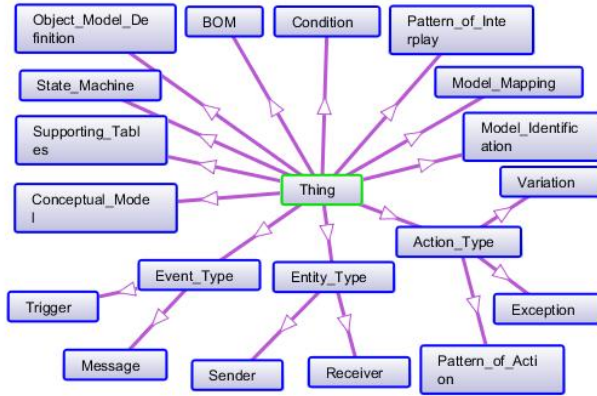
Fig. 1. BOM Ontology Classes in Protégé

Figure 1 is a graphical view of the concepts of BOM defined in Protégé Ontology Editor and shows the hierarchy of concepts that have been modeled. This graph is automatically extracted from the implemented ontology design.

3   Model Composition Approach

In this section we describe our process for BOM ontology-based model development. The basic thought is as follows: During the simulation development process, formalized simulation scenario sets needed in target simulation system are acquired via analyzing flows contained in certain conceptual models. After parsing, we get event classification models. Then we start search, matching and composition(SMC) process, which is shown at figure 2 as the follow steps:

1) The query is built based on the tag of the first event and sent to the model repository. The repository returns a set of potential candidates corresponding to the query.

2) The candidate models are matched syntactically (number of parameters and event name) and semantically (parameter data type and entity type) and the irrelevant models are filtered.

3) Then search models about the next event, the result including candidate models can be used as refined condition for the last result.

4) Iteratively execute the steps above until all events have been done. We gain a set of final models, which are processed under unitary matching and composition.
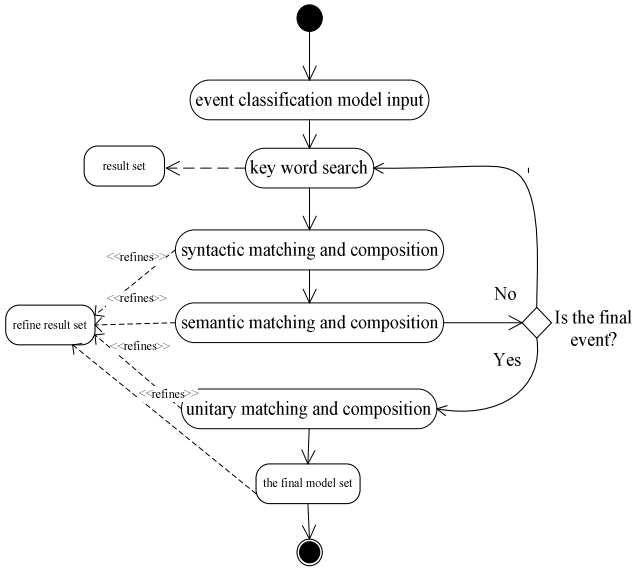
Fig. 2. Iterative search, matching and composition process

3.1   Model Transformation

This phase starts with a description of the target simulation scenario, which is decomposed into a series of atomic processes(expressed as event classification model) as input of the next phase via model parsing. Atomic process is a collection of entity names, their condition and corresponding tags.

The transformation from conceptual model to event classification model is divided into seven steps: 1) identify entities included in the simulation scenario; 2) search sub processes in the event according to model identification in the BOM Ontology; 3) based on the name of a sub process, we discover corresponding entities; 4) generate state-machine information via searching pre/post condition of each action; 5) build the tags including names of the entities and events, their classification, functional tag and so on, then atomic processes are built; 6) after ensuring the decomposition is completed, we make the sequence of these atomic processes with the relationship of events in pattern of numbering ways containing four events: sequential execution event, parallel execution event, circular execution event, branched execution event; 7) Finally, a unitary numbering is need and event classification model is generated.
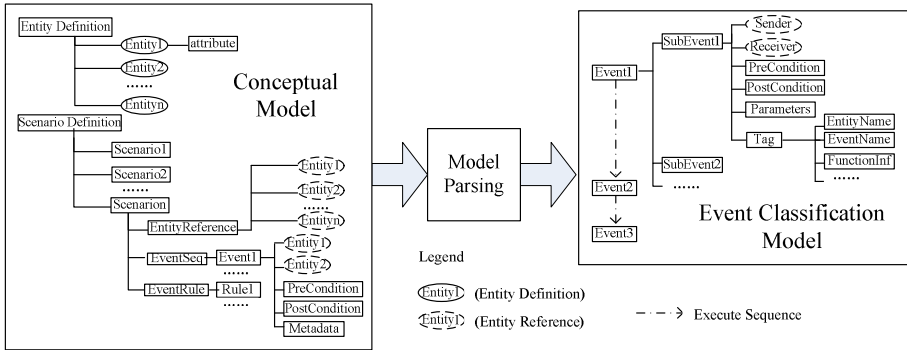
Fig. 3. Transformation from conceptual model to event classification model

3.2   Model Search

As described earlier, models in the repository are expressed as BOM ontology using OWL. Thus, search engine adopted here is based on the structure of the repository, which means that specifying a few basic types including entity name, event name, type and attribute as key words. This is mainly syntactical information and employed to filter out relevant models. This is a very naive search algorithm. However, it can be improved if some semantic filtering is also applied during the search phase.

A hierarchy of entity types can be defined in the ontology. As a result, that one can also retrieve all the models containing entities which are sub-class or super-class of the queried entity type. More models would be filtered out if we check the data type of message parameters. The accuracy of the search process is dependent on the degree of model description using ontology represented in OWL.

3.3   Model Matching and Composition

In this part, based on two candidate sets ({M1} and {M2}) got from the same event, we should find the set of model pairs ({m1, m2 | m1∈M1, m2∈M2}) which can meet the requirement of composability through the model matching between them. Because there are three layers of composition including syntactic, static semantic and dynamic semantic, we adopt rule engine to verify in the three layers considering the matching rules are different from the others' layers.

In the syntax layer, composition is verified by checking syntactic aspects of actions and messages. The "mode" of an action indicates whether the action initiates the interaction or it is invoked as a result of the interaction. The number of message parameters and event name should be the same for both "In" action and respective "Out" action.

In the static semantic layer, composition is verified by checking semantics of the events by using the semantic information. This semantic information consists of data type and unit of each parameter as well as the ontology of the event initiator and receiver entities.

In the dynamic semantic layer, composition is verified by checking the "plugin-pre" defined as the relation between post-condition of an "Out" action and the pre-condition of the corresponding "In" action. Thus the "plugin-pre" implication (PreCondsend ∧ PostCondsend → PreCondreceive) should be held between the actions of a send event and the corresponding receive event.

After the above matching, the behavior of entities in the interaction still remains unchecked: Entities expect events to be fired or to be received at right state. So the order of interactions among entities should match the current states of the involved entities.

Since post conditions are directly resulted from state machine transitions, both of state machine and dynamic semantic matching are done at the same time. In order to discover both types of mismatches, state machine of all involved entities will be executed based on given interactions in the simulation model. The state machine execution follows two major goals:

- All events, stated in the simulation scenario, can be applied to the state-machine of the involved entities such that each entity accepts or initiates the designated events and possibly changes its state.

- There is no conflict in understanding the semantic of interactions (Out action and the corresponding In action) among the entities.

4   Case Study

In order to evaluate our composition scheme, a simple scenario can be seen at figure 4 in sequence diagram format. There are five entities in the scenario: Submarine, Surfaceship, Torpedo, Jammer and Decoy. Each is supported by an BOM ontology as stated in section 2. To develop the ontologies, we build the models using Protégé, which is an open source application created by Stanford University. It is a Java-based standalone application with an extensible architecture. The tool offers capabilities for graphically-oriented ontology development using the Protégé Editor, as well as other services, such as merging, aligning, various types of visualization, and exporting/importing, among others [6].

The ontologies involved in underwater confrontation antagonistic simulation are shown at figure 5. To develop the simulation system with composition, Firstly the conceptual models need to be transformed to the event classification models. Then the search engine gets the sequence of sub events and does the search in order. Take the event of launch of a torpedo in a submarine as example, the sender and receiver are marine and torpedo. And these two kinds of ontologies are searched in repository by building key words such as the name of entities, type and attribute. Next is the matching and composition phase. In syntax layer, the models of submarine and torpedo must have the interface of "launch torpedo" and there should be four parameters(position, sound level, velocity and heading) in the interface. When in static semantic matching, three rules ought to be followed: 1. the candidate submarine models and torpedo models are in the same level; 2. the units are set consistently; 3. the data type of the parameters should be consistent or can be transformed for each other. Thereafter, the process ends up with the state machine composition.
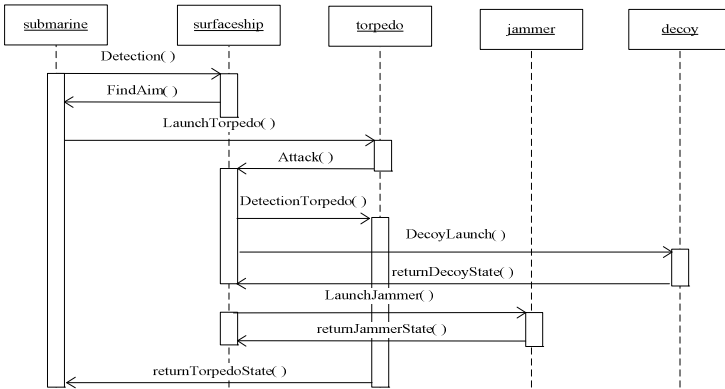
Fig. 4. Simulation Scenario in form of Sequence Diagram
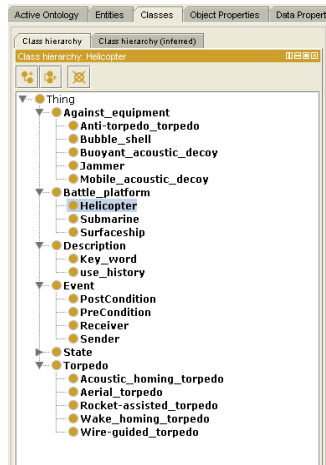
Fig. 5. Ontology description based on OWL

In the example, the pre-condition is that the target is detected successfully and post-condition is that the state of torpedo is startup. After finding the target, the submarine stays in "weapon launch" state and sends the event of launching torpedo to the torpedo. As a result, the torpedo receives the event and goes into "starting" state. At the same time, the submarine starts to guide the torpedo. Here the conditions are also checked and the "In" and "Out" actions are matched both syntactically and semantically. The models in our example pass all the matching.

5   Summary and Future Work

The goal of this study is to promote the development of simulation system based on composition. While BOM standard is an attempt to ease reusability and composition of simulation models. To the problem of not sufficient semantic information in BOMs, We present our approach for enhancement of the semantic contents of BOMs using ontology via using OWL. To support the development of composability simulation system, a novel process is presented. Firstly the event classification model is transformed from conceptual model as the input of the model search phase to reduce the complexity. Next we search models need in the model repository. Finally models based on BOM ontology are matched and composed in three layers. In this paper we have presented one case study. The preliminary results show that it can improve composition with ontology-based BOMs and SMC process.

The benefit of our proposal lies in the use of the original BOM standard to facilitate the semantic-level operations needed for composition by designing a BOM ontology to capture the semantics of the concepts without adding extensions in the BOM notation. As for the future research, we plan to consider the OWL and SWRL capabilities for further refinements in the exploitation of the elements of the entire BOM ontology to obtain richer semantic reasoning and inferences.

References

1. Weisel, E.W., Petty, M.D., Mielke, R.R.: Validity of Models and Classes of Models in Semantic Composability. In: Proceedings of the Fall 2003 Simulation Interoperability Workshop, 03F-SIW-073, Orlando, FL (2003)
2. Petty, M.D., Weisel, E.W., Mielke, R.R.: Overview of a Theory of Composability. Virginia Modeling Analysis & Simulation Center. Old Dominion University (2004)
3. Mojtahed, V., Andersson, B., Kabilan, V., Zdravkovic, J.: BOM++, a semantically enriched BOM. In: Proceedings of the Spring 2008 Simulation Interoperability Workshop, 08S-SIW-050, Orlando, FL (2008)
4. World Wide Web Consortium (W3C): OWL Web Ontology Language Overview, http://www.w3.org/TR/owl-features/
5. Simulation Interoperability Standards Organization (SISO): Guide for Base Object Model (BOM) Use and Implementation. SISO-STD-003.0-DRAFT-V0.11, SISO (2005)
6. Protégé Ontology Editor, http://protege.stanford.edu/
7. Moradi, F., Nordvaller, P., Ayani, R.: Simulation Model Composition using BOMs. In: Proceeding of the 10th IEEE International Symposium on Distributed Simulation and Real-Time Applications (2006)

8. Mahmood, I., Ayani, R., Vlassov, V., Moradi, F.: Statemachine Matching in BOM based model Composition. In: Proceeding of the 13th IEEE/ACM International Symposium on Distributed Simulation and Real-Time Applications (2009)
9. Moradi, F., Ayani, R., Mokarizadeh, S., Shahmirzadi, G., Tan, G.: A Rule-based Approach to Syntactic and Semantic Composition of BOMs. In: Proceeding of the 11th IEEE Symposium on Distributed Simulation and Real-time Applications (2007)
10. Peng, Y., Huang, J., Huang, K.: Research on Conceptual Model Semantics Based on BOM Ontology. J. System Simulation Technology 6(2), 147–152 (2010)

Research on IDEF0 and UML Combination Based Modeling of Equipment Support

Haihong Wang[1,2], Guanghong Gong[1], Jiayu Xie[2], Shu Cai[2], and Yicheng Zheng[2]

[1] School of Automation Science and Electrical Engineering, Beihang University
Beijing 100191
[2] Naval Academy of Armament Beijing 102249
1212whh@163.com

Abstract. The concept and merit of the simulation for equipment support is expatiated, and the requirement of modeling and simulation of the equipment support systems is analyzed. And then, a method for modeling of the equipment support systems is proposed which combines the IDEF0 and UML methods. With the navy equipment support as an example, a design of the simulation structure is introduced, and the model of a selected typical support activity is characterized graphically. A research method is provided for modeling and simulation of equipment support, which establish a foundation for farther simulation study of the equipment support.

Keywords: equipment support, IDEF0, UML, simulation modes system, support activity flow.

1   Introduction

With the promotion of information technology, the joint combat and equipment support is becoming an important research focus in the area of military modeling and simulation application technology, and the major means by which the army saves the acquisition cost, shorten the equipment development cycle, and improve the equipment support capability.

Modeling is the foundation of simulation, so for effective simulation of the equipment support, a complete modeling framework is needed which is the first issue to be addressed. Functional analysis of the equipment support simulation, and modeling of the inter-relationship and the time sequence of the support activities are the foundation of research of the equipment support simulation. Currently, the classical method for this problem is to employ the IDFF0 or UML technology to make some studies. For example, Zhang [1] established a model of the equipment repair and support system based on UML, Wang [2] established a model of the repair process based on IDEF0, Hong and Zhang [3] modeled an equipment information management system of the military school based on UML, and Li and Ruan [4] made research on modeling of the equipment support schema based on UML. However, all the methods mentioned above are aimed at some part of the whole equipment support

problem, not the entirety. Based on the analysis of the related works about the equipment support, we can find that the combining IDEF0 and UML methods is an effective means to achieve the simulation framework, the task decomposition and the task activity modeling. For example, by combining IDEF0 and UML, Zhang and Song [5] modeled the combat system, and Du and Ge [6] modeled the sale-store system, and so on. These works indicate that the combining method is a useful for equipment support systems modeling and simulation. So, in this article, we propose a method that could offer functional analysis of the equipment support system and graphical modeling of the equipment support activities.

2 Notation and Analysis of Equipment Support Modeling and Simulation Methods

2.1 Equipment Support

Equipment support is a series of activities by which the equipment support forces plan and make use of the support resources, to maintain a good combat effectiveness, support the equipment used in combat [7]. Some aspects included in this problem are shown as bellows:

   (1) To recover the repair of the damaged equipment.
   (2) To support the military supplies needed in fire attack.
   (3) To support the equipment supplies needed in usage, maintenance and repair of the equipment.

With the historical mission of our army is more and more widening, there might be many kinds of security threat and military actions, so that the equipment support confronts severe challenges, such as the support task is more complex, support range is wider, support time constraint is stricter, support objectives which include much technical content and support activities are more hard to achieve.

2.2 Modeling and Simulation of Equipment Support

Equipment support modeling and simulation is to describe and simulate the support activities on computer [8]. As an important work of the informationization of equipment support, equipment support modeling and simulation is a method which applies modern simulation technology to simulate the equipment support activities to support the decision in combat and training in usual, based on the abstract of the equipment support activities in real world.

Because of the complexity of the equipment support in nowadays, tremendous changes are undergoing in equipment support theories and technologies. Some characteristics of equipment support modeling and simulation are concluded as follows:

(1) The objects of equipment support are complex and uncertain, and this brings to modeling and simulation much more difficulty.
(2) The theoretical basis is not very mature, so that the theoretical model established by system analysis is not satisfactory practice.
(3) The new support modes emerging are somehow hard to described and analyzed quantitatively.

2.3   Combination of IDEF0 and UML

There are many methods can be used to model the equipment support systems. However, these methods are some partial for the equipment support simulation problems. By analysis of the characteristics of the equipment support, we proposed a method which combines IDEF0 and UML to describe the equipment support simulation task graphically.

(1) IDEF0 method

IDEF0 is a top-down analysis method, which takes simple graphics to describe the activities of a system and the inter-relationship between these activities structurally [9].

(2) UML method

UML is a method with a vast development these years, and it can be used in system understanding, design, review, configuring and maintenance, as well as information controlling. UML is most often used in software system modeling, including some other systems [10]. A UML model consists of views, graphics, model elements and common mechanisms. The view is a set of the model elements which is used to describe some specific characteristics of the system, and it is composed of one or more graphics, to abstract the system in specific perspective. Common mechanism includes entities and there relationships.

(3) Combination of IDEF0 and UML

IDEF0 don't describe the time sequence flow, but the activities. The set of inputs, outputs and control variables can be used to describe the activities. However, no time sequence and initial conditions are considered in this way, so it's not good at modeling information flow. UML can be used to describe the task execution by activity graphics, and describe the organization, resource structure and information relation, but is not good at describing the system function and inter-relationship. Therefore, neither IDEF nor UML is good enough to describe the complex and strong interacted equipment support tasks clearly. Combination of two can be developed to model the equipment support tasks, by which IDEF0 is used to model the functions and support factorial relation, and UML is used to model the support activities. The leaf-function (functions which are can't be decomposed) will be described in detail to achieve the modeling of the equipment support tasks better.

3   Establishment of the Task Oriented Equipment Support Simulation Models

Two factors should be considered in construction of the equipment support simulation models, one is identification of the support factors included in the model system according to the division of functions, and the other is modeling of support activity flow for simulation of the support tasks and process.

3.1   Identification of the Critical Factors of the Equipment Support Model System

By identification of the support object, some critical factors will be considered in equipment support simulation, such as simulation scenario, modeling and simulation of support system, modeling and simulation of support activity and evaluation. The equipment support system is composed of support objects, support entities, support resources and support environment. The support simulation scenario is a description or constraint of the support problems before the simulation run, to model the combat environment, combat task, support deployment, support resource configuration and support process. The combat environment considers the effect of geography, meteorology and electromagnetic on equipment support process. Support object includes descriptions of variety of failures and repairs of the support entities. Support organization is description of the entities and the relation between these entities which organizes the support activities. Support resource is a description of usage of the resources. Support activity is a set of the support objective oriented activities to be undertaken. Support evaluation is made by analyzing whether the support plan can satisfy the combat requirement. All aspects of activities of the factors are joined in the simulation to make the support simulation successful [11] [12].

We take an equipment support task of the navy as example, and give a design of the simulation model system of it (see Fig. 1).
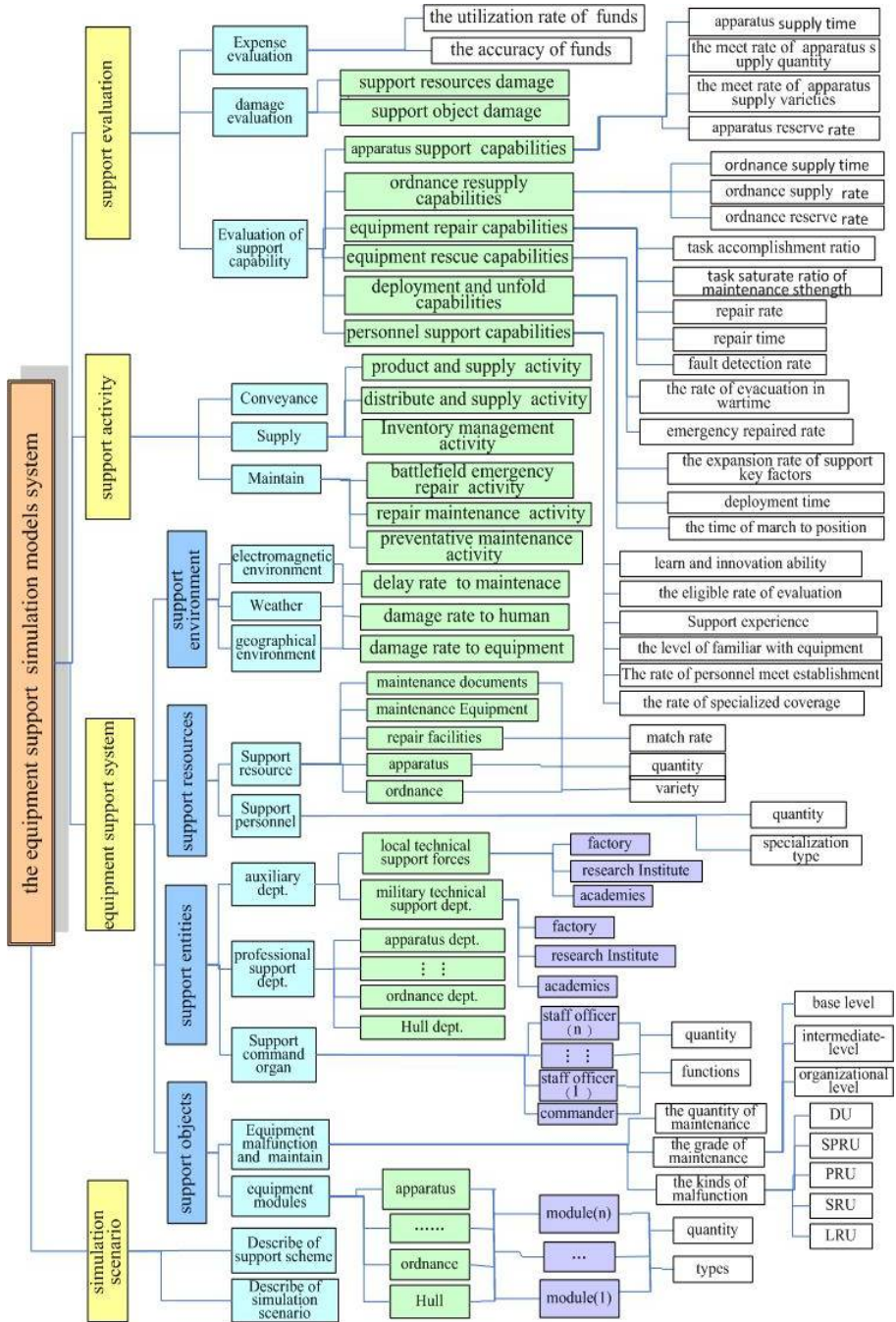
Fig. 1. Models system of Equipment support simulation

The factorial relations of equipment support modeled based on IDEF0 are described as Fig. 2.
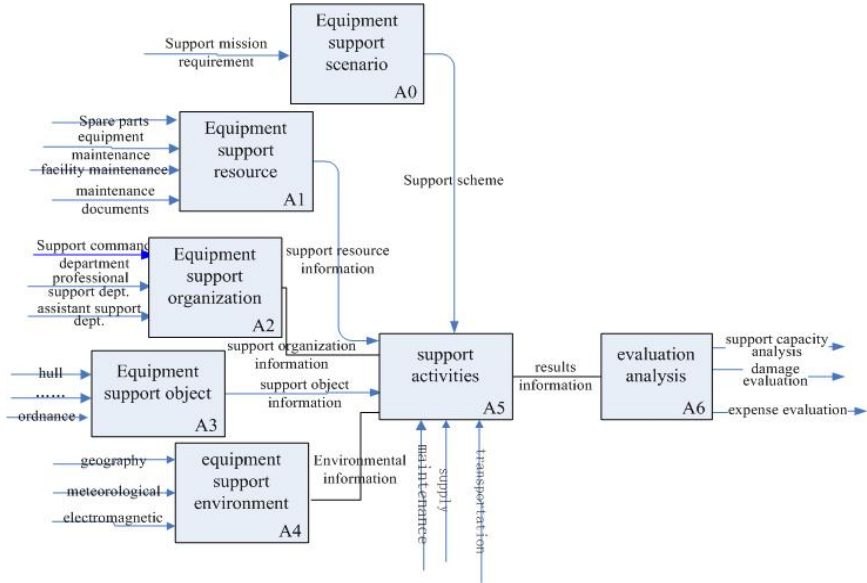
Fig. 2. Function modules of Equipment support system

Base on the top-down method of IDEF0, we decompose the support organization further, as Fig. 3 shows.
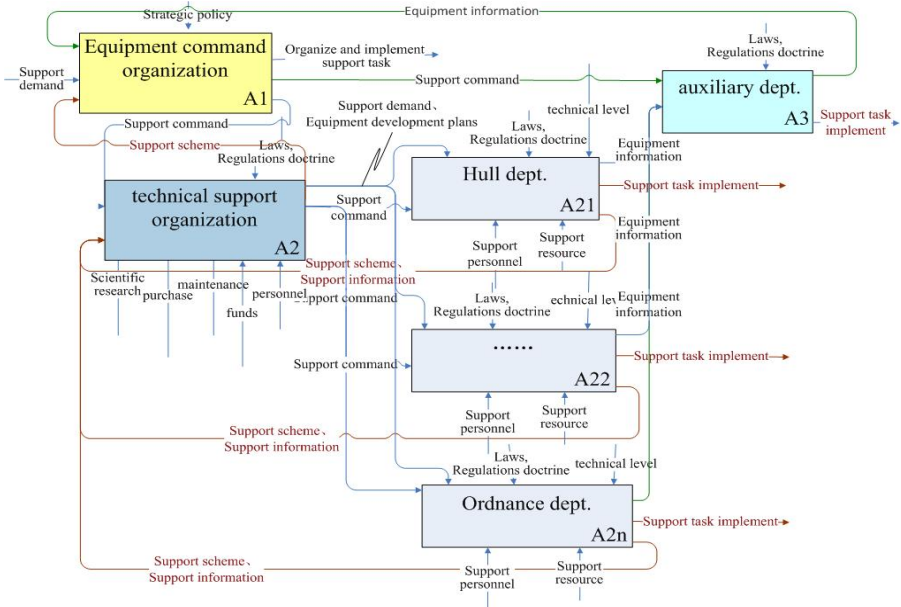
Fig. 3. Decomposition of the support organization

3.2   Graphical Description of Equipment Support Activities

Support activities referent equipment repair, supply support and transport support. By describing these activities, establish the relationship between the support factors based on analysis of support scenario, to describe the whole support related activities completely. This includes various support activities. We take a representative activity-the supply of ammunition as example, based on UML modeling method, to give the graphical description of the support process.

It's ammunition support that is to supply ammunition to the army in combat, according to the weapon ration, configuration, storage, assigning plan, material resources, consuming, combat requirement and ordnance loss in combat. In modern combat, as the high technology weapon is put to use, the ammunition consumption will be unprecedentedly huge. Because of increasing of the battlefield range, the combat forces will be maneuvering. In addition, the joined weapons and forces will be in large quantities, so the support of ammunition will be much difficult. To get victory in the war, the accurate, timely and appropriate ammunition will be needed. Therefore, the ammunition support is a critical work in equipment support.

Ammunition process is composed of some basic activities, and they are identification of the means of transport according to the amount of ammunition, requesting of transport to the support department, transferring from configuration point to ammunition depot, loading the ammunition on vehicles, transferring to support position, unloading the ammunition from vehicles, and traveling back to the original position. The commander decides the support time according to the operational plan and the ammunition consumption. After receiving the support order, the support forces supply the ammunition quickly according to the support plan and the order. This process is graphically presented in Fig. 4 and 5.
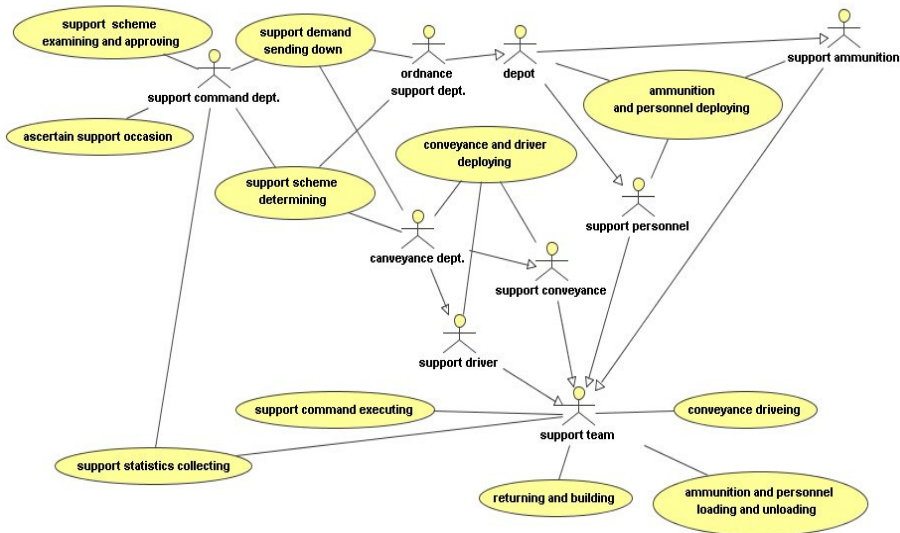
Fig. 4. Use case diagram of ammunition support
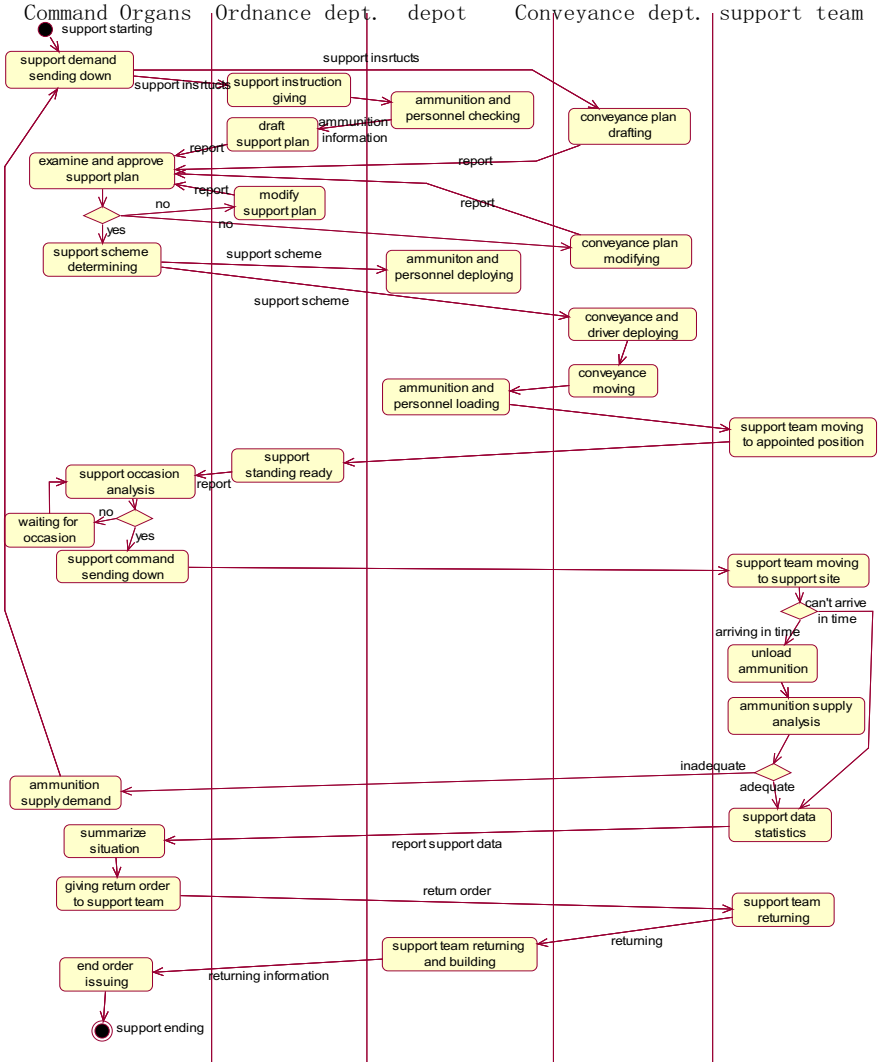
Fig.5. Activity view of ammunition supply

4   Conclusions

Modeling and simulation of equipment support is a difficult but meaningful project. We firstly proposed a modeling framework on system level, summarized the content and range of equipment support simulation, and laid a foundation for the further research on equipment support simulation, which include the simulation platform design, model development and model integration. Secondly, we analyzed the IDEF0

and UML method, and then proposed the combination method, with introducing a modeling process based on the combination method. Taking the navy support simulation for example, we proved that our method can be used to describe the equipment support tasks and the time sequence of the support activities. Generally speaking, it can be declared that the combination method has better promotion and practicability.

References

1. Zhang, Z.-H.: Function Modeling of Materiel Maintenance Support System (MMSS) Based on IDEF0 Method. Journal of Ordnance Engineering College 12, 15–19 (2007)
2. Wang, S.-Z., Gong, S.-Y., Fu, J.-Y.: Maintenance Plan Flow of Equipment Based on IDEF0 Mode. School of Information System & Management, National University of Defense Technology, pp. 14–20 (2006)
3. Hong, L., Zhang, F.-G.: Equipment Management Information System in Military Academy Based on UML. Computer Systems & Applications 9, 119–121 (2009)
4. Li, Z., Ruan, Y.: Research on Modeling of Equipment Support Plan Based on UML. Value Engineering 30, 14–16 (2011)
5. Zhang, J., Song, H.-X., Fu, M., Li, M.-M.: Operational Mission Modeling Method Combining IDEF with UML. Command Control & Simulation 6, 18–21 (2010)
6. Du, X.-M., Ge, S.: Administrate and Development of the Buying and Selling System Based on IDEF and UML. Journal of Computer Applications 12, 33–36 (2009)
7. Xu, Z.: Equipment Supportability engineering and management, pp. 2–3. National Defence Industry Press (2006)
8. Zheng, H., Song, H.: Equipment Command Learn, pp. 15–17. Academy of Equipment Command & Technology, Beijing (2002)
9. Tang, D.-D., Yu, D.-Y., Xiao, Z.-G., Wang, J.-Y.: The Function Modeling of IFEF0 and its Application on the Communication Countermeasures Decision System. Electronic Warfare 3, 46–48 (2011)
10. Zhang, M.-H., Li, A.: Integrated modeling and information mechanism of collaborative production information. In: 2008 International Conference on Information Management, Innovation Management and Industrial Engineering, pp. 67–70 (2008)
11. Bu, Y.-Y., Huang, J.-F., Ye, Y.-Q., Luo, B.-W.: IDEF&UML Based System Modeling Method and Mapping Rule Study. Control and Automation Publication Group 26, 16–18 (2010)
12. Kim, C.-H., Weston, R.H.: The complementary use of IDEF and UML modeling methods. Computers in Industry, 35–56 (2003)

Closed-Loop Subspace Identification Algorithm of EIV Model Based on Orthogonal Decomposition and PCA

Jianguo Wang, Yong Guo, and Juanjuan Wang

Department of Automation, School of Mechatronics Engineering and Automation, Shanghai Key Laboratory of Power Station Automation Technology, Shanghai University, Shanghai 200072
jgwang021@gmail.com, gy10721073@126.com

Abstract. In this paper, after analysis of the reason why some existing subspace methods may deliver a bias in the closed-loop conditions, a new SIM for closed-loop system based on orthogonal decomposition and principal component analysis is proposed by adopting the EIV model structure. Then, the underlying reason why SIMPCA-Wc delivers a bias estimate is explained from realization theory of closed-loop system based on orthogonal decomposition. At last, simulations show that the proposed method ORT_PCA-Wc used for closed-loop EIV system is effective and feasible.

Keywords: subspace identification methods (SIMs), closed-loop identification, orthogonal decomposition, principal component analysis (PCA), error in variable (EIV).

1 Introduction

In order to identify a state space model with closed-loop data, a couple of closed-loop subspace identification methods (SIMs) have been proposed in the last two decades. Closed-loop SIMs can be broadly divided into two groups: (i) closed-loop SIMs based on higher order Auto-regression model [1-4], and (ii) closed-loop SIMs based on orthogonal decomposition [5-9]. They have the advantages and disadvantages respectively. The former do not need to know reference input or the external excitation, but need to estimate the higher order ARX model while the latter is on the opposite. Moreover, the latter is better than the former for efficiency and complexity. So the paper is just considering the latter for closed-loop EIV system.

Chou and Verhaegen [10] developed an instrument variable subspace identification algorithm. They claimed that the algorithm could work for closed-loop systems if there is at least one sample time delay in the controller, which is restrictive in practice. To eliminate biased estimate of open-loop error in variable (EIV) system, Wang and Qin[5,11] proposed an instrument variable SIM via parity space and principal component analysis(SIMPCA), which can also be applied to closed-loop identification with colored input excitation. Huang [6] developed a new closed-loop subspace identification algorithm (CSOPIM) by adopting the EIV model structure of SIMPCA, and proposed a revised instrumental variable method to avoid identifying

the parity space of the feedback controller. Wang and Qin [8] present a new SIMPCA with appropriate column weighting, referred to SIMPCA-Wc, which is equivalent to SOPIM. Unfortunately, it also delivers a biased estimate for closed-loop identification through the simulation results. Katayama and Tannaka [9] derived a closed-loop subspace identification method (2ORT), which is a subspace version of the two-stage method of Van den Hof and Schrama [12]. However, most of these SIMs just consider or even don't consider output errors and assume the input variables are noise-free. Obviously, this assumption is far from the case in practice where all measurement variables contain noise. Presently, there are very few researches on SIMs for closed-loop EIV system are except Chou and Varhaegen's [10] and Wang and Joe's [5] studies, so it's an urgent desired to study SIMs for closed-loop EIV system .

In this paper, a new SIM for closed-loop system based on orthogonal decomposition and principal component analysis is proposed by adopting the EIV model structure of [5], and extended observability matrix/lower block-Toeplitz is used to extract system models. Then the underlying reason why SIMPCA-Wc delivers a bias estimate is explained from realization theory of closed-loop system based on orthogonal decomposition. At last, the simulation proves that the proposed method ORT_PCA-Wc used for closed-loop EIV system is effective and feasible.

The remaining parts of the paper are organized as follows. Section 2 gives the problem formulation and review several subspace notations adopted throughout this paper. Closed-loop SIM based on orthogonal decomposition and principal component analysis is proposed, and the reason why SIMPCA-Wc delivers a bias estimate is briefly explained in Setion3. Simulation results are presented in Section 4. The final section concludes the paper.

2   Problem Formulation and Notations

2.1   Problem formulation

Consider the identification of the closed-loop EIV system shown in Fig.1. P(z) and C(z) are, respectively, process plant and controller transfer function while r(k), u(k), y(k) and $\omega(k)$ stand for reference input, input and output free from noise, and process noise, respectively. Assume that P(z) and C(z) are discrete-time linear time-invariant (LTI) model. Thus the plant and controller are expressed as a state space model in the innovation form as Eqs. (1)-(2) and Eqs. (3)-(4), respectively.

$$x(k+1) = Ax(k) + Bu(k) + Ke(k) \tag{1}$$

$$y(k) = Cx(k) + Du(k) + e(k) \tag{2}$$

$$x_c(k+1) = A_c x_c(k) + B_c y(k) \tag{3}$$

$$u(k) = r(k) - C_c x_c(k) - D_c y(k) \tag{4}$$

where $x(k) \in \mathbb{R}^n$, $u(k) \in \mathbb{R}^m$, $y(k) \in \mathbb{R}^p$ and $e(k) \in \mathbb{R}^p$ is white noise innovation sequence with covariance K_e. (A_c, B_c, C_c, D_c) are the system matrix of controller with appropriate dimension. For EIV system as Fig.1, the relationships as following hold.

$$u^*(k) = u(k) + v(k) \tag{5}$$

$$y^*(k) = y(k) + s(k) \tag{6}$$

where $u^*(k)$, $y^*(k)$, $v(k)$ and $s(k)$ are, respectively, measured input and output, input and output measured noise. Define future data block-Hankel matrices of $v(k)$ and $s(k)$ denoted by V_f and S_f, future data block-Hankel matrices of $y^*(k)$ and $u^*(k)$ denoted by Y_f^* and U_f^*. Define short-hand notations: $M_f = \begin{pmatrix} V_f^T & S_f^T \end{pmatrix}^T$, $W_f = \begin{pmatrix} Y_f^T & U_f^T \end{pmatrix}^T$ and $W_f^* = \begin{pmatrix} Y_f^{*T} & U_f^{*T} \end{pmatrix}^T$, then using these short-hand notations, Eqs. (5) and (6) can be written as

$$W_f^* = W_f + M_f \tag{7}$$
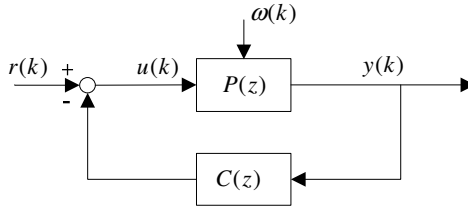
Fig. 1. Closed-loop system

To ensure the system can be identified, the following assumptions are introduced:

A1. The eigenvalues of (A-KC) are strictly inside the unit circle.

A2. The system is minimal in the sense that (A, C) is observable and (A, [B K]) is controllable.

A3. The Reference input signal $r(k)$ is uncorrelated with innovation sequence $e(k)$ and is persistently exciting of order (f+p), where f and p stand for future and past horizons, respectively, to be defined later.

A4. There is no feedback from (u(k), y(k)) to r(k).

The purpose of SIM for closed-loop system is deriving a subspace method to identify state-space models of the plant P (z) based on a finite data set {r(k), u(k), y(k), k =0,1, … , N=2i+j-2}.

2.2   Subspace Equations and Notations

Following the standard subspace notation and through the iterative substitution of Eqs. (1) and (2), one can derive the subspace matrix equation of plant and controller as

$$Y_f = \Gamma_i X_f + H_i^d U_f + H_i^s E_f \tag{8}$$

$$Y_p = \Gamma_i X_p + H_i^d U_p + H_i^s E_p \tag{9}$$

$$X_f = A^i X_p + \Delta_i^d U_p + \Delta_i^s E_p \tag{10}$$

$$U_f = R_f - \Gamma_i^c X_f^c - H_i^c Y_f \tag{11}$$

$$U_p = R_p - \Gamma_i^c X_p^c - H_i^c Y_p \tag{12}$$

where subscript p denotes past horizon and f denotes future horizon. The past and future reference input block-Hankel matrices are defined as

$$R_p \triangleq \begin{pmatrix} r(0) & r(1) & \cdots & r(j-1) \\ r(1) & r(2) & \cdots & r(j) \\ \cdots & \cdots & \cdots & \cdots \\ r(i-1) & r(i) & \cdots & r(i+j-2) \end{pmatrix} \quad R_f \triangleq \begin{pmatrix} r(i) & r(i+1) & \cdots & r(i+j-1) \\ r(i+1) & r(i+2) & \cdots & r(i+j) \\ \cdots & \cdots & \cdots & \cdots \\ r(2i-1) & r(2i) & \cdots & r(2i+j-2) \end{pmatrix}$$

Block-Hankel matrices U_p, U_f, Y_p, Y_f, E_p and E_f are defined in the same way. Moreover, we define $R = \left(R_p^T \quad R_f^T \right)^T \in \mathbb{R}^{2pi \times j}$, $U = \left(U_p^T \quad U_f^T \right)^T \in \mathbb{R}^{2mi \times j}$, $Y = \left(Y_p^T \quad Y_f^T \right)^T \in \mathbb{R}^{2pi \times j}$, $W_p = \left(Y_p^T \quad U_p^T \right)^T \in \mathbb{R}^{i(p+m) \times j}$ and $W_f = \left(Y_f^T \quad U_f^T \right)^T \in \mathbb{R}^{i(p+m) \times j}$. The extended observability matrix Γ_i and the lower triangular block-Toeplitz matrices H_i^d and H_i^s, where $\Gamma_i \in \mathbb{R}^{pi \times j}$, $H_i^d \in \mathbb{R}^{pi \times mi}$ and $H_i^s \in \mathbb{R}^{pi \times pi}$, are respectively defined as $\Gamma_i \triangleq \left(C^T \quad (CA)^T \quad \cdots \quad (CA^{i-1})^T \right)^T$

$$H_i^d = \begin{pmatrix} D & 0 & 0 & \cdots & 0 \\ CB & D & 0 & \cdots & 0 \\ CAB & CB & D & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ CA^{i-2}B & CA^{i-3}B & CA^{i-4}B & \cdots & D \end{pmatrix} \quad H_i^s = \begin{pmatrix} I & 0 & 0 & \cdots & 0 \\ CK & I & 0 & \cdots & 0 \\ CAK & CK & I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ CA^{i-2}K & CA^{i-3}K & CA^{i-4}K & \cdots & I \end{pmatrix}$$

3   SIM for Closed-Loop System

Wang and Qin [8] point out that subspace identification usually performing an oblique projection of Eq. (8) along the row space U_f onto the row space of W_p, i.e.

$$Y_f /_{U_f} W_p = \Gamma_i X_f /_{U_f} W_p + H_i^d U_f /_{U_f} W_p + H_i^s E_f /_{U_f} W_p \tag{13}$$

$U_f /_{U_f} W_p = 0$ holds true according to the property of the oblique projection. It's easy to see $E_f U_f^T / j \to 0$, $E_f W_p^T / j \to 0$ as $j \to \infty$ for open-loop system while $E_f W_p^T / j \to 0$ is no longer true for closed-loop system because future disturbance is independent of past input/output.

Theory of orthogonal decomposition for closed-loop system is briefly reviewed in subsection3.1, and then the realization steps of closed-loop SIM based on orthogonal decomposition and PCA is derived in detail in subsection3.2. Instrument variable methods based on PCA are analyzed and discussed subsection3.3.

3.1   Theory of Orthogonal Decomposition for Closed-Loop System

Picci and Katayama [13] developed stochastic realization theory with external input, and applied it to subspace identification. So system identification can be divided into two parts: deterministic component identification and stochastic component identification. Katayama and Picci [7] applied stochastic realization theory with external input to closed-loop identification.

We briefly review orthogonal decomposition of the joint input-output process (u, y) into deterministic and stochastic components [13]. Define $w(k) = \left(y^T \quad u^T \right)^T \in \mathbb{R}^{m+p}$. We introduce Hilbert spaces generated by second-order random variables of and joint input-output signals and the exogenous inputs, which are, respectively, denoted by $H = \overline{span}\{w(k) \mid k \in \mathbb{Z}\}$ and $\mathfrak{R} = \overline{span}\{r(k) \mid k \in \mathbb{Z}\}$. The orthogonal projection of w onto $\mathfrak{R}$ is $w_d(k) = \widehat{E}\{w(k) \mid \mathfrak{R}\}$ and its complement is $w_s(k) = w(k) - w_d(k)$, where $\widehat{E}\{\cdot \mid \mathfrak{R}\}$ denotes the orthogonal projection onto $\mathfrak{R}$, and where $w_d(k) = \left(y_d^T(k) \quad u_d^T(k) \right)^T$, $w_s(k) = \left(y_s^T(k) \quad u_s^T(k) \right)^T$. Deterministic components $w_d(k)$ and stochastic components $w_s(k)$ is uncorrelated since $w(k)$ and $r(k)$ is independent, namely, $E\{w_s(k)w_d(\tau)^T\} = 0$, $\forall k, \tau = 0, \pm 1, \ldots$. Katayama and Tanaka [9] developed realization theory for closed-loop based on a finite data set. Suppose that a set of finite input-output data, together with exogenous inputs, i.e. r(k), u(k), y(k) for k = 0,1, …,N. Let the subspace of finite history of second-order random variables of exogenous inputs be defined by $\mathfrak{R}_{[0,N]}$, we can conclude that the orthogonal projections of the input and output are, respectively, given by $\hat{u}_d(k) = \widehat{E}\{u(k) \mid \mathfrak{R}_{[0,N]}\}$ and $\hat{y}_d(k) = \widehat{E}\{y(k) \mid \mathfrak{R}_{[0,N]}\}$. The readers refer to [9] for the detailed theory.

3.2   Identified Algorithm Based on Orthogonal Decomposition and PCA

System matrices of plant {A, B, C, D} can be extracted from deterministic component $u_d(k)$ and $y_d(k)$ based on realization theory for closed-loop system. Different closed-loop subspace methods can be generated by using different methods for the deterministic data acquired from orthogonal decomposition. Subspace identification based on principal component analysis is a good choice because of its good efficiency. It's known that the system information is contained in the extended observability matrix Γ_i or in the state X_f from Eq. (8). Therefore it's essential to acquire the unbiased estimate of Γ_i or X_f. The realization procedure of SIM for closed-loop EIV system based on orthogonal decomposition and principal component analysis is as follows:

Step1: Compute the deterministic component using orthogonal decomposition for closed-loop data.

$$W_f^d = W_f^* / R = (W_f + M_f) / R = W_f / R \tag{14a}$$

$$W_p^d = W_p^* / R = (W_p + S_p) / R = W_p / R \tag{14b}$$

where $W_f^d = \left((Y_f^d)^T \quad (U_f^d)^T\right)^T$, $W_p^d = \left((Y_p^d)^T \quad (U_p^d)^T\right)^T$.

Step2: Perform SVD decomposition
Move the term related to U_f into the left hand side (LHS) of Eq. (8), then it yields a new equation in which input and output variables are in the same side.

$$[I \ -H_i^d]\begin{pmatrix} Y_f^d \\ U_f^d \end{pmatrix} = [I \ -H_i^d]W_f^d = \Gamma_i X_f + H_i^s E_f \tag{15}$$

Pre-multiplying Eq.(15) by $\frac{1}{j}(\Gamma_i^\perp)^T (W_p^d)^T$ to remove the state and noise variables, then yields

$$\frac{1}{j}(\Gamma_i^\perp)^T [I \ -H_i^d]\begin{pmatrix} Y_f^d \\ U_f^d \end{pmatrix}(W_p^d)^T = 0 \tag{16}$$

Perform SVD decomposition of $\frac{1}{j}\begin{pmatrix} Y_f^d \\ U_f^d \end{pmatrix}(W_p^d)^T$:

$$\frac{1}{j}\begin{pmatrix} Y_f^d \\ U_f^d \end{pmatrix}(W_p^d)^T = (U_1 \quad U_2)\begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix}$$

Then
$$\left(\left(\Gamma_i^{\perp}\right)^T [I \quad -H_i^d]\right)^T = U_2 \tag{17}$$

where $U_2 \in \mathbb{R}^{(pi+mi)\times(pi-n)}$

Step3: Extract the estimate of the extended observability matrix Γ_i and the lower triangular block-Hankel matrix H_i^d according to Eq. (17)

Let $U_2 = \begin{pmatrix} P_y \\ P_u \end{pmatrix}$, where $P_y \in \mathbb{R}^{pi\times(pi-n)}$, $P_u \in \mathbb{R}^{mi\times(pi-n)}$. Then Eq. (17) can be written as

$$\begin{pmatrix} \left(\Gamma_i^{\perp}\right) \\ -\left(H_i^d\right)^T \Gamma_i^{\perp} \end{pmatrix} = \begin{pmatrix} P_y \\ P_u \end{pmatrix} \tag{18}$$

The remaining problem is to solve Γ_i and H_i^d, and then to extract the system matrices {A, B, C, D} from Γ_i and H_i^d. The reader can refer to [5] for detailed solution procedure.

The subspace identification method is closed-loop SIMPCA-Wc method based on orthogonal decomposition and principal component analysis (abbreviated as ORT_PCA-Wc), which is slightly different from CSIMPCA [6]. CSIMPCA doesn't remove the effect of stochastic component by orthogonal decomposition. In the next subsection, we will explain the reason why some instrument methods based on PCA deliver biased estimate from the viewpoint of orthogonal decomposition.

3.3   Analysis of Instrument Variable Methods Based on PCA

Huang [6] analyzes identifiability and consistency of SIMPCA-Wc from the viewpoint of instrument variable. He introduced instrument variable W, and derived Eq. (19) from Eq. (8)

$$\frac{1}{j}(\Gamma_i^{\perp})^T [I \quad -H_i^d]\begin{pmatrix} Y_f \\ U_f \end{pmatrix} W = \frac{1}{j}(\Gamma_i^{\perp})H_i^s E_f W \tag{19}$$

Huang pointed out that a biased estimate might be delivered for closed-loop data to SIMPCA-Wc method, which adopted $W = \Pi_{W_p} = W_p^T(W_p W_p^T)^{-1}W_p$ as instrument variable. He claimed that the reason could be simply explained by Eq. (20). Huang Holds the view that $(\Gamma_i^{\perp})^T[H_i^c \quad I]$, which contains the essential information of the controller is also part of the subspace orthogonal to column space of $W_f W^T$ when the instrument variable used by SIMPCA-Wc method is adopted. Moreover, He concluded that the instrument variable W must satisfy two conditions as Eq. (21)

$$\begin{pmatrix} I & -H_i^d \\ H_i^c & I \end{pmatrix}\begin{pmatrix} Y_f \\ U_f \end{pmatrix} = \begin{pmatrix} \Gamma_i & 0 \\ 0 & -\Gamma_i^c \end{pmatrix}\begin{pmatrix} X_f \\ X_f^c \end{pmatrix} + \begin{pmatrix} H_i^s & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} E_f \\ R_f \end{pmatrix} \qquad (20)$$

One should pay attention to Eq. (20), which is slightly different from that proposed by Huang because of considering different closed-loop system. Eq. (20) can be easily got by combining Eqs. (8) and (11).

$$\frac{1}{j}E_f W^T = 0, \quad \frac{1}{j}R_f W^T \neq 0 \qquad (21)$$

For SIMPCA-Wc, even though the two conditions are satisfied when reference input r(k) is colored noise, SIMPCA-Wc would deliver a bias estimate. Huang have found the phenomenon from the simulation. Unfortunately, further explanations are not clearly given.The underlying reason why SIMPCA-Wc delivers a bias estimate is explained from realization theory of closed-loop system based on orthogonal decomposition as follows.

Orthogonal decomposition of SIMPCA-Wc can be viewed as $w^d(k) = \widehat{E}\{w(k) \mid H_k^-\}$, $w^d(k) = w(k) - \widehat{E}\{w(k) \mid H_k^-\}$. Define past and future Hilbert subspace of joint input-output, respectively, as $H_k^+ = \overline{span}\{w(\tau) \mid \tau \geq k\}$ and $H_k^- = \overline{span}\{w(\tau) \mid \tau < k\}$. $\Re^+$ and $\Re^-$ are defined conformably with H_k^+, H_k^-. Realization theory pointed out that deterministic and stochastic component can be acquired only if the present time input be included in orthogonal projection variables. So deterministic component and stochastic component must be uncorrelated for closed-loop system.

$$\widehat{E}\{w^s(k)w^d(\tau)^T\} = \widehat{E}(w(k) - \widehat{E}\{w(k) \mid H_k^-\})\widehat{E}\{w^d(\tau) \mid H_k^-\}^T$$

$$= \widehat{E}(w(k))\widehat{E}\{w^d(\tau) \mid H_k^-\}^T - \widehat{E}\{w(k) \mid H_k^-\}\widehat{E}\{w^d(\tau) \mid H_k^-\}^T \qquad (22)$$

It's easy to see $\widehat{E}(w(k)) \neq \{w(k) \mid H_k^-\}$ since the present w(k) is not only related with H_k^-, but also with past and present reference input, consequently, for Eq.(22) $\widehat{E}\{w^s(k)w^d(\tau)^T\} \neq 0$.Therefore, SIMPCA-Wc would deliver a biased estimate no matter reference input signal is white noise or colored noise.

4　　Simulations

Several SIMs for closed-loop system will be compared through numerical simulation. ORT_PCA-Wc(Proposed in this paper) will be compared with three well-known methods: CSOPIM [6], SIMPCA-Wc [8], 2ORT [9] through simulation of a SISO 5th –order system, which is adopting the EIV structure of Wang and Qin [5].

We consider a 5th-order plant, which is used in [14]. The block diagram is shown in Fig. 1, which is slightly different from Huang. Plant and controller are given in innovation form as follows.

$$A = \begin{pmatrix} 4.40 & 1 & 0 & 0 & 0 \\ -8.09 & 0 & 1 & 0 & 0 \\ 7.83 & 0 & 0 & 1 & 0 \\ -4.00 & 0 & 0 & 0 & 1 \\ 0.86 & 0 & 0 & 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0.00098 \\ 0.01299 \\ 0.01859 \\ 0.00330 \\ -0.00002 \end{pmatrix}, C = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}^T, K = \begin{pmatrix} 2.3 \\ -6.64 \\ 7.515 \\ -4.0146 \\ 0.86336 \end{pmatrix}, D = 0$$

$$A_c = \begin{pmatrix} 2.65 & -3.11 & 1.75 & -0.39 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, B_c = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, C_c = \begin{pmatrix} -0.4135 \\ 0.8629 \\ -0.7625 \\ 0.2521 \end{pmatrix}^T, D_c = 0.63$$

where e(k) is a Gaussian white noise sequence with variance; the reference signal r(k) is a Gaussian white noise sequence with variance 0.2 for CSOPIM, 2ORT and ORT_PCA-Wc while r(k) is auto-correlated external excitation for SIMPCA-Wc, which is filtered white noise with variance 1; the filter is first-order with a pole 0.9. v(k) and s(k) are Gaussian white noise for measured noise with variance 0.1. Each simulation generates 2000 data points. We generate 30 data sets by performing Monte-Carlo simulations, each time with the same reference input r(k) but with a different noise e(k). Also, we take past and future horizons p = f = i = 20, and assume that the dimension of the plant is known. One can refer to [13] for the choice of system order.
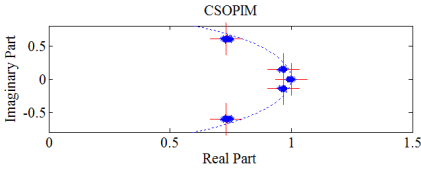
Fig. 2a. Pole plots of the estimated plant by CSOPIM
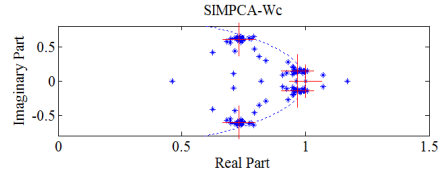
Fig. 2b. Pole plots of the estimated plant by SIMPCA-Wc
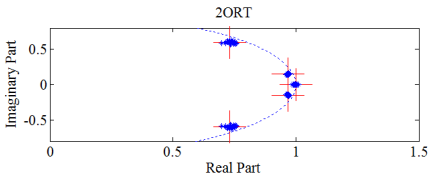
Fig. 2c. Pole plots of the estimated plant by 2ORT
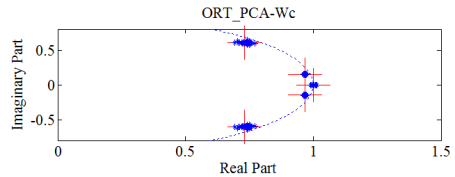
Fig. 2d. Pole plots of the estimated plant by ORT_PCA-Wc
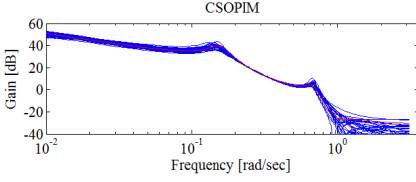
Fig. 3a. Bode plots of the estimated plant by CSOPIM
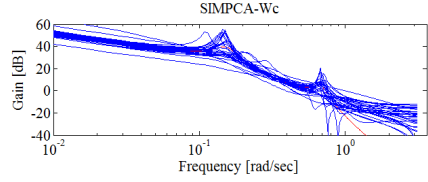
Fig. 3b. Bode plots of the estimated plant by SIMPCA -Wc
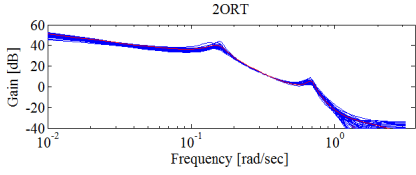
Fig. 3c. Bode plots of the estimated plant by 2ORT
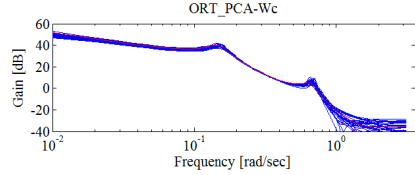
Fig. 3d. Bode plots of the estimated plant by ORT_PCA-Wc

Table 1. Comparison of the efficiency for CSOPIM, 2ORT and ORT_PCA-Wc

SIMs	Simulation Times	Time consuming
CSOPIM	30	27.55s
2ORT	30	5.88s
ORT_PCA-Wc	30	5.80s

Pole plots of the estimated plant by CSOPIM, SIMPCA-Wc, 2ORT and ORT_PCA-Wc are displayed, respectively, from Fig.2a to Fig.2d, where the true poles are denoted by +. Similarly, Bode plots of the estimated plant by CSOPIM, SIMPCA-Wc, 2ORT and ORT_PCA-Wc are displayed, respectively, from Fig.3a to Fig.3d, where the true Bode plot is covered by the estimated curves.

The results show that ORT_PCA-Wc proposed in this paper for closed-loop EIV system gives unbiased estimate, which identified accuracy is almost the same as CSOPIM and 2ORT. From Fig.2b and Fig. 3b, we observe that SIMPCA-Wc delivers a biased estimate even though it satisfies the two conditions proposed by Huang where the reference signal is colored noise. So it's not sufficient to explain the reason why SIMPCA-Wc gives a biased estimate from the point of instrument variable. Then, from the view of identified efficiency, Table 2 shows that ORT_PCA-Wc has better identified efficiency than CSOPIM obviously, and is slightly better than 2ORT in efficiency.

5   Conclusions

In this paper, by adopting the EIV model structure, a new SIM based on orthogonal decomposition and principal component analysis (abbreviated as ORT_PCA-Wc) is proposed. Through the realization theory of closed-loop system based on orthogonal decomposition, The underlying reason why SIMPCA-Wc delivers a bias estimate is explained, which is complement of Huang's theory for choice of instrument variable when the reference is auto-correlated noise. The performances of the proposed algorithm are compared with CSOPIM and 2ORT through simulations of a well-know 5th order system. The results of experiment demonstrate the feasibility and efficiency of the proposed algorithms.

References

1. Ljung, L., McKelvey, T.: Subspace identification from closed loop data. Signal Processing 52, 209–215 (1996)
2. Jansson, M.: Subspace identification and ARX modeling. In: Proceedings of the 13th IFAC SYSID Symposium (2003)
3. Qin, S.J., Ljung, L.: Closed-loop subspace identification with innovation estimation. In: Proceedings of SYSID 2003, Rotterdam (2003)
4. Chiuso, A., Picci, G.: Consistency analysis of some closed-loop subspace identification. Automatica 41(3), 377–391 (2005)
5. Wang, J., Qin, S.J.: A new subspace identification approach based on principal component analysis. Journal of Process Control 12(8), 841–845 (2002)
6. Huang, B., Ding, S.X., Qin, S.: Closed-loop subspace identification: an orthogonal Projection approach. Journal of Process Control 15, 53–66 (2005)
7. Katayama, T., Kawauchi, H., Picci, G.: Subspace identification of closed-loop systems by orthogonal projection method. Automatica 41(5), 863–872 (2005)
8. Wang, J., Qin, S.: Closed-loop subspace identification using the parity space. Automatica 42, 315–320 (2006)
9. Katayama, T., Tanaka, H.: An approach to closed-loop subspace identification by orthogonal decomposition. Automatica 43, 1623–1630 (2007)
10. Chou, C., Verhaegen, M.: Subspace algorithms for the identification of multivariable dynamic errors-in-variables models. Automatica 33(10), 1857–1869 (1997)
11. Wang, J., Qin, S.J.: Principal component analysis for errors-in-variables subspace identification. In: Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, Florida, pp. 3934–3941 (2001)
12. Van Den Hof, P., Schrama, R.: An indirect method for transfer function estimation from closed loop data. Automatica 29(6), 1523–1527 (1993)
13. Katayama, T., Picci, G.: Realization of stochastic systems with exogenous inputs and subspace identification methods. Automatica 35(10), 1635–1652 (1999)

Design of Intelligent UUV Model Based
on the Command Mechanism

Honghong Li[1,2], Fengju Kang[1,2], Hao Gu[1,2], and Huizhen Yang[1,2]

[1] School of Marine, Northwestern Polytechnical University, Xi'an 710072, China,
[2] National Key Laboratory of Underwater Information Process and Control,
Xi'an 710072, China
3h3h3h@163.com

Abstract. Compared with traditional BDI structure Agent which is controlled by individual consciousness to accomplish tasks, Multiple UUV combat receives command from carrier or submarine and takes action according to its surrounding information. Combined with research works from Agent theory and military combat command system, this paper brings up a structure which is based on Agent model and illustrates the formula language to build up the command oriented intelligent UUV model. UUV model realized the MAS UUV Fleet Combat simulation system and provide technical support to the realization of MAS underwater network centric warfare systems.

Keywords: Command, UUV, Agent organization, Model, Combat simulation.

1 Introduction

MAS-based multiple UUV combat fleet submarine aircraft carrier receives instruction to complete the assigned task. In this process, UUV individuals are supposed to be guided by command received, and take optimal scheme to complete tasks. At present, the research on the intelligent Agent body structure can be divided into mainly three types, namely deliberation type, reaction type and mixed structure type. Purely deliberative type and reaction type are not the best choice of structure body, and BDI-based hybrid structure which is subject to individual consciousness[1]; can not adapt to describe such intelligent UUV battle group. To overcome the weakness of the hybrid structure, three Germans Fischer, Muller and Pischel combine reaction, deliberative and cooperative ability to develop a mixed structure called InteRRap[2]. In the military field, [Literature 3] of intelligent command and control system of warship formation was studied, but was only limited to the original BDI structure entered combat level identity to describe the entity level characteristics, and can not fundamentally solve the command and individual consciousness conflict. [Literature 4] for the government and military fields is mandatory on mission based Agent model to undertake initial exploration.

Through the research of the military combat command system, This paper focuses on constituting the BDI structure in the Italian atlas with formation command set, in

order to build a new command based Agent model structure and to give the model description language ABL and the definition of semantic. It is based on the commands to the intelligent UUV model, to realize the MAS UUV Fleet Combat Simulation system.

2   Bring Agent Model Based on the Command

In military operations, there is a clear hierarchical relationship at all levels of their respective mission. Orders from the Superior officers must be strictly carried out by the Inferior officers. However, traditional structure of the BDI agent is not suitable for the model[5]. The BDI structure intends to describe the target of the central part which will change intentions set in real time. In the UUV combat fleet, goals of the UUV are often issued by a superior, thus, this paper emphasizes on weakening strengthening superior orders, and brings up a new BCD model that can be widely- used in military operation. On this basis of belief, commitment an the role of mental factors in this new BCD model, the establishment of intelligent model of the UUV can be realized.

As mentioned above, BCD model consists of the belief, desire, command (including social obligations), ability, commitment, and the role of the composition. Belief includes surroundings, self and group status information. The Command includes the orders received, the obligations and the intention of the agent. The ability includes individuals executable action of Agent individuals; The desire means to select the target commitments of the Agent; Roles include Agent in the group hierarchy, to determine the ability to set the composition; Commitment is based on the beliefs, goals and ability to apply the commitment rules desire; Commitment rule is that the Agent made a commitment in the form of basis; Planning is based on the desire and ability to an action sequence.
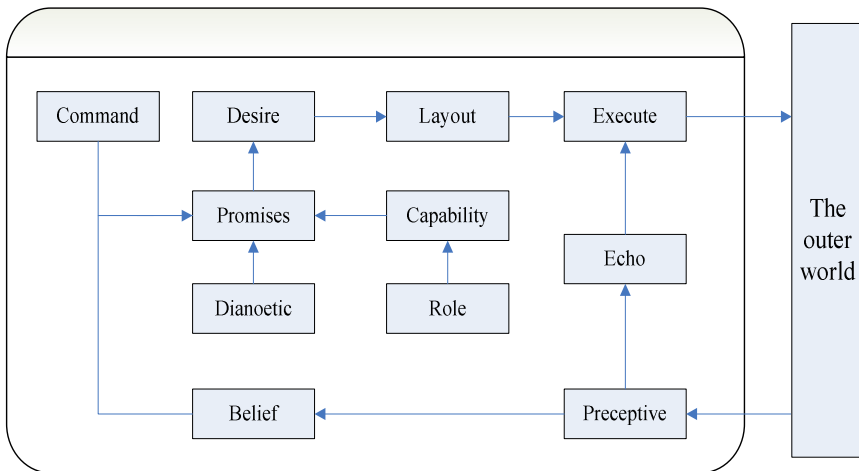
Fig. 1. Agent BCD model

3   Logic of BCD Model

3.1   ABL Syntax

In order to describe the Agent BCD Model, firstly this paper put forword Agent BCD Model Describe Language (ABL)[6].

ABL take predicate logic as the basis, including Tense Arithmetic, Action Arithmetic and BCD Arithmetic. Tense Arithmetic include: L(Last), N(Next), P(Previous), F(Future), H (Hardorwould), G(Give). The Action Arithmetic include: ACHIEVED, DONE. BCD Arithmetic include: BEL(Belief), DES(Desire), COM(Command), CAP(Capability)and so on. Then gives the following definition for ABL:

Definition 1: ABL language(For short La)constructed as follows:

1) Individual constants: C1, C2, c3... ... ;
2) Individual variables: x1, x2, x3... ... ;
3) Predicate symbol: P1, P2, P3... ... ;
4) Proposition connector: $\neg$, $\rightarrow$, $\vee$, $\wedge$, $|$, " ; " ;
5) BCD Arithmetic: BEL, DES, COM, CAP;
6) Action Arithmetic: ACHIEVED, DONE
7) Tense Arithmetic: L, N, P, F, H, G;
8) There exists: $\forall$, $\exists$;
9) Brackets: (,)

The definitions of $\neg$, $\vee$, $\wedge$ are the same to conventional definitions. $\rightarrow$ denote causal relationship, P$\rightarrow$Q denote if P then Q. "|"connect two choose action. ":" connect two sequential execution action. ACHIEVED （φ, p, φ'） denote complete P in the state of φ ,reach the goal state φ '; DONE (φ ,P) denote complete P in the state of φ . $CMD(i, j, \alpha)$ denote Agenti command Agentj to do behavior α . $CAP(i, \alpha)$ = 1 means Agenti have the ability to do α .

Definition 2: La formula is defined as follows:

1).Each atomic formula is La formula;
2).If φ and ∂ are La formula, then $\neg\varphi$, $(\partial \vee \varphi)$, $\forall$ xφ , $\exists$ xφ and X (φ) are the formula, where X is any arithmetic word. given by definition 1.
3).The non-empty string what only given through the above1) and 2) are the L a formula.

3.2   AML Semantic

The semantic model of ABL make up eight groups M= $(S,T,\phi ,\pi ,B,DE,C,N,CO)$, where S is the collect of state,T is the state sequence followed time.The elements of T are used to reflect the state change over time. ϕ are atomic formula set. $\pi \subseteq \phi \times T \times Z$ are true

value assignment function for atomic formula. CO is a constant assigned. $B \subseteq T \times Z \times T$ are accessibility relation for the beliefs; $DE \subseteq T \times Z \times T$ are the accessibility relation for desire; $N \subseteq T \times Z \times T$ are the accessibility relation for ability. $C \subseteq T \times Z \times T$ are the accessibility relation for command

The definition of true value assignment function π are:

$$B(i, j, \alpha) = \begin{cases} 1 & \text{if } C(i, j, \alpha) = 1 \\ 0 & \text{else} \end{cases}. \tag{1}$$

$$C(i, \alpha) = \begin{cases} 1 & \text{Get } B(j, i, \alpha) = 1 \text{ by axiom} \\ 0 & \text{else} \end{cases} \tag{2}$$

$$DE(i, \alpha) = \begin{cases} 1 & \text{Get } B(j, i, \alpha) = 1 \text{ by experience} \\ 0 & \text{else} \end{cases} \tag{3}$$

$$C(i, j, \alpha) = \begin{cases} 1 & \text{Agent}_i \text{ send command}_\alpha \text{ to Agent}_j \\ 0 & \text{else} \end{cases} \tag{4}$$

Set variable assignment V. The function I for the interpretation of individual words, as when a constant, variable, as in,. ABL formula for semantic interpretation[7]:

1) $(M, \pi, T, V) \models P_K(x_1, x_2, x_3 \ldots)$ iff
   $(P_K, i(x_1), i(x_2), i(x_3) \ldots)) \in \pi$

2) $(M, \pi, T, V) \models \neg \varphi \text{ iff} (M, \pi, T, V) \not\models \varphi$;

3) $(M, \pi, T, V) \models \varphi \vee \phi \text{ iff} (M, \pi, T, V) \models \varphi \text{ or} (M, \pi, T, V) \models \phi$

4) $(M, \pi, T, V) \models \varphi \wedge \phi \text{ iff} (M, \pi, T, V) \models \varphi \text{ and} (M, \pi, T, V) \models \phi$

5) $(M, \pi, T, V) \models \varphi \rightarrow \phi \text{ iff} (M, \pi, T, V) \models \neg \varphi \text{ or} (M, \pi, T, V) \models \phi$

6) $(M, \pi, T, V) \models BEL(\varphi) \text{ iff } \varphi \in B(\pi, t)$

7) $(M, \pi, T, V) \models DES(\varphi) \text{ iff } \varphi \in DE(\pi, t)$

8) $(M, \pi, T, V) \models CAP(\varphi) \text{ iff } \varphi \in N(\pi, t)$

9) $(M, \pi, T, V) \models CMD(\varphi) \text{ iff } \varphi \in C(\pi, t)$

10) $(M, \pi, T, V) \models L_i(\varphi) \text{ iff } i=0 \text{ or} (M, \pi, T, V) \models L_{i-1}(\varphi)$

11) $\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models N_i(\varphi)\,\mathrm{iff}\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models N_{i+1}(\varphi)$

12) $\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models P_i(\varphi)\,\mathrm{iff}\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models P_j(\varphi),\,j\le i$

13) $\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models F_i(\varphi)\,\mathrm{iff}\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models F_j(\varphi),\,j\ge i$

14) $\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models H_i(\varphi)\,\mathrm{iff}\,\forall j\in Z,\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models H_j(\varphi),$

15) $\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models G_i(\varphi)\,\mathrm{iff}\,\forall j\in Z,\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models G_j(\varphi),$

16) $\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models \exists x\varphi\,\mathrm{iff}\left(M,\pi,\mathrm{T},\mathrm{V}\right)\models \varphi,$

3.3   AML Promises Rules

UUV promises module generate desire according to promises rules.The initial rule make up by ABL axioms and rules.The ABL axioms generated in using UUV are added to the promises mechanism.

Promises rules:

A1   CMD（φ）∧BEL（φ）∧CAP（φ）→DES（φ）

A2   DES（φ）∧CAP（φ）→DONE（φ）

A3   CMD（φ）→BEL（φ）

A4   DES（N（φ））→N（φ）

A5   DES（φ）→BEL（DES（φ））

A6   CMD（φ）→BEL（CMD（φ））

A7   L(φ)→BEL(L(φ))

A8   CMD(φ)→F(-CMD(φ))

A9   CAP(φ)→BEL(φ)

A10 CMD(φ)→CAP(φ)

A11 CAP(φ)→BEL(CAP(φ))

A12 CMD(φ)→CAP(CMD(φ))

A13 DES(φ)→CAP(DES(φ))

A14 G($\varphi\to\phi$)→(G(φ)→G(ϕ))

A15 H($\varphi\to\phi$)→(H(φ)→H(ϕ))

A16 G(φ)→φ

A17 H(φ)→φ

A18 L(φ)→P(φ)

A19 N(φ)→F(φ)

A20 φ→H(F(φ))

A21 φ→G(P(φ))

A22 P(φ)→H(F(φ) $\vee$ φ $\vee$P(φ))

A23 F(φ)→G(P(φ) $\vee$ φ $\vee$F(φ))

A24 P(φ)→G(P(φ))

A25 F(φ)→F(F(φ))

A26 F(φ)→H(F(φ))

A27 $\neg$G(φ $\wedge$ $\neg$ φ))

A28 $X(\forall x\varphi) \rightarrow \forall xX(\varphi)$

A29 $\forall xX(\varphi) \rightarrow X(\forall x\varphi)$

A30 $\exists xX(\varphi) \rightarrow X(\exists x\varphi)$

4 Design of BCD UUV Model

4.1 BCD Model Composition

The main composition of BDC UUV are as follows:

1).Communication module It is responsible for receiving command information, state information. receiving really interactive information from other Agent sended, and feedbacking state information.

2) Set of Command Stored orders came from the superior Agent and the duties in groups. Also with the ability to analysis orders and real-time adjustment target.
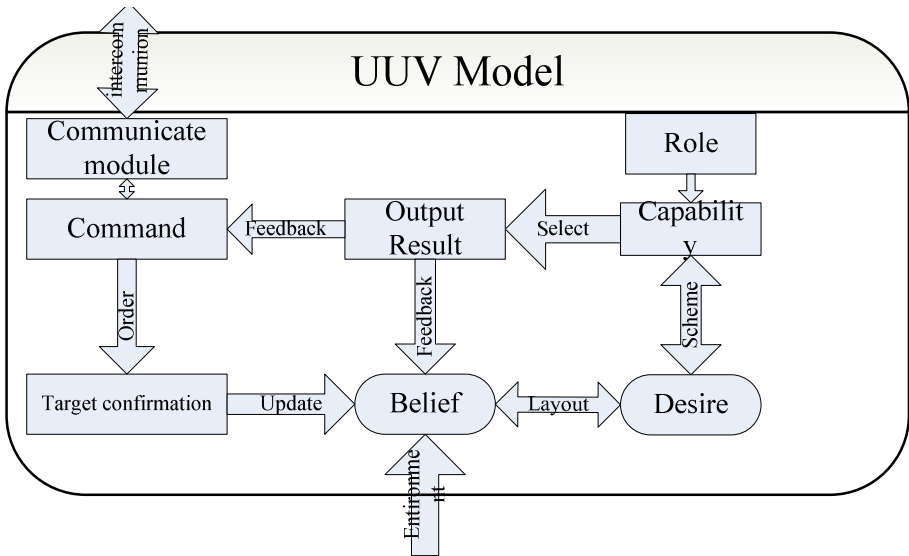
Fig. 2. UUV model composition

3).Target confirmation module   According to the target given by commands set to determine the exact targets that can be implemented.

4).Set of Beliefs   Storing environmental information,and Agent. And According to these information reasoned out the solutions for fulfilling goals

5).Set of Desire Stored individual interest desire rules, accept scheme sent by the belief, according to individual desires to choose the best step. And real-time feedback to the set of beliefs.

6).Capacity sets   Stored individual capacity determined by role. Receiving the scheme from desire, according to the ability to determine the step for implementation.

7).Results output implement the action sended by the set of capacity.And real-time feedback the state to the command set and the set of beliefs.

4.2   UUV Model

The main UUV model can be described by the following elements:

1).a set of beliefs about the world;

2).UUV want to achieve a set of goals

3).a rule base, describes how to complete the command and belief change;

4).a command base, describing UUV received command and individual consciousness.

Then give a simple interpreter, as shown below:

```
BCD-Interpreter
Initialize-state();
do
options:=option-generator(event-queue,B,C,N);
selected-option:=deliberate(options,B,N,C);
update-desires(selected-options,D);
execute(D);
get-external-events();
drop-successful-attitudes(B,C,N);
drop-impossible-attitudes(B,G,N);
until quit
```

4.3   UUV Run

In order to fully reflect the characteristics of higher orders as the center UUV run mainly by the superior command driven, rather than by environmental,UUV runner is as follows:

1) Initial state;
2) the perception module real-time apperceive environmental information and updated the set of beliefs. For some emergency using reflection mechanism direct action.
3) accept the higher command, to build target;
4) Give the desire according to the environment information, ability to set ,target and commitment rules;
5) if the desire to set is not empty, planning implementation plan,according to capability set.
6) behavioral assessment, update commitment rules and ability base.to3).

5   Simulation

This paper adopts Visual C + +2003.NET UUV BCD model to the formation of the BCD model-based multi-UUV fleet. The UUV formation control of a distributed simulation platform developed in the HLA platform. System architecture is as follows: The system consists of the nodes of the two-dimensional momentum, track mapping, the simulation clock and 9 UUV nodes. HLA / RTI communicate between the nodes. Two-dimensional momentum node shows the formation of the operational status of the overall track graph nodes in real time and draws the trajectory of each UUV. Clock node controls the simulation time forward, and the UUVs node on the respective UUV exercise settlement control.
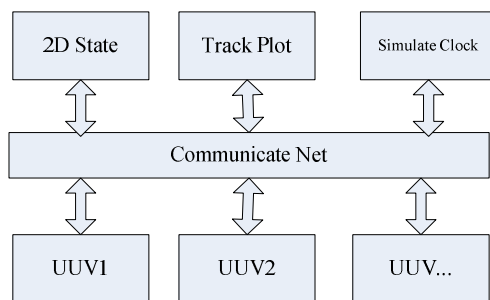
Fig. 3. Simulation system structure

In this design scenario, nine UUVs are randomly distributed in the system after the start of the simulation artificial to the UUV fleet orders in formation through a narrow Strait to reach the designated area. The firstly- joined UUV take their respective roles once received commands , then UUV consultation by the system to determine the superior UUVs, decision-making based on superior UUV formation To detect the narrow Strait of time by the superior UUV issued a directive to transform formation through the Strait. Recover through the Strait of chevron formation navigation to the target area. The diagram is shown below:

Fig. 4. The interface of UUV
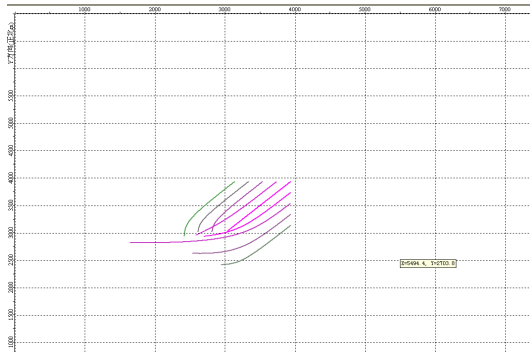
Fig. 5. The route of UUV
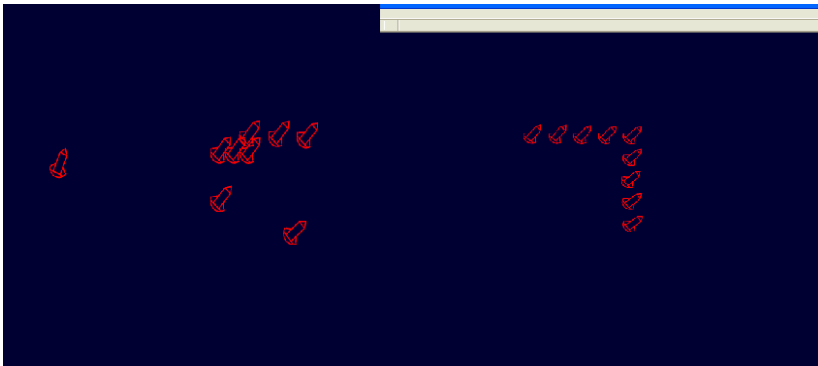
Fig. 6. The initial position and chevron formation of UUV

From the simulation results, the UUVs correctly accept the orders of superiors, and take it as the core tasks to take appropriate action depending on the surrounding

environment information program to complete the task. It is showed that the proposed the UUV new model is feasible and effective.

6 Summarize

Using the instructions of the commands in the combat system to form a set of commands, this paper tries not only to constitute the BDI structure set with a new command-based Agent Model and organizational structure, but also gives the model description language definition and semantic interpretation of the ABL, and build intelligent UUV model based on the command mechanism. VC + + fleet of intelligent UUV control simulation system results show that use of the Model UUV operational fleet command and control, to complete the orders of superiors efficient to Provide technical support for building the underwater network-centric warfare system based on MAS.

References

1. Yang, S.: Study on Rational Model Based on BDI Agent. Computer & Digital Engineering 38(7), 77–80 (2010)
2. Jennings, N.R.: On Agent based Software Engineering. Artificial Intelligence 117(2), 277–296 (2000)
3. Zhao, X.-Z., Wang, C.: Research on Modeling of Formation Cooperative Combat Simulation System Based on Multi-agent System. Journal of System Simulation 21(8), 2381–2385 (2009)
4. Li, X.-G., Li, Z.-X.: Agent Organizational Model Based on Mission. Computer Engineering 34(23), 187–195 (2008)
5. Qiu, H.-P., Xiao, D.-H., Lian, X.-L., Wang, L.-L.: A New Architecture Based on Agent. Computer Technology and Development 20(1), 50–57 (2010)
6. Wang, Y., Gao, Y.-L., Wang, J.: A Formal Specification for Agent Structure. J. of Wuhan Uni. of. Sci. & Tech. 29(6), 636–639 (2006)
7. Cao, J.-H., Zhang, H.-M., Xiong, G.-L.: Formalization Description Method of Agent Behaviors for Multi-Agent Simulation. Journal of System Simulation 16(11), 2398–2400 (2004)

Analysis of Vacuum Casting Pressure Time and Its Influence on Casting Quality

Nannan Xu, Zhuangya Zhang, Haiguang Zhang, Tian Lv,
Yuanyuan Liu, and Qingxi Hu

[1] College of Mechanical Engineering and Automation
[2] Shanghai Key Laboratory of Mechanical Automation &Robotics
Shanghai University, Shanghai 200444
huqingxi@shu.edu.cn

Abstract. Differential pressure vacuum casting process exist some characteristics, such as nonlinear, time-varying, lag and so on, which is a dynamic process with parameter changing. In the casting process, in order to make parameters meet requirements, the compression model of casting system must be analyzed correspondingly. Pressure time formula of compression process is derived by means of vacuum technology theory, and V450N-VD Vacuum Casting Machine designed by Shanghai University is used to do experiments and its results show that the cross-sectional area of inlet valve and the initial pressure of vacuum chamber are the important parameters of affecting pressure time during compression process. With the motorcycle headlight shell as the experimental mold during the vacuum casting process, experiments show that reasonable pressure time is beneficial to improving the casting quality during the compression process.

Keywords: differential pressure, vacuum casting, process parameter, pressure time, filling velocity.

1   Introduction

In recent years, with the development of new product technology, vacuum casting technology with advantages of short cycle, high efficiency and good adaptability is widely used in aerospace industry, automobile design, electronic instruments and household electrical appliances, building design, decoration design, mold design and manufacturing and other fields. At the same time, the requirements of customers about vacuum casting quality are increasingly high. If enterprises are willing to seek survival and development during the increasingly fierce market competition, the product quality must be ensured under the premise of rapid response to market demands [1, 2].Based on V450N-VD Vacuum Casting Machine developed by Shanghai University, with the help of theoretical analysis and experimental verification, put forward the important parameter having influence on pressure time during the compression process of differential vacuum casting, and analyze how pressure time has influence on casting quality by experiment.

2 Problems Existed during Vacuum Casting Process

Vacuum casting technology is a method used in vacuum environment to make the mixed material vacuumed, mixed, bubble-eliminated, then pour the mixed material into the mold cavity, and form a mold eventually. The traditional vacuum casting process depends only on the gravity of material to cast. In the casting process of the parts especially with thin wall, complex and large pieces and other characteristics, because of not regulating and controlling the casting pressure, so the filling pressure and the filling speed are too low only relying on gravity casting, resulting in material-flow poor and easily giving rise to phenomenon such as insufficient injection or shrink marks [3]. As shown in Fig.1 (a) (b).

(a) Insufficient filling (b) Shrink marks (c) Spilling (d) Air pocket
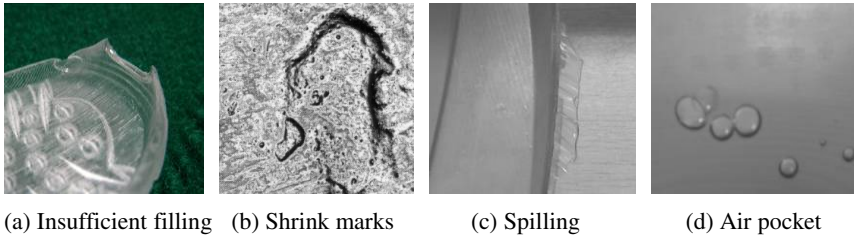
Fig. 1. Common quality defect of casting

Differential pressure vacuum casting technology is developed base on the traditional vacuum pouring process. In the casting process, make mixing chamber and pouring chamber form a certain differential pressure, and then the mixed material fast fills into the mold cavity fully under the common action of both weight and pressure-difference. To some extent, differential pressure vacuum casting avoids the defect of parts with thin wall, complex and large pieces in the casting process caused by the insufficient filling, making the casting with high density [4, 5]. However, differential pressure vacuum casting is a dynamic process with the parameters changing constantly. During the compression process, the air is so difficult to control that the differential pressure between mixing chamber and casting chamber is not constant. If differential pressure is too small, casting easily gives rise to defects such as insufficient injection or shrink-marks [1], as shown in Fig.1 (a) (b). If differential pressure is too great, casting easily gives rise to defects such as spilling and air pocket, as shown in Fig.1 (c) (d).

3 Theoretical Analysis of Pressure Time

Compression process is a gas process. That is after the separation between mixing chamber and casting chamber because of casting material pouring into funnel, and from the time when gas fills into mixing chamber to the time when a specific differential pressure is formed between mixing chamber and casting chamber. In the process of differential pressure vacuum casting, the speed of liquid dropping down funnel and the speed of the mixed material filling into mold cavity are determined by the speed of gas

pressure without considering the action of gravity. The faster the pressurized gas is, the faster the speed of filling is [6, 7]. In the process of differential pressure vacuum casting, the pressure speed should selected according to casting system design, structure cavity, casting temperature, mold wall-thickness, casting cooling velocity, material composition and other specific conditions.

For the same kind of mold, the corresponding cavity structure and casting material are known, with the environmental and anthropogenic factors under the constant condition, the final differential pressure between the mixing chamber and the pouring chamber can be regarded as a constant value which ensures the casting having comprehensive performance. Then the longer the pressure time is, the smaller the corresponding pressure speed and the filling speed are; conversely, the pressure speed and the filling speed are greater. During the casting and filling process, reasonably controlling the pressure time can control the pressure speed and the filling speed. Therefore, the analysis of parameters affecting pressure time has great significance in differential pressure casting process.

Under the condition of atmospheric pressure, when air through the valve hole into the vacuum container, no matter how the pressure inside the vacuum vessel changes, the state of flow can only be viscous flow. In the state of viscous flow, the flow of the gas through hole is:

$$Q = AP_\alpha \left(\frac{P}{P_\alpha}\right)^{\frac{1}{K}} \sqrt{\frac{2K}{K-1} \cdot \frac{RT}{M} \left[1 - \left(\frac{P}{P_\alpha}\right)^{\frac{K-1}{K}}\right]} \tag{1}$$

The flow formula under the condition of vacuum is:

$$Q = \frac{PV}{t} \tag{2}$$

As to the compression process, as known from Formula (1), the gas flow through hole is related with the pressure of vacuum container. But on one point, the flow can be considered as a constant value, so the integral method can be used to calculate value about the kind of problem. Formula (2) can be expressed as a differential form:

$$dt = \frac{VdP}{Q} \tag{3}$$

Both sides are integrated to get the formula (4) about air time during compression process, which is as follows:

$$t = \frac{V}{A}\sqrt{\frac{2k}{k-1} \cdot \frac{M}{RT}} \sqrt{1 - \left(\frac{P_1}{P_\alpha}\right)^{\frac{k-1}{k}}} - \sqrt{1 - \left(\frac{P}{P_\alpha}\right)^{\frac{k-1}{k}}} \tag{4}$$

In usual case, adiabatic exponent K, gas constant R, gas molar mass M, atmospheric pressure Pa and gas temperature T in formula (4) are constant value. Therefore, the filling time t of the vacuum container is related with the vacuum container volume V,

the valve cross-sectional area A, the initial pressure P1 of vacuum container and the pressure P after gas filling into the vacuum.

4   Influence on Casting Quality Caused by Pressure Time

4.1   Experimental Device

The experimental device is V450N-VD Vacuum Casting Machine designed by Shanghai University. The simplified model is as shown in Fig.2. The experimental parameters are as follows.

Mixing chamber volume: 454 x 452 x 387(mm3); Pouring chamber volume:454 x452x307(mm3); Maximum diameter of regulated valve:15mm; Caliber of balance-valve: 20mm.
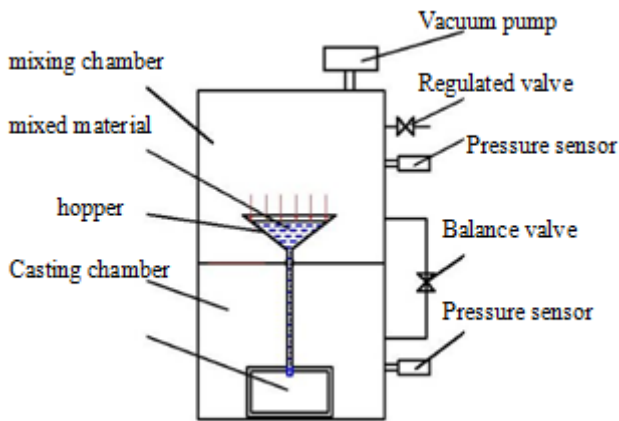
Fig. 2. Simplified Model of V450N-VD Vacuum Casting Machine

4.2   Pressure Time Records during Compression Process

Compression process is a dynamic process, the pressure of vacuum chamber is recorded by high precision digital pressure sensor. Table 1 gives the value which describes the required pressure time of forming 0.01MPa differential pressure under the condition of different initial pressure of vacuum chamber when the valve open-degree is at 45. The pressure time directly reflects the pressure speed during compression process.

In Table.1, when the initial pressure of vacuum chamber is -0.09MPa, the pressure time of forming the specified differential pressure is 0.9s; the initial pressure is -0.07MPa, the time is 1.2s; the initial pressure is -0.04MPa, the time is 1.4s. Because experiment exists errors, and the difference among the pressure time of forming 0.01MPa differential pressure under the condition of different initial pressure is very

little, so Table.1 is not able to adequately describe the relation between the initial pressure of vacuum chamber and the pressure time.

Regulate the valve open-degree and repeat experiment when other conditions and experimental parameters remain unchanged. Table2 gives the value which describes the required pressure time of forming 0.01MPa differential pressure under the condition of different initial pressure of vacuum chamber when the valve open-degree is at 9. In Table2, when the initial pressure of vacuum chamber is -0.09MPa, the pressure time of forming the specified differential pressure is 30.6s; the initial pressure is -0.07MPa, the time is 38.1s; the initial pressure is -0.05MPa, the time is 52.3s. Ignore the errors of factors such as equipment, human and external environment during the experiment process, as known from Table2: when the valve open-degree is constant, the smaller the initial pressure of vacuum chamber is, the shorter the pressure time of forming 0.01MPa differential pressure is.

Table 1. Differential Pressure Time at Open-degree 45

Vacuum chamber Initial pressure （MPa）	Mixing chamber Pressure after compression (MPa)	Casting chamber Pressure after compression (MPa)	Differential pressure Time (s)
-0.09	-0.08	-0.09	0.9
-0.08	-0.07	-0.08	1.1
-0.07	-0.06	-0.07	1.2
-0.06	-0.05	-0.06	1.3
-0.05	-0.04	-0.05	1.4
-0.04	-0.03	-0.04	1.4
-0.03	-0.02	-0.03	1.5

Table 2. Differential Pressure Time at Open-degree 9

Vacuum chamber Initial pressure （MPa）	Mixing chamber Pressure after compression (MPa)	Casting chamber Pressure after compression (MPa)	Differential pressure Time (s)
-0.09	-0.08	-0.09	30.6
-0.08	-0.07	-0.08	40.9
-0.07	-0.06	-0.07	38.1
-0.06	-0.05	-0.06	49.6
-0.05	-0.04	-0.05	52.3
-0.04	-0.03	-0.04	58.5
-0.03	-0.02	-0.03	81.8

Calculate theoretically the pressure time of forming 0.01MPa differential pressure with the valve open-degree 45 and 9 by formula (4), make the comparison with the experimental data in Table.1, the error range meets the experimental requirements. Use the same method and sequentially record the pressure time when the valve open-degree is 18, 27 and 36. In order to effectively contrast the experimental data and visualize graph, fit the time of forming 0.01MPa differential pressure under the condition of different initial pressure when valve open-degree is 18, 27, 36 and 45. Fitted curve is as shown in Fig.3. As known from Fig.3: when the initial pressure of vacuum chamber is constant, the valve open-degree directly affects the pressure time. The smaller the open-degree is, the longer the pressure time is. In addition, when the valve open-degree is at 27, 36 and 45, even if the initial pressure of vacuum chamber before differential pressure is different, the difference among pressure time is very little. When the valve open-degree is at 19, the initial pressure of vacuum chamber has a remarkable influence on the pressure time.
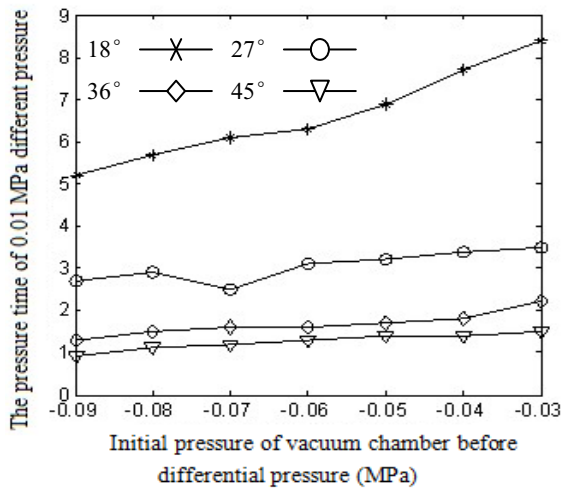
Fig. 3. Pressure Time Curve with Different Valve Open-degree

4.3   Influence on Casting Quality Caused by Pressure Time

As seen from Fig.3, when the valve open-degree is 18, the initial pressure of vacuum chamber has a significant influence on pressure time, and has more efficient compared with open-degree 9(as shown in Table.2). In order to analyze the influence on casting quality by pressure time, take motorcycle headlight as an example and do experiments about the vacuum casting process with valve open-degree 18.

(1) Experimental Material
The required materials are curing agent and resin. According to the regulated ratio between mold weight and casting materials (Based on the following equation: That

resin weight pluses curing weight is equal to that mold weight pluses casting allowance, and determine allowance according to the mold size, generally 60-80 grams), weigh quantitative required curing agent and resin material, pour them into A, B cup, then put A, B cup filled with material and sealed mold into oven to preheat.

(2) Experimental Parameters

The initial pressure of mixing chamber and casting chamber before the compression process:

$$P_{mc} = P_{cc} = P_i = \text{-0.06MPa}$$

The differential pressure between mixing chamber and casting chamber after the compression process:

$$0.005MPa \leq \Delta P \leq 0.01MPa$$

(3) Experimental analysis

V450N-VD Vacuum Casting Machine is as shown in Fig.4. Do experiments according to the steps and requirements of casting process, the casting corresponding to the differential pressure time with valve open-degree 18 is as shown in Fig.5.

As seen from Fig.5 (a), when pressure time is 5s, casting material can not fill the whole mold cavity, which makes casting appear incomplete defect. Combine the experimental process and the casting defect to analyze, that the shell of motorcycle headlamp is much thinner makes the flow resistance of casting material in filling process much bigger than that of thicken shell molds. The differential pressure corresponding to pressure time 5s is too small to overcome the material flow resistance by mold-wall during the filling process, which leads to the casting material liquidity poor and makes the casting present the incomplete defects. Compared with Fig.5 (a), the differential pressure corresponding to pressure time 10s is sufficient to overcome the material flow resistance by mold-wall during the filling process. The mold cavity is filled completely and the casting quality is improved greatly, as shown in Fig.5 (b).There is a little of air bubble on the casting surface in Fig. 5(b), the reason may be that the pressure time is so long that gas is involved into the casting material during the filling process. The above results indicate that reasonable selection of pressure time is beneficial to improve the casting quality during the compression process.

(a) Pressure time5s   (b) Pressure time：10s
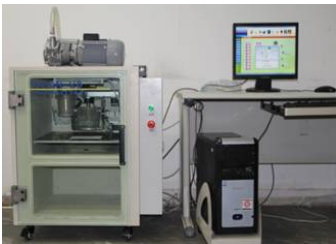
Fig.4. Experimental Prototype   **Fig.5.** Casting corresponding to Differential pressure Time with Valve Open-degree 18
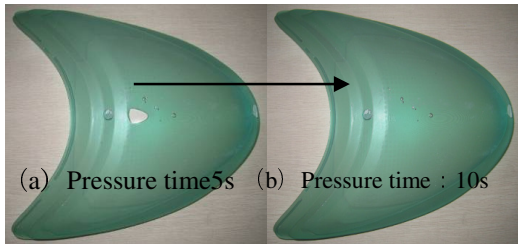
5   Conclusion

During the vacuum casting compression process, it is necessary to select a reasonable pressure time according to mold wall-thickness, cavity structure, casting system design, casting temperature, casting cooling speed, material composition and other specific circumstances. By means of vacuum technology theory and experiment, the conclusions are as follows:

(1) With the initial pressure of vacuum chamber constant, the smaller the valve open-degree is, the longer the pressure time of forming a specific differential pressure is; conversely, the shorter the pressure time is. With the valve open-degree constant, the smaller the initial pressure of vacuum chamber is, the shorter the pressure time is; conversely, the longer the pressure time is.

(2) As to the vacuum casting process of some molds similar to a motorcycle headlight shell with characteristics of thin mold-wall, large and complex cavity structure, it is common that the casting material can not fill into the whole mold cavity completely because of the flow resistance caused by mold-wall, so it is effective to avoid some defects by extending the pressure time that, which improves the casting quality greatly.

References

1. Zhong, H.R.: Warpage Optimization Based on Moldflow and DOE Technology. Plastic 40(4), 90–93 (2011)
2. Ou, Z.H.: Vacuum Casting Process CAE Analysis of Silicone Mold, pp. 45–61. Ph.D. Tianjin University, China (2008)
3. Qian, X., Jin, Y.F.: Plastic Injection Product Defects and CAE Analysis, pp. 20–25. Chemical Industry Press, China (2010)
4. Gao, Y.Z., Zhao, J.: Plastic molding CAE Technology, pp. 103–118. China Light Industry Press, China (2011)
5. Kiam Thyaqo, M., Pereira Nilson, C.: Study of injection-compression molded part using CAE analysis. Polimeros 17(1), 16–22 (2007)
6. Shen, Y.K., Liu, J.J., Chang, C.T., Chiu, C.Y.: Comparison of The Results For Semisolid and Plastic Injection Molding process. Int. Comm. Heat Mass Transfer 29(1), 97–105 (2002)
7. Xie, H.S., Li, H.J., Wang, Y.Q.: Filling Characteristics and the Influence Factors of Differential Pressure Casting for Thin Wall Part. Journal of Chongqing University 25(10), 20–22 (2002)

Role-Based Vehicle-Level CGF Entity Model Code Framework

Xiao Song, Xuecheng Shi, Guanghong Gong, Lidong Qian, and Shaoyun Zhang

Science and Technology on Aircraft Control Laboratory, School of Automation Science,
Beihang University, Beijing, P.R. China
`songxiao@buaa.edu.cn`

Abstract. Vehicle-level CGF entity model with high intelligence usually has very complex logic programming code architecture. Nowadays most simulation model generating tools can not generate CGF entity models efficiently. To overcome this shortcoming, we propose a XML based role configurable and reusable code framework. It consists of command part and action part, corresponding to behavior model and physical model respectively. The configurable assembling module assembles the two parts and forms the CGF entity. Preliminary application verifies that the code frame can satisfy the real-time requirement of current large scale system-of-systems CGF simulation, having good configurable and reusable properties.

Keywords: Computer Generated Forces (CGF), code framework, XML, role configurable, command part, action part.

1 Introduction

With the development of computer simulation technology, Computer Generated Forces (CGF) technology has been widely used in military exercises, personnel training, weapon equipment, and system research. Therefore it achieves the purpose of reducing the cost and the risk, shortening the cycle time and improving the quality.

Along with the development of combat simulation, CGF systems advance with the increase of the type and the number of combat entities. If we have no uniform and stable code architecture of CGF entities, it will make the development of each CGF system a heavy load for developers.

Moreover, CGF entities consist of not only the vehicles' physical characteristics, but also the operations of the virtual commanders and operators. Meanwhile, it responds to the event in the virtual battlefield without intervention, which makes that the implementation of its code must contain a very complex logical reasoning and business processes as well as its code framework architecture. In that case, the general generation tool of simulating models is difficult to meet the functional requirements.

The well-known One Semi-Automated Forces (OneSAF) system provides the entity editor to create, configure, check and modify the parameters of entities. But OneSAF cannot edit the process models of business behavior of entities, including the threat assessment model and decision-making of command and control.

Motivated by this, we establish a code framework and define the reusable modules of the entities model in CGF system, which can offer a stable code architecture. The developers can quickly navigate to the position of the models need to be modified so that it can endow the system with a better configurability and reusability.

2 Architecture of CGF Entity

CGF entities correspond to vehicles in the real battlefield. The vehicles can be tanks, airplanes, ships, missiles and so on. CGF entity models can be divided into two sub-models, physical model and behavioral model. Physical model refers to the physical characteristics of devices such as engine, chassis and armor etc. Behavioral model refers to the simulation model of Command and Control (C2) system. It is used to describe command and control operations and decisions of the virtual commanders.

In order to establish a general simulation model or module adaptive framework suitable for various CGF entities, we propose the framework to include the *Entity Assembling and Simulation Control module, C2 role Configuration module and the Action Process Configuration module.*

- The *Entity Assembling and Simulation Control module* uses XML format parameter to configure the role of CGF entities and integrate the C2 and Action modules. Role in this context includes two levels of meaning, the role of vehicles such as tank and airplane and the role of its troop organizations such as division, corps, battalions, companies and so on.
- The *C2 role Configuration module* corresponds to the *behavioral model* and the *Action Process Configuration module* corresponds to the *physical model*.

The CGF entity model architecture is shown in Figure 1. The architecture is divided into three layers. The ground layer of the operation support module includes the support module of comprehensive battlefield environment, the Battle Management Language (BML) and HLA/RTI. The comprehensive battlefield environment should generally include terrain, atmosphere, ocean and other environment models. BML is the command language in the battlefield which is used for information interaction. And it is also the command and control language, providing the situation awareness and sharing of intentions of combat operations based on XML file. RTI is the framework of communication to realize the distributed and interactive interoperations.

The middle layer is the module for function realization. It consists of C2 simulation module and action simulation module. The command and control simulation module includes many sub-modules such as decision-making, intelligence, communication and unit communications. Action simulation module includes such sub-modules as maneuver, firepower, protection and damage.

The top layer is the module for simulation entity configuration. It consists of C2 Role Configurable module, Action business Process Configurable model, entity assembly operation and simulation control module. Among them, the business process configurable module is used to configure the operations for all kinds of C2 business. It is the method to realize the C2 operation through the basic C2 function. For example, the collaborative business can be divided as "report the collaborative request, get instructions from the superior, contact cooperative units, command and control the cooperative units".
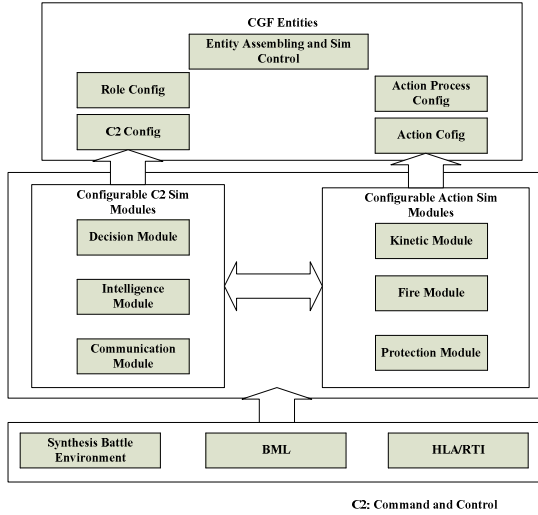
Fig. 1. Configurable architecture of CGF Entity

3   Role Configuration of CGF Entity

As shown in figure 1, modules in CGF entity model can be configured specifically. The method used in this article is to define entities' role and attributes in the entity generation level using XML files in specific format. The Document Type Definition (DTD) of the XML files is shown in Table 1.

Let EP be the set of entity parameters, it is composed of local entity list (VehicleList, which vehicle refers to aircraft, tank etc.) and task list (TaskList). The VehicleList element is the parameters of C2 role configure module, configurable C2 simulation module, and configurable action simulation module. It changes with the entities' role, C2 equipments and weapons, and it should be set accordingly. TaskList element describes the standardized procedure of C2 tasks. It defines the superior, subordinate, executing position and time etc.

EP ::= <VehicleList, TaskList>

VehicleList is composed of several Vehicles which contain some pairs of attributes and attribute-values, i.e. Vehicle::=<Attr, Attr-Value>

For example, in Figure 2, Vehicle contains attributes such as ID, Type, Man, Ammo, Oil, InitPos and EndPos.

TaskList is composed of Exec, Tar, Order, Time, ExecPos and Route.

TaskList::=<Exec, Tar, Order, Time, ExecPos, Route>

The CGF entities load the configuration files in XML before simulation and use the Config function to acquire the configuration parameters, and then use them to assign attributes in the modules such as C2 roles, Decision, Information, Action, Fire and Tasks. After all above, the CGF entities' parameter configuration is completed.

4   C2 Module

After parameter configuration, the CGF entities contain C2 part and Action part corresponding to some roles. The C2 part is composed of Decision Module, Information Module and Communication Module.

(1) Decision Module: Decision is the core of C2 system, and it simulates commander's behavior of formulating or modifying action plan according to battlefield information and the assigned tasks. The reasoning mechanism of the system is based on rules repository which is saved in XML files.

(2) Information Module: This module simulates the general functionality of the Information Department. It presets several information roles (mainly contain information collection role, information delivery role and information processing role), and the information role is finally set by the role configuration module.

(3) Communication Module: Communication in simulation includes RTI communication, process communication and thread communication. Federates communicate via RTI publish and subscribe, and they communicate with their subordinates via process communication. The internal roles in the entities share information, so they need no communication. The module aims to realize transparently processing of communication in simulation, and establish abstract communication class therefore all communication can derive from it and finally realize the information communication between combat units.

5   Action Module

The Action Module contains the module of the Kinetic module, Firepower and protection / damage.

Kinetic Module: Kinetic is the core function of the physical effects of weapons. Almost every command of the Command and Control module cannot be separated from the operations of kinetic. It consists of two modules of the kinematics and dynamics. CGF entities in large-scale system simulation can be regarded as Particles. We just need to simulate its kinematic characteristics such as speed, location, etc. Dynamics simulation of the entity is more complex. The interaction between tanks and terrain, aircraft and air requires complex computing formulas.

Firepower Module: to search targets, to analyze threatening level of the targets, to shoot targets and determine the target hitting situation. This module includes target detection model, target selection model, shells selection model, aiming and shooting model, external trajectory and projectiles spread model.

Protection Module: The features in this module are: to calculate our damage situation when our entity has been attacked by enemy's ammunition. The core functionality of the module is to classify the damage degree of the CGF entity、to calculate and determine the damage parts of the entity according to trajectory of the damage weapons. Firstly, the damage level of the simulated entity can be classified into three levels: 1.K-level of damage (destroyed) 2.M-level of damage (no mobility) 3.F-level

of damage (, loss efficacy of fire weapon). Then we will calculate the damage degree and location with features of assault weapons.

6   Assembly Module

The engineering background of this study is the development of a generating forces system of armored equipment. The system provides realistic combat environment for operational performance assessment of a certain type of armored equipment and deduction of the tri-service joint digital tactical. The main CGF entity type is troops entity of a certain type of armored equipment, including fighters, bombers, helicopters, artillery, missiles, etc.

In the system development process, the architecture of the CGF entity model uses the above model architecture and implementation. Furthermore, in the CGF entities assembly layer which is above the C2(command and control) simulation module and the operation simulation module, zzAssemble class uses configuration item to create C2 department(zzC4ISRPart) and operation department of combat unit (zzActionPart). At the same time it creates zzEntity (CGF entity) to coordinate and manage the communication and simulation process between the two internal parts.
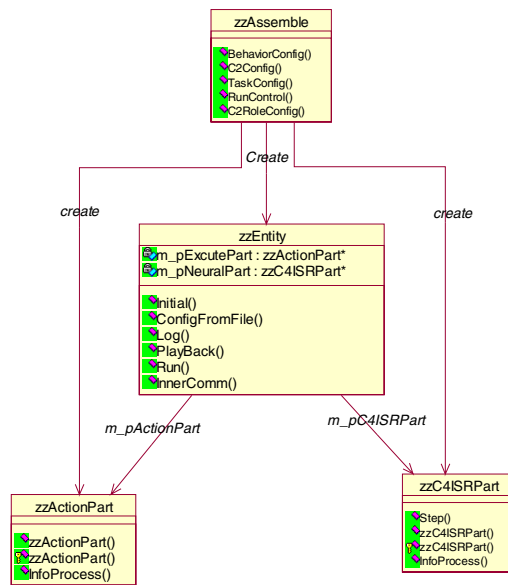
Fig. 2. UML Class Diagram of CGF Entity

Communication mechanism between the two divisions: two divisions classes provide constructor with parameters of zzEntity * and record the pointer m_pEntity; two division classes shield the default constructor; two division classes have a information processing function named void InfoProcess (byte *) to accept the

information from the opponent; when need to communicate, they reference method of zzEntity(InnerComm(byte *)) through m_pEntity to send internal information.

The system implements the information flow of the C2 department and the operation department through zzEntity. It also implements configurability of the entity through zzAssemble. A complete application is constituted by the master control units, terrain database server, multi-dimensional information display, artillery CGF, tank division, regiment, battalion, company and platform-level CGF node. In the system, all members from a division to the company and platoon use the same application to run. It just defines its role in the initial XML description file at the beginning of the program.

7   Conclusions

The complexity of CGF entity model's function decides the complexity of implementation code. Currently there is no better modeling tools. In this situation defining a more clear, strongly reusable code framework has great practical significance.

The practice shows that the CGF entity model configurable framework can be assembled into the armored force, air force, artillery and other CGF entity. It is a effective solution to the current large-scale virtual battlefield system with multi-member and multi-role.

References

1. Jin, W.: Large-Scale Simulation System. Publishing House of Electronics Industry (2004)
2. Wen, W., He, Y., Li, M.: Research on CGF Entity Framework Structure Based on the Parameter Database. Journal of System Simulation 19(6), 1231–1237 (2007)
3. Rathke, C.: Object-oriented programming and frame-based knowledge representation. In: Proceedings of the International Conference on Tools with Artificial Intelligence, pp. 95–98 (1993)
4. Guo, Q., Yang, L., Yang, R.: Computer Generated Forces Introduction. National Defence Industry Press (2006)
5. Guo, G.: Research on Modeling and Simulation of Natural Environment. National University of Defense Science and Technology (October 2004)

QVGA OLED Display Control Module
with High Gray-Level

Meihua Xu, Shihao Weng, and Mengwei Sun

Shanghai University, Shanghai, China
DanielWengSH@gmail.com

Abstract. OLED has a lot of advantages like simple structure, ultrathin, self-luminescence, high brightness, short response time, wide viewing angle, low operation voltage and so on, which is widely applied in cell phone, PDA, DC, on board display and military field. This paper represents a real-time video display system on OLED based on the detailed analysis of OLED panel electrical characteristics and various gray scale scanning principles of the OLED scan and drive circuit. FPGA is the core control device in the whole system, the DVI decoded signal is processed and real-time displayed on the OLED. 240×RGB（H）×320（V） QVGA, 256 gray scale is implemented and frame frequency is 60Hz~100Hz. The power drive has 7 outputs, which is adjustable to fit the external environment. Among them, sub-field scanning working mode is adopted in the design, gray scale is selectable.

Keywords: OLED Controller, FPGA, selectable grayscale, sub-field, RAM Manager.

1   Introduction

Organic light-emitting diodes have been discovered by 1936, when the French Destriau manufactured a film, which was the first light-emitting diodes, by dispersing the fluorescent compounds into polymer. In the 1960s, Pope achieved light-emitting with anthrancene single crystal [1]. However, the future of OLED technology was unlighted until the Kodak Company applied for a patent for the bi-layer structure of small molecule OLED instrument in 1987. In 1997, Cooperating with the Universal Display Company, the scientists from Princeton University and The University of Southern California made the first OLED display instrument sample; In the same year, the Pioneer Company brought out a 256×64 car potable screen; Motorola was the first for applying the colored OLED screen into cordless telephone [3]. With the rise of OLED technology, more and more samples and products were brought out since 2000.

OLED is consisted as the strongest competitor against LCD, the last ten years was the most important period of time for this technology. It has provided new opportunities for flat panel displays used in various products such as mobile communication device, car potable device, computer and television. In china, the OLED technology has been set a high value by the government and the related companies, and the industrialization is in progress.

This paper represents a design of high gray scale OLED video display control system based on FPGA. It was founded on basis of the electronic characteristic of the OLED display Floater from AFD Company, as well as various gray scale scanning theories. The main control device is modularized into four parts, including DVI Signal Receiving Module, DVI Signal Writing Module, RAM Management Module, and Scanning Module; the core display module adopts the sub-field scanning method. It allows both 256 and 64 levels of gray scale, achieving the high gray scale OLED video display.

The design accomplished the entire circuit system, from the video input to OLED display driver output. The system uses the FPGA as a core control device; it will handle the decoded DVI signals and send them to the OLED screen for real time display. The system can achieve 240×RGB (H) ×320 (V) QVGA for resolution, 256 levels for gray level, and 60Hz~100Hz for frame scanning frequency. There are 7 power outputs on the power driver board, which can make the brightness of the screen controllable according to the external environment.

2 The Design and Implementation of OLED Video Display Control System

2.1 Overall Design of System

Fig.1 shows the system structural drawing. The Data Receiving Module is made up of DVI-D slot, SiI1161 DVI data receiving chip, 24LC21A chip. It achieves the decoding of DVI digital signal from the graphics card. The Core Control Module is made up of Cyclone II FPGA and SRAM; it will receive the graphics data and the signals including every space and row, then accomplish the real time data storing, format transferring, scanning control etc.
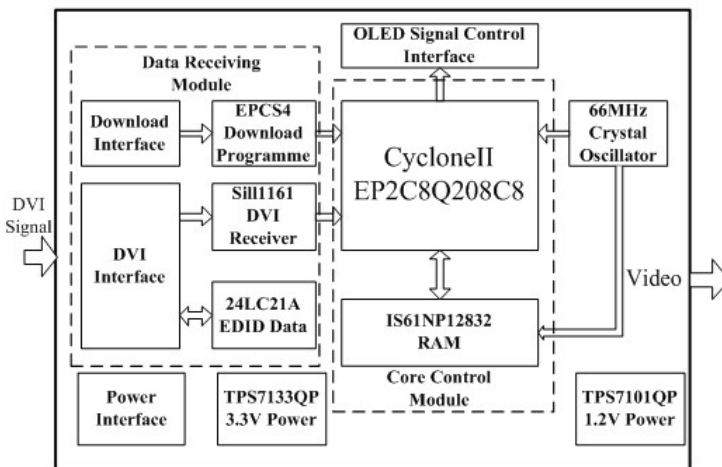
Fig. 1. System Structural Drawing

The digital signal from the graphics card is sent by the transmitter to the receiver under the TMDS coding protocol through the TMDS channel. The function of DVI signal receiving chip is: Decoding the TMDS signal from DVI interface into pixel clock signal CLK, RGB data, synchronous row signal HSYNC, synchronous space signal VSYNC, enable signal DE and control signal CTL etc. that are input data for back-end processing circuit.

The chip 24LC21A stores the EDID (Extended Display Identification Data Standard) data of DVI display instrument, which includes the standard data information transferred from the monitor to the computer through DDC. The graphics card reads the setting information of the display instrument from the chip with I2C protocol through DDC.

The FPGA will get the DVI data and finish the conversion; these transferred data should be stored in the external SRAM chip IS61NP12832. Then, the FPGA will output these image data to the OLED screen with sub-field scanning method. The frequency of the active crystal oscillator is 66MHz, the power management chips are TPS7133 and TPS7101, supplies the 3.3V and 1.2V voltage separately.

2.2   Core Control Module

The core control module accomplished the following jobs:

- The synchronization and transformation of dynamic image data;
- Reading and buffering the data;
- The control of scanning driver circuit;

Fig.2 shows the structural drawing of core control module, including DVI receiving module, DVI writing module, RAM management module, and scanning module.
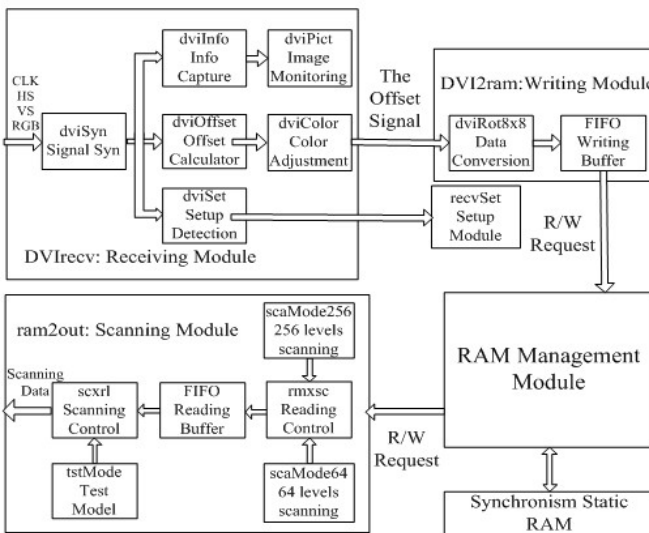
Fig. 2. Structural Drawing of Core Control Module

When the system is on, the logic circuit will first get the image data from the graphic card and read the display statement of each pixel in different sub-field according to the relationship between them. After that, these data is moved into SRAM buffer. At last, the system will start the display driving on experimental screen when the data processing is done.

The ensuing chapters will give out a detailed introduction of the RAM Management Module and Scanning Module.

2.3   RAM Management Module

On the storage of pixels, the system adopts the bit plane method, which will combine bits with the same weight of the different pixel, and store them in the continuous memory. Under the help of sub-field scanning method, the entire data will be rebuilt and reproduced as a multi-gray-level image.

The function of RAM Management Module includes: saving the signals for display scanning; controlling the reading and writing of the RAM; saving sub-field data, and send them into OLED screen. In the design, the frame buffer is a bi-frame structure, which means there are two frame buffers for image receiving and scanning separately. They work in turn in case of data loses during scanning. Each frame buffer is built by 2 pieces of SDRAM. The bi-frame buffers achieve the ping-pang structure; the SDRAM controller processes the R/W on two frame buffers in turn. The FPGA will store the sub-fieldd frame into the A frame buffer, in the meantime, another frame will be sent to OLED screen from the B frame buffer for display. When the R/W process is finished, the A buffer and the B buffer will exchange their role under the control of bus switch. The odd SDRAM of each frame buffer is separated into 8 pieces base on the logic address, as well as the double SDRAM, as one frame is divided into 8 spaces.

With the help of aaWrite and bbWrite signal, the address bus of the Group.1 RAM is connected to writing address generator while the data bus to data output bus, when the RAM is under writing statement. The writing address generator will control the system to store the image data from 8 sub-fields into the corresponding unit address.

When the next space flag signal GNT is come, the Group.1 RAM will be switch to reading statement and the Group.2 to writing statement. It's easier to get use of continuous memory address at a high reading speed. The I/O interface of RAM unit switch circuit contains the address generator port, data inputs and outputs, ports of aRAM and bRAM and the space flag signal.

Fig.3 shows the principle scheme of RAM Management Module. The inputs contains the port signal of aRAM and bRAM, the writing request from the converted data that is waiting for being moved from FIFO to SRAM, and the reading request from the gray-level scanning reading control module that is asking for the SRAM data in particular address. The outputs are interface signals of outer RAM, including pinA, pinCE, pinWEn and pinOEn etc. The ram32SVA Module will coordinate writing request from DVI side and the reading request from the Scanning side, controlling the access right of outer RAM with the requesting/authorizing mode, in order to reach the largest coefficient of utilization.

Fig. 3. Principle Scheme of ram32SVA – RAM Management Module

In Fig.3, aaREQ and bbREQ are request signals for accessing the RAM, while aaGNT and bbGNT are replay signals which will arbitrate if the writing module gets the access to visit the RAM or not. The arbitrage process comes through three phases: First, seeing whether the SRAM is occupied – if the aaGNT and bbGNT are on, it means the SRAM is occupied, otherwise, it's in idle state. Second, checking whether there is a request or not – the module will see if the aaREQ or bbREQ is on. When there is an accessing request and the SRAM is in idle state, the request will be accepted and aaGNT or bbGNT will be turn on to show the SRAM is in busy state. At last, when aaREQ or bbREQ is off, which means the memory visiting is finished, the module will retract the access, and the SRAM will go back to idle state. In the RAM Management Module, the aaREQ holds a higher priority level than the bbREQ. When aaREQ and bbREQ are both on, the system will deal with the aaREQ first, which is the writing request from asynchronous FIFO buffer.

2.4   Scanning Module

2.4.1   Scanning Module (ram2out)

Fig.4 shows the principle scheme of ram2out module, the inputs are system clock signal oscCLK, rstn; rtsF, rstA[16:0],rstQ[23:0],rstRead, rstWrite, bbrstNew from RAM; and the test signal – tsetData[47:0], tsetNew. The outputs includes the interface signals linking to bRAM, and the screen driver signals – SCK, SCKB, SSP, PWC, GCK, GCKB, G1SP, G2SP, MR[8:1], MG[8:1], MB[8:1], MNC[4:3].
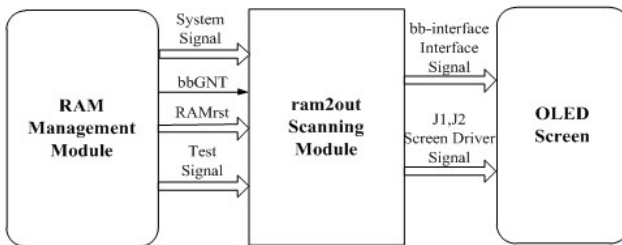
Fig. 4. Principle Scheme of ram2out – Scanning Module

2.4.2 Core Imaging Module

The active OLED screen scanning frequency in this system can be set to 100Hz or 60Hz. There are 11 spaces for every frame; the first ten ones are data spaces and the last DS sub-field, which is used for filling in order to reach the frequency, is a connection with no content. The ten sub-fields are SF3-SF11-SF4-SF21-SF12-SF5-SF13-SF22-SF14-SF6 in order; each of them should transfer 320 rows of data and 16 empty rows, and every row transmission contains 30 times of data and 6 times of empty column. Because 24 bits of data can be transferred in a SCK/SCKB clock, there are 240 columns×3 colors of data during the 30 times transmission for each row.

Among the ten sub-fields, SF11~SF14 hold the highest weight (32) of gray-level data; each of them can illume 168 time units, which is the longest display period. SF21~SF22 hold the weight of 16; and they can illume a longest display period as well. SF3 only carries the weight of 8, but it can also illume a longest display period. SF4 weights 4 only, it is able to illume half of the longest display period (the G2SP signal will turn off the display after 84 time units). SF5 weights 2, can illume 1/4 of the longest display period (42 time units). SF6 only weights 1, which illumes 1/8 of the longest display period (21 time units).

Fig.5 shows the internal structure of the module, the address generation module will accomplish all the control signals, including SCK, SCKB, GCK, GCKB, G1SP, G2SP, that are connected to the active OLED screen. The address is generated by three counters – row counter (with the counting region of 0~337), column counter (0~35) and sub-field counter (0~10), whose counting region is set base on the size of the screen and the partition of the sub-fields. The row address (RowAddr), column address (ColAddr) and bit weight address (BitAddr) that are brought out by the address generator will be sent into the reading request module, asking RAM for reading the corresponding RGB data and sending them into the buffer. When the control signals, such as SCK, are stored into the buffer as well, all the data should be wrote into the double-port RAM, and sent to the active OLED screen synchronously. In case of the overflow of the buffer, there is an extra module called buffer data comparison, which will send a flag signal to tell the address generator if the buffer is available. The scanning module also can choose the particular working frequency by switch the transfer speed of the data that is sending to the OLED screen.
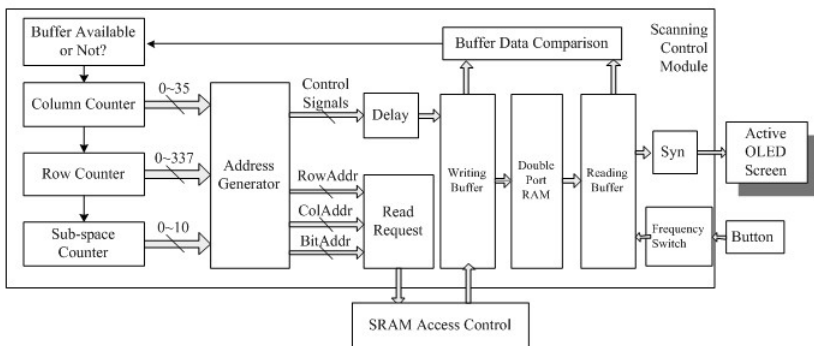
Fig. 5. Internal Structure of Core Imaging Module

There are two options for gray-level in the system: 256 levels and 64 levels. Taking 64 levels for example, 6 bits are enough for RGB data. In the imaging module, the lower 2 bits in the address are neglected, which means using [7:2] and ignoring the effect of [1:0]. This shall produce a smaller effect on the original gray-level than neglecting the higher 2 bits.

2.4.3   Scanning Sequence Module (rmxsc) and Scanning Output Module (scxrl)

The scaMode256 Module produces the 256-level scanning control sequence, there are gCLK, rstn, paraGrayLen, GetNext for inputs and rowAddr, colAddr, bitAddr, readENA, outFlag for outputs. When the internal sub-frame counter frmCT reaches different numbers, the bitAddr will be given various paraGrayLen (which presents the gray-level parameter) from 0~7. While the counter reaches 16, one frame will be over. The scaMode64 Module works in the same way. The ports of rmxsc module are system clock OSC; the communication signals to SSRAM: ramREQ, ramGNT, reqF[NF-1:0], reqA[16:0], reqRead, rstF[NF-1:0], rstA[16:0], rstQ[23:0], rstRead, rstNew; scanning mode signals: GetNext, rowAddr[8:0], colAddr[7:3], bitAddr[2:0], readENA, outFlag[7:0]; and the scanning data buff control signals: cswAddr[9:0], cswData[49:0], cswENA, cswBlock[4:0], csrBlock[4:0]. The rmxsc module can be set to a kind of scanning sequence, which is used for producing the memory accessing address sequence. The reading request will be sent to the RAM Manager base on the address sequence. After getting the access right, data and control information are wrote into the FIFO buffer. The scxrl module possesses the ports including system clock signal OSC; scanning data buffer control signals: csrAddr[9:0], csrENA, csrq[49:0], csrBlock[4:0], cswBlock[4:0]; and the scanning data outputs: rscFlag[7:0], rscR[7:0], rscG[7:0], rscB[7:0]. This module reads the scanning data from FIFO and sends them into the OLED display screen at a stable speed; in the test mode, tstMode module is used for test data generation.
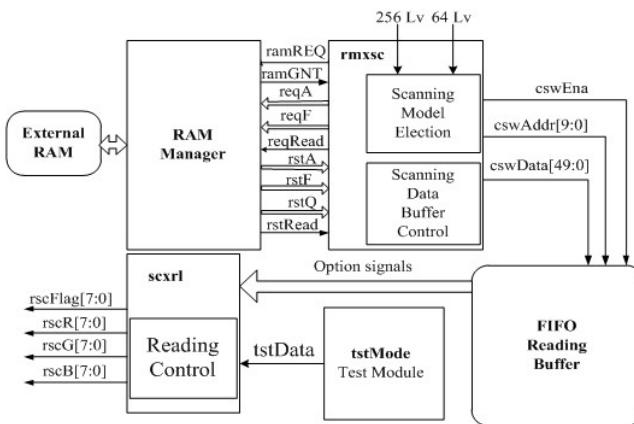
Fig. 6. Principle Scheme of rmxsc and scxrl Module

3   Experiment Result and System Verification

As shown in Fig.7, oscCLK is the overall system clock signal. In the DVI receiving chip, there are data clock dCLK, row syn. control signal dHS, column syn. control signal dVS. The dQE outputs the 24 bits image data. The dSCDT is used for triggering the external circuit to process syn. checking. Synchronous RAM possesses the 32-bit syn. data inputs and outputs, address signal rA, syn. R/W control signal, chip access signal and output access signal. As the sub-field scanning method is adopted, the data is converted and offset.
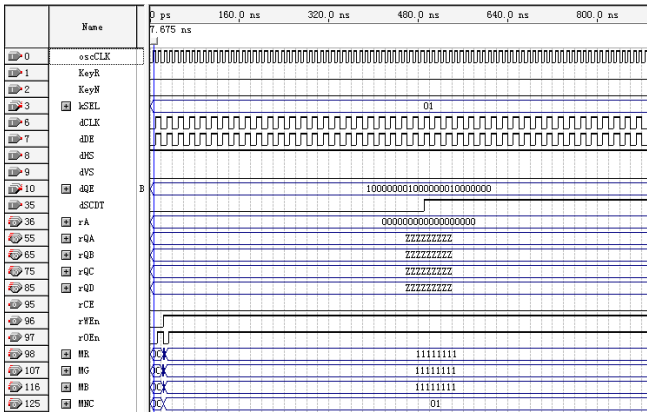
Fig. 7. Logic Design Co-Simulation

This paper accomplished the entire circuit system, from the video input to OLED display driver output. The system uses the FPGA as a core control device. The decoded DVI signals will be sent to the OLED screen for real time display. The system can achieve 240×RGB（H）×320（V）QVGA for resolution, 256 levels for gray level, and 60Hz~100Hz for frame scanning frequency, 16777216 colors. There are 7 power outputs on the power driver board, each of them can work separately. The main control board gets the external image signal from the graphic card. The OLED system work clock pulse is generated by crystal oscillator.

References

1. Sun, M.: Research and Realization of Display System Based on Organic Light Emitting Diode (OLED). Master's Thesis of East China Normal University
2. Jiang, Q.Q.: OLED technology and it's market development. TV Engineering 247, 53–55 (2003), doi:CNKI:SUN:DSSS.0.2003-01-019

3. Wang, L.: OLED display technology. Electronic Engineering & Product World 206, 91–93 (2006), doi:CNKI:SUN:DZCS.0.2006-01-041
4. Notarangelo, G., Gibilaro, M., et al.: Low power strategy for a TFT controller. In: Proceedings of the Euromicro Symposium on Digital System Design, vol. 4-6, pp. 351–354 (2002)
5. Blades, C.D.J., Walker, A.B.: Simulation of organic light-emitting diodes. Synthetic Metals 111-112, 335–340 (2000)
6. Guo, B.: Driving Method of sub-frame and current-ratio for PM-OLED. Chinese Journal of Liquid Crystals and Displays 17, 182–188 (2002), doi:CNKI:SUN:YJYS.0.2002-03-009

Wave-Based Reflections Reducing Approach for Bilateral Teleoperation

Min Zheng, Wei Xiao, and Qinghai Chen

College of Mechatronic Engineering and Automation, Shanghai University
Shanghai Key Laboratory of Power Station Automation Technology,
Shanghai 200072, China

Abstract. The wave-based method can guarantee the passivity of the bilateral teleoperation system, but the ensuing wave reflections will deteriorate the operation performance. This paper presents a new wave-based bilateral teleoperation structures to attenuate the wave reflection, at the same time, by scaling the structure parameters to improve the force and speed tracking performance. And according to the wave scattering theory, passivity of the new structure is analyzed. Experiments show that the proposed method not only guarantees the stability of the system, but also improves the position tracking and force feedback performance.

Keywords: Teleoperation, Wave variable, Passivity.

1   Introduction

Teleoperation plays an important role in some challenging environments such as outer space, exploration in deep sea, nuclear industrials and so on, in which a human operator conducts a task in a remote environment via master and slave manipulators. Nevertheless, the instability of teleoperation can occur because of the time-delay between the two sides of the teleoperation. Anderson and Spong introduced the concept of scattering variables, which was inspired by transmission line theory, to the problem of bilateral teleoperation [7]. Through this theory, the passivity of teleoperation can be guaranteed at any constant time-delay [1-2]. A human operator manipulate at the master side, and then, the commands about position and velocity are sent to the slave side, while the slave manipulator works on the environment and the information of force reflecting goes back to the master side. The architecture of the bilateral teleoperation is presented in fig.1. The wave variables are as follows,

$$u_m = \frac{1}{\sqrt{2b}}(f_{md} + b\dot{x}_{md})$$

(1)

$$u_s = \frac{1}{\sqrt{2b}}(f_{sd} + b\dot{x}_{sd})$$

(2)

$$v_m = \frac{1}{\sqrt{2b}}(f_{md} - b\dot{x}_{md}) \tag{3}$$

$$v_s = \frac{1}{\sqrt{2b}}(f_{sd} - b\dot{x}_{sd}) \tag{4}$$

We can also get the position tracking and force reflecting expressions, as follows,

$$\dot{x}_{sd}(t) = \dot{x}_{md}(t-T) + \frac{f_{md}(t-T) - f_{sd}(t)}{b} \tag{3}$$

$$f_{md}(t) = b(\dot{x}_{md}(t) - \dot{x}_{sd}(t-T)) + f_{sd}(t-T) \tag{4}$$

where $\dot{x}_m$ and f_{md} are velocity and force for the master manipulator, respectively ; f_{sd} is force of the motor employed to the slave manipulator ; $\dot{x}_{sd}$ denotes the velocity command from the slave manipulator ; f_h denotes the force employed by the operator ; f_e denotes the force employed by the environment ; u_m , v_s are the outgoing wave variables while u_s , v_m denote the corresponding incoming wave variable, respectively ; M is the mass inertia and b denotes the characteristic impedance of the wave transmission [3]

Wave-based teleoperation will unavoidably cause the reflections and the phenomena of the reflections will corrupt the useful information flow or even cause the oscillatory behavior [3]. As shown in fig.2, there exist four paths of reflections in basic wave-based teleoperation. Niemeyer introduced impedance matched method in order to eliminate the reflections [3-5], as shown in fig.3. The outgoing wave variable u_m simply contains the velocity information while the outgoing wave variable v_s simply contains the force information, as follows.

$$u_m(t) = \frac{b\dot{x}_{md}(t)}{\sqrt{2b}} \tag{7}$$

$$v_s(t) = \frac{f_{sd}(t)}{\sqrt{2b}} \tag{8}$$

As impedance matched, the outgoing wave variable u_m and v_s do not contain the information of u_s and v_s any more, i.e. they are decoupled. The phenomena of reflections occurring in fig.2 are eliminated. However, new equations of velocity tracking and force reflecting can be formulated with the element of matched impedance [3][5]. $\dot{x}_{sd}$ and f_{md} are as follows.

$$\dot{x}_{sd}(t) = \frac{b\dot{x}_{md}(t-T) - f_{sd}(t)}{2b} \tag{9}$$

$$f_{md}(t) = \frac{b\dot{x}_{md}(t) + f_{sd}(t-T)}{2} \tag{10}$$

Seeing from the equations (9) and (10) , when $f_{sd} \to 0$ occurs, then $\dot{x}_{sd} \to \dfrac{\dot{x}_{md}}{2}$;

When the master manipulator moves slowly, i.e. $\dot{x}_{md} \to 0$, then $f_{md} \to \dfrac{f_{sd}}{2}$. Position

tracking and force reflecting stay only half of the expected value, that is, teleoperation does not behavior well. Besides, matched impedance need to base on the force reflecting information containing the accurate damping or friction components. Therefore, complete matched impedance is impractical.
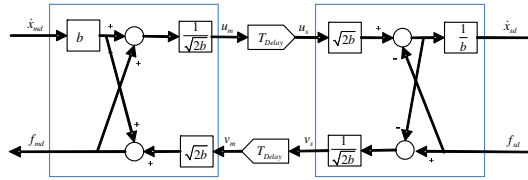
Fig. 1. Basic wave-based bilateral teleoperation architecture
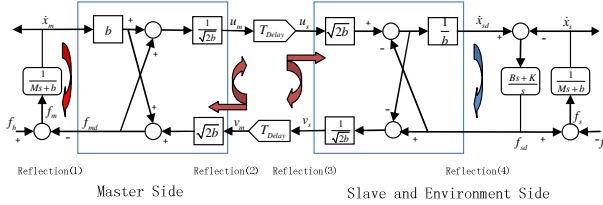
Fig. 2. Schematic diagram of paths of wave reflections
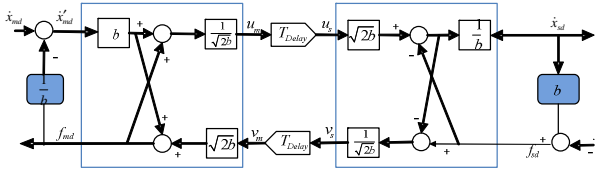
Fig. 3. Impedance matched teleoperator

A.C. Smith et al. proposed architectures consist of a novel pseudo two-channel nonlinear predictive controller and its variations that use neural networks for online estimation of the slave and environment dynamics to replicate the environment contact force at the master side using a similar local network [6]. However, to choose an appropriate neural network structure can be a tedious and time-consuming task. When estimating the environment is no effective, the performance of position
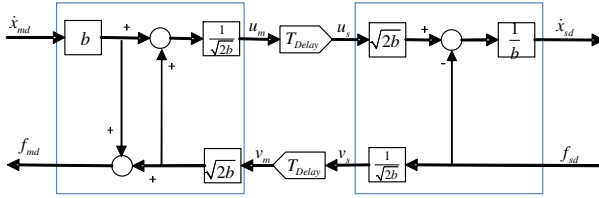
Fig. 4. Configuration of reducing wave-based teleoperator

tracking and force reflecting will be degraded. Besides, this method cannot guarantee the stability theoretically, so it risks in applications.

Teleoperation was developed quickly in the applications and concepts e.g. haptic transmission, are introduced into teleoperation systems [7-11]. However, these teleoperation systems cannot attenuate or cancel reflections. Furthermore position drift and force distorting occur in these systems and rigid conditions are required, e.g. using haptic method to describe the environment needs the prior knowledge but it is not possible for any environment. Kenji Natori introduced the notion of network disturbance and time-delay got compensated by estimating the communication disturbance in the bilateral teleoperation, which made it possible to transmit through the Internet [12]. Researches on teleoperation through a variety of unreliable networks were wildly observed [13-16]. All the methods mentioned above base on the prior knowledge and attenuating or eliminate reflections are merely concerned.

L.Bate et. al lately proposed a reducing reflections scheme [17], seeing fig.3. Using this method , reflections can be reduced, but when the characteristic impedance b gets greater, we cannot get effective force reflecting under the state of the master manipulator moving, because at that state we get,

$$F_m(t) = b\dot{x}_m(t) + F_s(t-T).$$

When the characteristic impedance b gets smaller, we cannot get effective velocity tracking, because at that condition, we get,

$$\dot{x}_s(t) = \dot{x}_m(t-T) + \left(F_s(t-T) - F_s(t)\right)/b.$$

A new architecture with adjusting factors is proposed in this paper in order to get better behavior. Based on scattering theory, the condition of system's passivity is also proposed. Finally, experiments verify the validness.

2   New Configuration of Teleoperation System and Its Stability Analysis

2.1   New Configuration of Teleoperation System

A new structured flowchart of wave-based teleoperation is proposed, see fig.5. And a new combination of wave variables is as follows.

$$u_m(t) = \frac{1}{\sqrt{2b}} (\beta f_{sd}(t-T) + b\dot{x}_{md}(t)) \tag{11}$$

$$u_s(t) = \frac{1}{\sqrt{2b}} (\beta f_{sd}(t) + b\dot{x}_{sd}(t)) \tag{12}$$

$$v_m(t) = \frac{1}{\sqrt{2b}} (f_{md}(t) - \alpha b\dot{x}_{md}(t))) \tag{13}$$

$$v_s(t) = \frac{1}{\sqrt{2b}} f_{sd}(t) \tag{14}$$

Equation(14) shows v_s does not contain any velocity information any more, i.e., the outgoing variable v_s of the slave side does not has anything to do with the incoming variable u_s of the slave side, which means the path 3 of reflections is eliminated with the information of force and velocity decoupled in fig.3. Whether through equation (11) or through fig.5, it is clear that the influence between v_m and u_m gets smaller with the reflection adujusting factor β , because $0 < \beta < 1$, i.e. the path 2 of reflections in fig.2 has been attenuated already. As the same way understanding, through fig.5, it is also clear that the influence between f_{md} and $\dot{x}_{md}$ becomes smaller with the reflection adujusting factor α , because $0 < \alpha < 1$, i.e. the path 1 of reflections in fig.2 has been attenuated already. The new equations of velocity tracking and force reflecting in fig.5 is supposed to be as follows.

$$\dot{x}_{sd}(t) = \dot{x}_{md}(t-T) + \frac{\beta(f_{sd}(t-2T) - f_{sd}(t))}{b} \tag{15}$$

$$f_{md}(t) = \alpha b\dot{x}_{md}(t) + f_{sd}(t-T) \tag{16}$$

Herein, we tranform equation (15) and (16) by Lapalace and use matix form to describe them as follows.

$$\begin{bmatrix} F_{md}(s) \\ -\dot{X}_{sd}(S) \end{bmatrix} = \begin{bmatrix} \alpha b & e^{-sT} \\ -e^{-sT} & \dfrac{\beta(e^{sT} - e^{-sT})}{be^{sT}} \end{bmatrix} \begin{bmatrix} \dot{X}_{md}(S) \\ F_{sd}(s) \end{bmatrix}$$

$$= H(s) \begin{bmatrix} \dot{X}_{md}(S) \\ F_{sd}(s) \end{bmatrix} \tag{17}$$

Where, *H(s)* is defined as mixture matrix [7].

Both position tracking and force reflecting in fig.5 performance better than those in fig.4 with reflection adjusting factors by comparing fig.4 and fig.5. Besides, the new proposal method need not estimate or model the environment in order to reduce reflections. Since this method do not require the prior knowledge of the environment, reflections can not be reinstated with different environment.
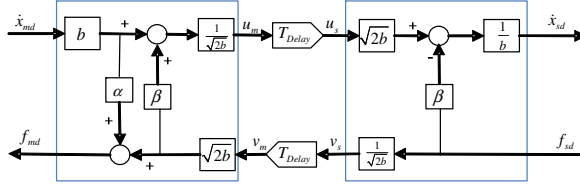
Fig. 5. Schematic diagram of a new wave-based teleoperator

2.2   Analysis of Passivity and Stability

The wave-based method can guarantee the passivity of the bilateral teleoperation system with communication time-delays. The form of scattering matrix shows below,

$$\Xi(s) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} (H(s)-I)(H(s)-I)^{-1} \tag{18}$$

Based on the passivity theory, we get the following proposition.

Proposition: As to the new architecture in fig.5, suppose that the characteristic impedance $b>0$, communication delay $T>0$, if there exist $0<\alpha<1$, $0<\beta<1$ and inequality $\left\| \Xi(s) \right\| \leq 1$ is satisfied, systems must be passive and stable.

Where,

$$\Xi(s) = \begin{bmatrix} \Xi_{11}(s) & \Xi_{12}(s) \\ \Xi_{21}(s) & \Xi_{22}(s) \end{bmatrix} \tag{19}$$

$$\Xi_{11} = 1 - \frac{2be^{Ts} + 4\beta\sinh(Ts)}{2b\cosh(Ts) + 2\beta\sinh(Ts) + \alpha b^2 e^{Ts} + 2\alpha\beta b\sinh(Ts)}$$

$$\Xi_{12} = \frac{2b}{2b\cosh(Ts) + 2\beta\sinh(Ts) + \alpha b^2 e^{Ts} + 2\alpha\beta b\sinh(Ts)}$$

$$\Xi_{12} = \Xi_{21}$$

$$\Xi_{22} = -1 + \frac{2\alpha b^2 e^{Ts} + 2be^{Ts}}{2b\cosh(Ts) + 2\beta\sinh(Ts) + \alpha b^2 e^{Ts} + 2\alpha\beta b\sinh(Ts)}$$

3   Simulation Results

In this chapter, experiments on position tracking and force reflecting are implemented. Fig.7 and fig.8 are position tracking and force reflecting of basic wave-based teleoperation architecture, respectively. The scattering norm value can be described by fig.6.
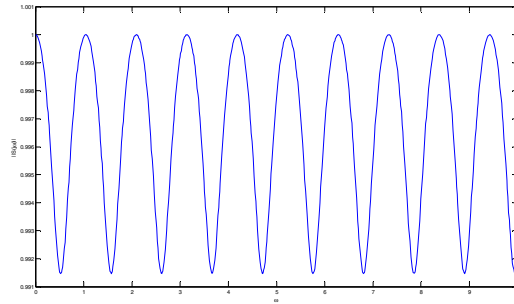
Fig. 6. Scattering norm with appropriate reflection scaling parameters
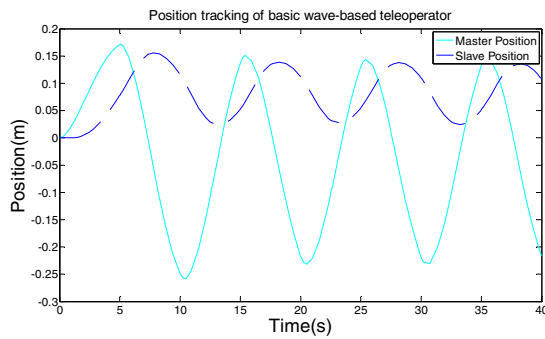
Fig. 7. Position tracking of basic wave-based teleoperator in fig.2
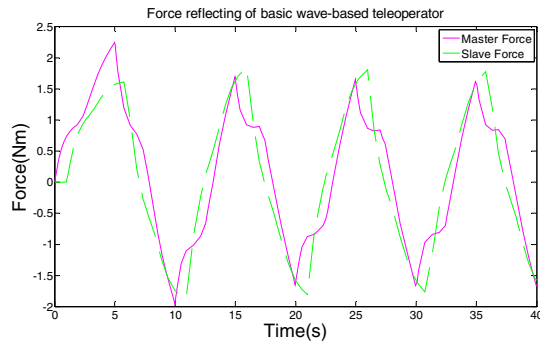
Fig. 8. Force Reflecting of basic wave-based teleoperator in fig.2

In fig.9 and fig.10, both position tracking and force reflecting of new proposed method are compared with that in fig.4. Herein, the reflection adjusting factors of new propose method are chosen as α=0.7 , β=0.8 . Comparing the two figures, both position tracking and force reflecting of new proposal are better.
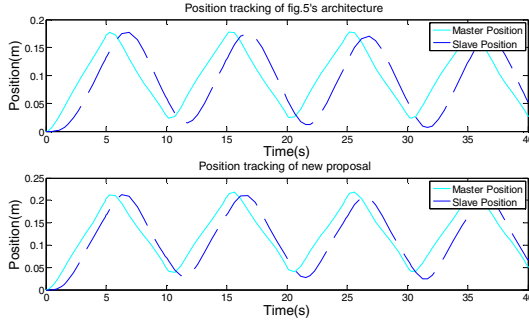
Fig. 9. Comparison of position tracking between fig.4's architecture and new-proposal
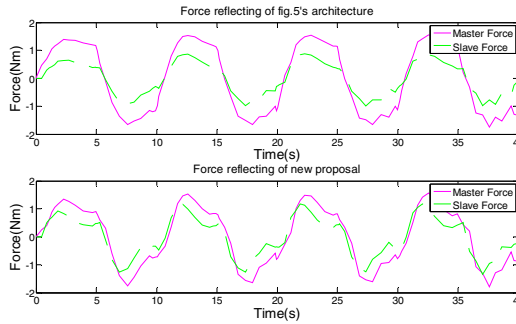
Fig. 10. Comparison of force reflecting between fig.4's architecture and new-proposal

4   Conclusions

A new architecture of wave-based bilateral teleoperation is proposed in order to attenuate or eliminate reflections and also improve the tracking performance. Rule for guaranteeing the passivity of the teleoperation systems are also presented. This method is compared to the conventional wave-based method and experimental results lead to the effectiveness of attenuating or eliminating wave reflections, not even knowing the prior knowledge of the environment. Better performances of position tracking and force reflecting are realized under unknown conditions.

References

1. Anderson, R., Spong, M.: Asymptotic stability for force reflecting teleoperators with time delay. In: Proceedings IEEE International Conference on Robotic Automation, vol. 3, pp. 1618–1625 (1989)
2. Anderson, R., Spong, M.: Bilateral control of teleoperators with time delay. IEEE Transactions on Automatic Control 34(5), 494–501 (1989)
3. Niemeyer, G., Slotine, J.: Stable adaptive teleoperation. IEEE Journal of Oceanic Engineering 16(1), 152–162 (1991)
4. Benedetti, C., Franchini, M., Fiorini, P.: Stable tracking in variable time-delay teleoperation. In: Proceedings IEEE International Conference on Intelligent Robots Systems, vol. 3, pp. 2252–2257 (2001)
5. Ching, H.: Internet-based bilateral teleoperation. Georgia Institute of Technology, Atlanta (2006)
6. Smith, A.C., Hashtrudi-Zaad, K.: Smith Predictor Type Control Architectures for Time Delayed Teleoperation. Journal of Robotic Research, 152–162 (2006)
7. Hokayem, P.F., Spong, M.W.: Bilateral teleoperation: An historical survey. Automatica 42(2), 2035–2057 (2006)
8. Uchimura, Y., Yakoh, T.: Bilateral robot system on the real-time network structure. IEEE Transactions on Industrial Electronics 51(5), 940–946 (2004)
9. Oboe, R.: Web-interfaced, force-reflecting teleoperation systems. IEEE Transactions on Industrial Electronics 48(6), 1257–1265 (2001)
10. Slama, T., Trevisani, A., Aubry, D., Oboe, R., Kratz, F.: Experimental analysis of an Internet-based bilateral teleoperation system with motion and force scaling using a model predictive controller. IEEE Transactions on Industrial Electronics 55(9), 3290–3299 (2008)
11. Tian, D., Yashiro, D., Ohnishi, K.: Haptic transmission by weighting control under time-varying communication delay. IET Control Theory 6(3), 420–429 (2012)
12. Natori, K., Tsuji, T., Ohnishi, K.: Time-Delay Compensation by Communication Disturbance Observer for Bilateral Teleoperation Under Time-Varying Delay. IEEE Transactions on Industrial Electronics 57(3), 1050–1062 (2010)
13. Nuno, E., Basanez, L., Ortega, R., Spong, M.W.: Position tracking for nonlinear teleoperators with variable time delay. I. The Journal of Robotic Research 28(7), 895–910 (2009)
14. Kawashima, K., Tadano, K., Wang, C., Sankaranarayanan, G., Hannaford, B.: Bilateral teleoperation with time delay using modified wave variable based controller. In: Proceedings IEEE International Conference on Robotic Automation, pp. 4326–4331 (2009)
15. Chopra, N., Berestesky, P., Spong, M.W.: Bilateral teleoperation over unreliable communication networks. IEEE Transactions on Control Systems Technology 16(2), 304–313 (2008)
16. Huijun, L., Aiguo, S.: Virtual-environment modeling and correction for forcereflecting teleoperation with time delay. IEEE Transactions on Industrial Electronics 16(2), 304–313 (2008)
17. Bate, L., Cook, C.D., Li, Z.: Reducing Wave-Based Teleoperator Reflections for Unknown Environments. IEEE Transactions on Industrial Electronics 58(2), 392–397 (2011)

Recovering Three-Dimensional Surfaces
with Multi-images Shape-From-Shading Method

Lei Yang and Ning Zhang

School of Mechatronic Engineering & Automation, Shanghai Key Laboratory of Power Station
Automation Technology, Shanghai University, NO.149,Yanchang Rd.
200072 Shanghai, China

Abstract. Three-dimensional (3-D) shape reconstruction is one of the
fundamental problems in the field of computer version. Most existing
shape-from-shading (SFS) methods are based on signal image and orthogonal
projection. But the reflectance map equation is a nonlinear partial differential
equation about two random variables. So the SFS is an ill-posed problem. Further
more, orthogonal projection used to simulate the imaging processes of camera is
not very accurate. This paper proposes a new SFS method under perspective
projection with multi-images. Three images with different lighting source
directions are captured by camera firstly. Following three reflectance map
equations which are described by Lambertain model are established. Then the
gradient vectors of the 3-D surface are calculated by solving the reflectance map
equations. The gray constraint and gradient component constraint conditions are
used to construct target function, and the corresponding Eulor-Poision equations
are derived. Simultaneously, discrete difference is used to approximate
differential operation. New iterative 3-D shape reconstruction algorithm is
proposed by the discrete difference equation. Three pixel values are used to solve
certain gradient value in our method. So the ill-posed problem in traditional SFS
which solves a single reflectance map equation can be avoided. At last,
experimental results of 3-D reconstruction show that the proposed method is
effective.

Keywords: SFS method, perspective projection, multi-images, point lighting
source, reflectance map equation.

1 Introduction

Three-dimensional shape recovery for images is becoming more and more important
because of its widely potential applications. Such as, industrial production, medical
image reconstruction, SAR image topographic survey, and so on. In the early 1970s,
Horn has already proposed SFS algorithm to reconstruct object surfaces. The original
SFS methods proposed by Horn minimize the total error function consisting of one or
several terms of the brightness constraint, the smoothness constraint by using the
principle of variations [1]. Woodham first introduced photometric stereo for Lambertain
surfaces assuming known albedos and known lighting directions. The method was
based on the use of the reflectance maps in the form of look-up tables. Three images

were used to solve the reflectance equation for recovering surface gradients and albedos of Lambertain surface [2]. In recent years, finite element algorithm, neural network method, level set method and viscosity solution algorithm have been proposed [3-7], and so on.

The photometric stereo technique uses several images of the same surface taken from the same viewpoint under different lighting directions. From obtained images, the changes of the intensities in the images depend only on the local surface orientation that can be recovered by combining that information [8]. Su proposed a shape from shading method under orthogonal projection by three images to calculate the surface normal vectors. Through some reflectance equations, the only gradient vector can be obtained and the method is effective to avoid the ill-posed problem [9].

In this paper, we present a simple photometric stereo method for shape recovery of three- dimensional object under perspective projection. As we know, perspective projection is more accurate than orthogonal projection in which the light source is assumed to be located at infinite distance. In the last ten years, perspective projection was gradually adopted in SFS [10-12]. On the one hand, to calculate reflectance map equations by three images gray value is effective to deal with the difficult problem which will be caused by single image. On the other hand, the only obtained gradient vector and iterative 3-D shape reconstruction algorithm are used to get the deep information of object surface easily. Remainders of this paper are organized as follows. Section 2 formulate image model and a series of reflectance equations is given. In Section 3, the normal vectors and the only gradient value are acquired based on images gray value under three different point lighting source directions. And recovering object surface information by iterative calculation is completed. In Section 4, experimental result of synthetic ball and vase are presented. Finally, conclusions and future work are drawn in Section 5.

2   The Formulated Model and Perspective Reflectance Equations

It is almost impossible to establish accurate and unique imaging model in SFS methods because of different imaging conditions in practice. However, recovering object shape by single image is to calculate reflectance equation:

$$E(x, y) = R(p(x, y), q(x, y)) \tag{1}$$

where $E(x, y)$ is the captured image, $R(p(x, y), q(x, y))$ is imaging reflectance map, $(p(x, y), q(x, y))$ is the gradient vector of object surface in point (x, y). If following assumptions are made: the lighting source is located at infinite distance with a known direction $\vec{s} = (p_0, q_0, -1)$, surface normal $\vec{n}$ is denoted by $(p, q, -1)$, ρ is constant albedo of Lambertain reflectance surface and orthogonal projection

under Lambertain model is used to simulate the imaging process, following reflectance map equation is established:

$$R(p,q) = \rho \frac{pp_0 + qq_0 + 1}{\sqrt{p^2 + q^2 + 1}\sqrt{p_0^2 + q_0^2 + 1}} \tag{2}$$

If we let the captured image and reflectance equation normalization, then following reflectance map equation is established:

$$I = \frac{E - E_{min}}{E_{max} - E_{min}} = \frac{\vec{n} * \vec{s}}{\|\vec{n}\|\|\vec{s}\|} = \frac{pp_0 + qq_0 + 1}{\sqrt{p^2 + q^2 + 1}\sqrt{p_0^2 + q_0^2 + 1}} \tag{3}$$

where I is the normalized image, E_{max} and E_{min} denote the maximum and minimum grey values of the captured image E, and $(p,q) = (\partial z/\partial x, \partial z/\partial y)$, $z = z(x, y)$ is the surface in scene. After Horn proposed the reflectance map equation, different mathematic techniques were adopted to solve the equation [13].

Because of near-scene imaging situation, these conditions are reasonable : 1. Perspective projection is applied, 2.Point light source is located at the optical center. The surface S in scene is defined as $z = z(x, y)$, and pixel is denoted as (u, v) at imaging plane $z = f$, where f is focal length. The perspective projection equations is

$$\frac{x}{u} = \frac{y}{v} = \frac{z}{f} \tag{4}$$

From Eq.(4), we have

$$\begin{cases} x = u\dfrac{z}{f} \\ y = v\dfrac{z}{f} \end{cases} \tag{5}$$

The surface S is parameterized in imaging plane such as

$$(x, y, z) = z(u, v) * (\frac{u}{f}, \frac{v}{f}, 1) \tag{6}$$

By using the difference geometry theory, the normal of the surface is given by

$$\vec{n}(u,v)=\vec{S}_u\times\vec{S}_v=\begin{vmatrix} i & j & k \\ \dfrac{\partial x}{\partial u} & \dfrac{\partial y}{\partial u} & \dfrac{\partial z}{\partial u} \\ \dfrac{\partial x}{\partial v} & \dfrac{\partial y}{\partial v} & \dfrac{\partial z}{\partial v} \end{vmatrix}=\begin{vmatrix} i & j & k \\ \dfrac{z_u u+z}{f} & \dfrac{z_u v}{f} & z_u \\ \dfrac{uz_v}{f} & \dfrac{z_v+z}{f} & z_v \end{vmatrix} \tag{7}$$

$$=(-f\frac{\partial z}{\partial u},-f\frac{\partial z}{\partial v},u\frac{\partial z}{\partial u}+v\frac{\partial z}{\partial v}+z(u,v))*\frac{z(u,v)}{f^2}$$

So the unit normal of the surface is

$$\vec{n}=\frac{(-fz_u,-fz_v,uz_u+vz_v+z)}{\sqrt{(uz_u+vz_v+z)^2+f^2(z_u^2+z_v^2)}} \tag{8}$$

where $(z_u,z_v)=(\partial z/\partial u,\partial z/\partial v)$ and if the point lighting source vector is $\vec{n}_i=(p_i,q_i,-1)$. So the reflectance map equation of Lambertain model under perspective projection is concluded from Eq.(3) as:

$$I(u,v)=\frac{\vec{n}_i*\vec{n}}{|\vec{n}_i|\|\vec{n}\|}=\frac{(u-fp_i)\dfrac{\partial z}{\partial u}+(v-fq_i)\dfrac{\partial z}{\partial v}+z}{\sqrt{p_i^2+q_i^2+1}\sqrt{(f\dfrac{\partial z}{\partial u})^2+(f\dfrac{\partial z}{\partial v})^2+(u\dfrac{\partial z}{\partial u}+v\dfrac{\partial z}{\partial v}+z)^2}} \tag{9}$$

Let $Z(u,v)=\ln z(u,v)$, $\partial Z/\partial u=Z_u,\partial Z/\partial v=Z_v$ and we denote $z(u,v)>0$,then we have

$$I(u,v)=\frac{(u-fp_i)Z_u+(v-fq_i)Z_v+1}{\sqrt{p_i^2+q_i^2+1}\sqrt{(fZ_u)^2+(fZ_v)^2+(uZ_u+vZ_v+1)^2}} \tag{10}$$

Equation (10) can be further written as

$$I(u,v)=\frac{(u-fp_i)p+(v-fq_i)q+1}{\sqrt{p_i^2+q_i^2+1}\sqrt{(fp)^2+(fq)^2+(up+vq+1)^2}} \tag{11}$$

where $p=Z_u,q=Z_v$ .Equation (11) is a first nonlinear partial differential equation, so multi-images reflectance map equations will be used to solve this question[14].

3   Calculate Surface Gradient Vector Based on Multi-images

The method to get object surface gradient vector based on multi-images, which has three steps following as: a. the camera location is fixed and we obtain three images from different point lighting sources. b. every pixel grey value is extracted from that images. c. we can construct three reflectance equations at every pixel point and get the only gradient vector. Avoiding matching three imaging corresponding point is one of our method advantages. We assume these imaging pixel points have not shadows. And if the shadow exists, we can choose more images in order to get three images whose grey values are not zero at same point..

We suppose the imaging resolution ratio is M*N, $i \in [1, M]$,corresponding x coordinate; $j \in [1, N]$,corresponding y coordinate. The three digital images at pixel point (i, j) under three different directions lighting source are described as $I_1(i, j), I_2(i, j), I_3(i, j)$. And the three point lighting source directions are $(p_1, q_1, -1), (p_2, q_2, -1), (p_3, q_3, -1)$. From Eq.(11), the reflectance equations are established.

$$\begin{cases} I_1(u, v) = \dfrac{(u - fp_1)p(i, j) + (v - fq_1)q(i, j) + 1}{\sqrt{p_1^2 + q_1^2 + 1}\sqrt{(fp(i, j))^2 + (fq(i, j))^2 + (up(i, j) + vq(i, j) + 1)^2}} \\ I_2(u, v) = \dfrac{(u - fp_2)p(i, j) + (v - fq_2)q(i, j) + 1}{\sqrt{p_2^2 + q_2^2 + 1}\sqrt{(fp(i, j))^2 + (fq(i, j))^2 + (up(i, j) + vq(i, j) + 1)^2}} \\ I_3(u, v) = \dfrac{(u - fp_3)p(i, j) + (v - fq_3)q(i, j) + 1}{\sqrt{p_3^2 + q_3^2 + 1}\sqrt{(fp(i, j))^2 + (fq(i, j))^2 + (up(i, j) + vq(i, j) + 1)^2}} \end{cases} \quad (12)$$

From the Eq.(12) we have the following formulation

$$\begin{bmatrix} [(c_1 - 1)u - (c_1 p_2 - p_1)f] & [(c_1 - 1)v - (c_1 q_2 - q_1)f] \\ [(c_2 - 1)u - (c_2 p_3 - p_1)f] & [(c_2 - 1)v - (c_2 q_3 - q_1)f] \end{bmatrix} \begin{bmatrix} p(i, j) \\ q(i, j) \end{bmatrix} = \begin{bmatrix} 1 - c_1 \\ 1 - c_2 \end{bmatrix} \quad (13)$$

where $c_1 = \dfrac{I_1(i, j)}{I_2(i, j)} \dfrac{\sqrt{p_1^2 + q_1^2 + 1}}{\sqrt{p_2^2 + q_2^2 + 1}}, c_2 = \dfrac{I_1(i, j)}{I_3(i, j)} \dfrac{\sqrt{p_1^2 + q_1^2 + 1}}{\sqrt{p_3^2 + q_3^2 + 1}}$. The 3-D

surface gradient vector $(p(i, j), q(i, j), -1)$ will be obtained by solving equations.

Actually, object height can be recovered by integrating surface gradient vector or solving functional extreme value which has some constraint conditions. We will choose later method to reconstruct three dimensional surface. $p(x, y), q(x, y), z(x, y)$ are regarded as independent functions and the target function E concluding regularization constraint condition is formulated as

$$E = \iint_D (R(x, y) - E(x, y))^2 + \lambda((p(x, y) - z_x)^2 + (q(x, y) - z_y)^2)dxdy$$

(14)

where D is the integrating area on object surface, .Let

$$F = (R(x, y) - E(x, y))^2 + \lambda((p(x, y) - z_x)^2 - (q(x, y) - z_y)^2)$$ (15)

the corresponding Eulor-Poision equation is

$$F_z - \frac{\partial F_{z_x}}{\partial x} - \frac{\partial F_{z_y}}{\partial y} = 0$$

(16)

The concrete formulation of the Eulor-Poision equation is as

$$z_{xx} + z_{yy} = p_x + q_y$$

(17)

Then, we use discrete difference approximate differential method such as

$$p_x = (p(i+1), j) - p(i-1, j)/2h, q_y = (q(i, j+1) - q(i, j-1))/2h$$

(18)

and

$$z_{xx} = (z(i+1, j) - 2z(i, j) + z(i-1, j))/h, z_{yy}$$
$$= (z(i, j+1) - 2z(i, j) + z(i, j-1))/h$$

(19)

imaging pixel distance h is seen as unit value ($h = 1$), so

$$(z_{i+1,j} + z_{i-1,j} + z_{i,j+1} + z_{i,j-1}) - 4z_{i,j} = \frac{1}{2}(p_{i+1,j} - p_{i-1,j} + q_{i,j+1} - q_{i,j-1})$$

(20)

The corresponding iterative computing formula to compute value of $z_{i,j}$ is

$$z_{i,j}^{(k+1)} = \frac{1}{4}(z_{i+1,j}^{(k)} + z_{i-1,j}^{(k)} + z_{i,j+1}^{(k)} + z_{i,j-1}^{(k)}) - \frac{1}{8}(p_{i+1,j} - p_{i-1,j} + q_{i,j+1} - q_{i,j-1})$$ (21)

where $k \in Z^+$ is the iteration time.

4   Simulation Experiment

Experimental results of both synthetic ball and vase are illustrated. All the experiments are completed under the following conditions: hardware CPU Intel Pentium (R) 3.20GHz, RAM 1.46GB; software Windows XP and Matlab7.0.

4.1   Experimental Result of the Hemisphere Function

The experiment uses Lambertain model to synthesis three hemisphere images. Figure 1. is the synthetic grey image under the lighting directions: (0, 0, -1), (0.03, 0, -1), (0, 0.04, -1) and focal length is set as 300. Pixels of images are 200*200. The imaging grey values are multiplied by 255 and chosen integer. Figure 2(a). is reconstructed object shape by our method, Figure 2(b). is 3-D shape of synthetic hemisphere, Figure 2(c). is the error between synthetic and reconstructed image.

(a)   (b)   (c)
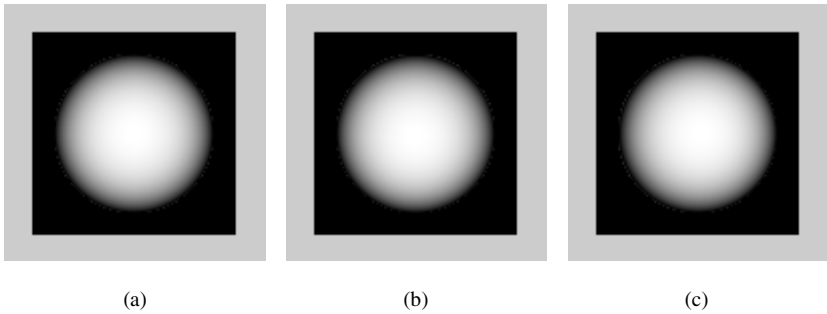
Fig. 1. The synthetic hemisphere images under different lighting directions

(a)   (b)   (c)
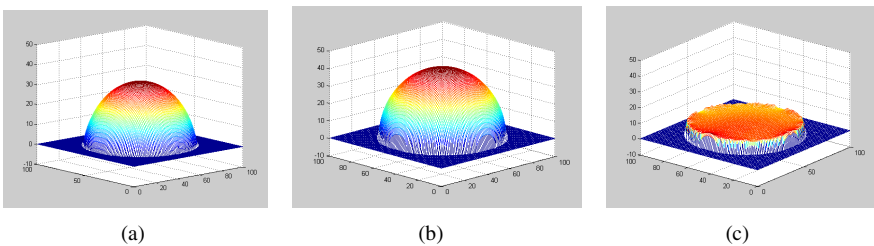
Fig. 2. Reconstruction result of synthetic image of hemisphere

4.2   Experimental Result of the Vase Function

The synthetic image of vase function Eq. (22) is used. Image is formulated by Eq. (11) using the surface functions, where lighting source directions are located at (0, 0, -1),

(0.03, 0, -1), (0, 0.04, -1) and focal length is set as 300. Pixels of images are 200*200. The imaging grey values are multiplied by 255 and chosen integer.

$$z(u,v) = f + 0.15 - \sqrt{(0.15 - 0.1v(6v+1)^2(v-1)^2(3v-2))^2 - u^2} \quad (22)$$

$(u,v) \in [-0.5, 0.5] \times [0,1]$.

Synthetic images of vase function are shown as figure 3(a-c). Figure 4(a). is reconstructed object shape by our method, Figure 4(b). is 3-D shape of synthetic vase, Figure 4(c). is the error between synthetic and reconstructed shape.

(a)   (b)   (c)
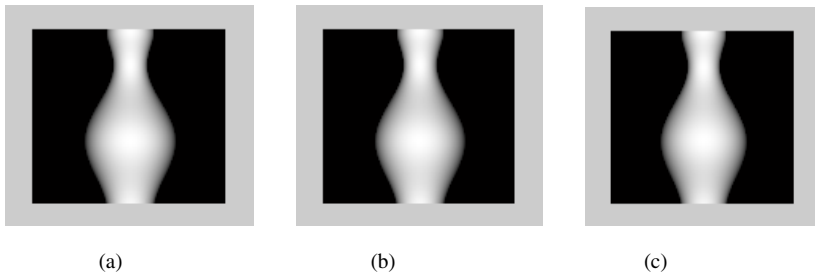
Fig. 3. The synthetic vase images under different lighting directions

(a)   (b)   (c)
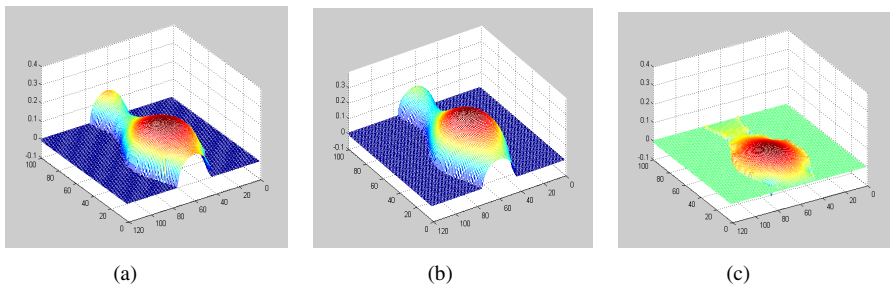
Fig. 4. Reconstruction result of synthetic image of hemisphere

5 Conclusions and Future Work

A new effective SFS method to recover object shape under perspective projection is proposed in this paper. This method can avoid a large number of errors caused by single image recovery efficiently. We just need get three images under different lighting sources. Through calculating three reflectance equations, we obtain the only and certain gradient vector. At last, iterative algorithm is used to recover 3-D object surface height. As we know, beside the numerical error produced by discrete difference

approximation, the error of imaging model is inevitable in SFS. Therefore, less computational complexity and calculating operation time, which demonstrates the convenience of our method. Experimental results illustrate the performance of the proposed method. In the future, our work is to establish more accurate reflected model and theory algorithm. Further more, many constraint factors will be considered in our method, such as highlight and shadow. Our experiment will relate with real object with near-scene point lighting source condition.

References

1. Horn, B.K.P.: Height and gradient from shading. Int. J. Computer Vision 5(1), 37–75 (1990)
2. Woodham, R.J.: Photometric Method for Determining Surface Orientation from Multiple Images. Optical Engineering 19(1), 139–144 (1980)
3. Horn, B.K.P., Brooks, M.J.: The variational approach to shape from shading. Computer Vision Graphics Image Process. 33(2), 174–208 (1986)
4. Lee, K.M., Kuocc, J.: Shape from shading with a linear triangular element surface model. IEEE Trans. Pattern Analysis and Machine Intelligence 15(8), 815–822 (1993)
5. Cho, S.Y., Chow, T.W.S.: A new color 3D SFS methodology using neural-based color reflectance models and iterative recursive method. Neural Computation 14(11), 2751–2789 (2002)
6. Ron, K., James, A.S.: Optical Algorithm for shape from shading and path planning. Journal of Mathematical Imaging and Vision 14(3), 237–244 (2001)
7. Prados, E., Camilli, F., Faugeras, O.: A unifying and rigorous shape from shading method adapted to realistic data and applications. J. Math. Imaging 25(3), 307–328 (2006)
8. Woodham, R.J.: Gradient and Curvature from the Photometric Stereo Method, Including Local Confidence Estimation. J. Optical Soc. Am. 11(11), 3050–3068 (1994)
9. Su, Q., Si, C.: Study on New Algorithm of Shape Reconstruction Based on Multi-images. Aeronautical Computing Technique 4(37), 17–19 (2007)
10. Yang, L., Han, J.Q.: 3-D shape reconstruction of medical images using perspective projection. International Journal of Computer Vision 63(1), 21–43 (2005)
11. Prados, E., Faugeras, O.: A generic and provably convergent shape-from-shading method for orthographic and pinhole cameras. Int. J. Comput. Vis. 65(1), 97–125 (2005)
12. Breuss, M., Cristiani, E., Durou, J.D., Falcone, M., Oliver, V.: Numerical algorithms for perspective shape from shading. Kybernetika 46(2), 207–225 (2010)
13. Zhang, R., Tsai, P.S., Cryer, J.E., Shah, M.: Shape from shading: a survey. IEEE Trans. PAMI 21(8), 690–706 (1999)
14. Yang, L., Ma, S., Tian, B.: New Shape-from-Shading Method with Near-Scene Point Lighting Source Condition. In: Wang, Y., Li, T. (eds.) Foundations of Intelligent Systems. AISC, vol. 122, pp. 653–664. Springer, Heidelberg (2011)

Balancing Methods on the Three-Axis Air-Bearing Platform

Shuai Wang[*], Jie Ma, and Shuang Gao

Control and Simulation Center, Harbin Institute of Technology, Harbin, 150001
wshigh@sina.com

Abstract. In order to improve the performance of the three-axis air-bearing platform (TAP) and to achieve higher simulation accuracy, better balancing methods are required. In this paper, we propose a complete balancing method considering other kinds of disturbance torque as well as the gravity torque. The gravity torque adjustment is divided into three steps: manual adjustment, rough automatic adjustment and precise automatic adjustment. Other kinds of disturbance torque, such as vortex torque, elastic distortion torque and dynamic unbalance torque are discussed. The effectiveness of the proposed method is validated through experiments.

Keywords: Three-axis air-bearing platform, Balancing methods, Disturbance torque.

1   Introduction

Three-axis air-bearing platform (TAP), with air bearing as its key component, is one of the most important equipments for the ground-based simulation of satellites [1]. It is widely used to create a frictionless and torque-free environment. Platform balancing is one of the most important aspects that must be considered before the use of TAP. Traditional balancing relocates the platform's center of mass (CM) into the geometrical center of the spherical air-bearing (CS), which can significantly reduce the external gravity torque and therefore generate an approximately gravity torque free environment. Therefore, under this condition, more accurate performance of TAP requires the consideration of eliminating disturbance torque other than gravity torque. Methods to eliminate other disturbance torque would further improve the TAP performance.

Previously, many researchers have done a lot of studies about balancing methods. Researchers mainly focused on the influence of gravity and ignored other kinds of disturbance torque. However, if we expect the unbalance torque to be close to $10^{-4} N \cdot m$ or less, we have to take the other disturbance torque into account, such as vortex torque, elastic distortion torque, dynamic unbalance torque and so on.

In this paper, we introduce the whole balancing process including gravity torque adjustment and other disturbance torque adjustments in detail.

---

[*] Graduate student.

2   Gravity Torque Adjustment

Before using the TAP, we need to adjust it to the balanced state. The gravity torque adjustment includes manual adjustment, rough automatic adjustment and precise automatic adjustment.

2.1   Manual Adjustment

In most cases, one needs to add or move mass blocks to adjust the balance as the first step. The effect of manual adjustment depends largely on the experience of the workers. This step may take a lot of time for an inexperienced worker. One needs to determine the balance of the TAP by visual inspection. The declination angles of both ends will help identify the level of the horizontal plane of TAP. Only when CM is under CS, the TAP can be stable. CM should be raised gradually until it is above CS. And then lower CM and repeat the above steps to ensure better balance state. Finally, the swing period should be used as an index to measure balancing results.

2.2   Rough Automatic Adjustment

After manual adjustment, the unbalanced gravity torque of a small TAP can be close to $0.01 N \cdot m$, which is still the dominant unbalanced torque. The other disturbance torque can be ignored. .And the TAP can be viewed as a 3-D pendulum. In order to further simplify the analysis, it is often regard as a compound pendulum. Fig. 1 shows the equilibrium state of the ξ-axis. The TAP is in equilibrium when it is in position 1. And the platform is in position 0 when the top surface of platform is horizontal. The angle between position 1 and position 0 is α.

According to the cycle formula of compound pendulum, we can get the offset of CM and CS in the ξ-axis

$$r_{\zeta} = -\frac{4\pi^2 I}{MgT^2 \sqrt{1 + \tan^2 \alpha + \tan^2 \beta}} \quad , \tag{1}$$

where β is the angle in the ζ-axis and M is the mass of the platform.

We can get r_{η} and r_{ζ} in the same method. When the offset between CM and CS is very small, the platform moves in a very low velocity, thus increasing the difficulty in identifying the equilibrium position. But with the help of accelerometers we can find the equilibrium position and the corresponding angles, since the platform is in equilibrium when the value of accelerometer is zero.

However, due to the time-consuming and inaccurate feature of this method, it is mostly used for preliminary estimation of the offset and unbalanced torque.
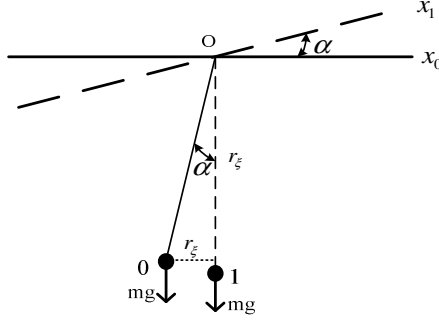
Fig. 1. The equilibrium state of the ξ-axis

2.3   Precise Automatic Adjustment

The compound pendulum model is a simplified one with reduced computation. In order to achieve a higher balancing accuracy, a more accurate model to describe the TAP is needed. So the details of the inverse dynamic model [1] are presented here.

First we establish an inertial coordinate system in the center of rotation and a body-fixed coordinate system on the TAP. Then we have

$$\dot{\omega} = \begin{bmatrix} \dfrac{mg}{I_{\xi\xi}}\left(-r_\eta \cos\phi\cos\theta + r_\zeta \sin\phi\cos\theta\right) \\[2mm] \dfrac{mg}{I_{\eta\eta}}\left(r_\xi \cos\phi\cos\theta + r_\zeta \sin\theta\right) \\[2mm] \dfrac{mg}{I_{\zeta\zeta}}\left(-r_\xi \sin\phi\cos\theta - r_\eta \sin\theta\right) \end{bmatrix} \tag{2}$$

When automatic adjustment begins, motion of the platform is very slow. According to Eq. (2), we have the following equations during the time interval of Δt.

$$\begin{cases} \omega_{\xi t2} - \omega_{\xi t1} = \dfrac{-Mg\Delta t}{2I_{\xi\xi}}\Big[\big((\cos\phi\cos\theta)_{t2} + (\cos\phi\cos\theta)_{t1}\big)r_\eta - \big((\sin\phi\cos\theta)_{t2} + (\sin\phi\cos\theta)_{t1}\big)r_\zeta\Big] \\[2mm] \omega_{\eta t2} - \omega_{\eta t1} = \dfrac{Mg\Delta t}{2I_{\eta\eta}}\Big[\big((\cos\phi\cos\theta)_{t2} + (\cos\phi\cos\theta)_{t1}\big)r_\xi + \big((\sin\theta)_{t2} + (\sin\theta)_{t1}\big)r_\zeta\Big] \\[2mm] \omega_{\zeta t2} - \omega_{\zeta t1} = \dfrac{-Mg\Delta t}{2I_{\zeta\zeta}}\Big[\big((\sin\phi\cos\theta)_{t2} + (\sin\phi\cos\theta)_{t1}\big)r_\xi + \big((\sin\theta)_{t2} + (\sin\theta)_{t1}\big)r_y\Big] \end{cases} \tag{3}$$

We make a simple replacement and Eq. (3) can be rewritten as:

$$\begin{pmatrix} \Delta\omega_\xi \\ \Delta\omega_\eta \\ \Delta\omega_\zeta \end{pmatrix} = \begin{pmatrix} 0 & \lambda_{12} & \lambda_{13} \\ \lambda_{21} & 0 & \lambda_{23} \\ \lambda_{31} & \lambda_{32} & 0 \end{pmatrix} \begin{pmatrix} r_\xi \\ r_\eta \\ r_\zeta \end{pmatrix} \tag{4}$$

Then Eq. (4) can be abbreviated as follows:

$$\Delta\omega = \Phi r. \tag{5}$$

The determinant of Φ is

$$|\Phi| = \lambda_{13}\lambda_{21}\lambda_{32} + \lambda_{12}\lambda_{23}\lambda_{31} = 0. \tag{6}$$

Since it is impossible to solve Eq. (4) directly, we use the least square method to identify the value of r. In order to use this method, we expand Eq. (4) by combining equations at different time as shown in Eq. (7)

$$\begin{pmatrix} (\Delta\omega_\xi)_{t0} \\ (\Delta\omega_\eta)_{t0} \\ (\Delta\omega_\zeta)_{t0} \\ (\Delta\omega_\xi)_{t1} \\ (\Delta\omega_\eta)_{t1} \\ (\Delta\omega_\zeta)_{t1} \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 & (\lambda_{12})_{t0} & (\lambda_{13})_{t0} \\ (\lambda_{21})_{t0} & 0 & (\lambda_{23})_{t0} \\ (\lambda_{31})_{t0} & (\lambda_{32})_{t0} & 0 \\ 0 & (\lambda_{12})_{t1} & (\lambda_{13})_{t1} \\ (\lambda_{21})_{t1} & 0 & (\lambda_{23})_{t1} \\ (\lambda_{31})_{t1} & (\lambda_{32})_{t1} & 0 \\ \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} r_\xi \\ r_\eta \\ r_\zeta \end{pmatrix} \tag{7}$$

Eq. (7) can be written as

$$\Delta\omega_N = \Phi_N r, \tag{8}$$

where N is the number of equations.

According to the least square method, we have the following equation:

$$r = \left(\Phi_N{}^T \Phi_N\right)^{-1} \Phi_N{}^T \Delta\omega_N. \tag{9}$$

This method is more efficient in calculating the bias between CM and CS, and suitable for more cases. On the basis of inverse dynamic model, using flywheel [2] as an excitation will help us get a new model.

3   The Other Disturbance Torque

In the process of adjusting the balance of TAP, various disturbance torque including gravity torque renders TAP unbalanced. However, the main factors influencing the balancing accuracy are changing during the balancing process. At first gravity torque can be the dominant factor. But the reduction of the offset of CM and CS resulting from preliminary balancing would leave gravitational force a less important one, in which case various other disturbance torque such as vortex torque, elastic

distortion torque and dynamic unbalance torque become relatively more important in determining the balancing accuracy. The disturbance torque and the relative eliminating approaches will be discussed in this part.

3.1   Vortex Torque

Externally pressurized spherical air bearings can offer a nearly torque-free environment, close to that of space, and for this reason it is the key component of the TAP. The static characteristic of spherical air bearings and the vortex torque affect the performance of the test-bed directly.

In order to find out various factors that influence vortex torque, we use computational fluid dynamics. And by means of FLUENT software, we establish a three dimensional model and analyze the impact of various parameters on the performance of the air bearing.
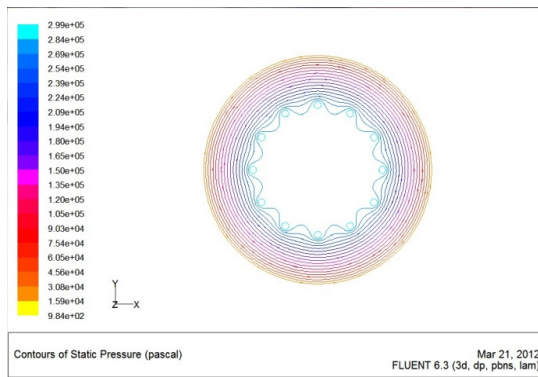
Fig. 2. Three dimensional finite element model of air bearing

After software analysis, we find out that the vortex torque is mainly caused by manufacturing error of air intake, ball head roundness error and ball socket installation error. And the impact of ball socket installation error on vortex torque is the greatest. For a ball bearing with a diameter of $300mm$, the socket installation errors should be limited within $\pm 0.3^{o}$ to ensure the vortex torque no larger than $10^{-5} N \cdot m$.

Improving manufacturing accuracy of the ball bearing can help reduce the vortex torque. But this is not an economical way. For the vortex torque caused by air intake manufacturing errors and ball socket installation errors, the compensating method is taking an independent air supply approach. And for the vortex torque arising from the ball head roundness error, the compensation method is adjusting CM of the test-bed. The compensation method is proved to be effective through simulation.

3.2   Elastic Distortion Torque

Hans F. Kennel pointed out that for a TAP almost 90 percent of the unbalance torque is caused by elastic distortion after gravity torque adjustment and vortex torque adjustment [3]. The platform has elastic deformation when the TAP is tilted, and there is a displacement of CM. And Yanbin Li has given the formula about the unbalance torque caused by elastic distortion

$$M = mg \begin{bmatrix} \dfrac{A_\eta - A_\zeta}{2} \sin 2\phi \cos^2 \theta \\[2mm] \dfrac{A_\xi - A_\zeta}{2} \cos \phi \sin 2\theta \\[2mm] \dfrac{A_\eta - A_\xi}{2} \sin \phi \sin 2\theta \end{bmatrix} \tag{10}$$

where A is the displacement of CM.

To eliminate the elastic distortion torque, Dong Xiang introduces a compensator. The structure of compensator is shown in Fig. 3, where k is spring and G is mass block.
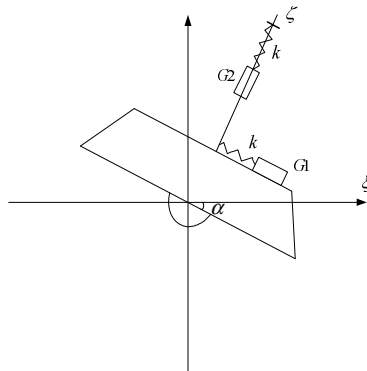
Fig. 3. Compensating system for the elastic deformation unbalance torque

According to Fig 3, the displacement of G2 along the ζ-axis is Δl_1.

$$\Delta l_1 = -\frac{G_1 \cos \alpha - f_1}{k} \tag{11}$$

where f_1 is the friction along ζ-axis, f_2 is the friction along ξ-axis and k is the coefficient of elasticity.

Then the compensation torque caused by G2 is

$$M_1 = G_1 \sin \alpha \cdot \Delta l_1 = -\frac{G_1^2}{2k} \sin 2\alpha + \frac{G_1 f_1}{k} \sin \alpha \tag{12}$$

Same argument, we have

$$M_2 = G_2 \cos\alpha \cdot \Delta l_2 = \frac{G_2^2}{2k}\sin 2\alpha - \frac{G_2 f_2}{k}\cos\alpha \qquad (13)$$

The total compensation torque is

$$M_\eta = \frac{G_2^2 - G_1^2}{2k}\sin 2\alpha - \frac{G_1 f_1 \sin\alpha - G_2 f_2 \cos\alpha}{k} \qquad (14)$$

If the friction is small enough, with some approximation, the second term of Eq. (14) can be negligible.

When 2α is small, we also have the following approximation :

$$\sin(2\alpha) \approx 2\alpha \qquad (15)$$

So if we let

$$\frac{G_2^2 - G_1^2}{k} = k_0 \qquad (16)$$

we get

$$M_\eta = \sin(k_0\alpha) \qquad (17)$$

If G and k are carefully chosen, the elastic distortion torque can be greatly compensated.

We should notice that the error of this compensator increases as the angle increases.

3.3   Dynamic Unbalance Torque

Dynamic unbalance torque is caused by the product of inertia. If the axis of inertia and the rotation axis do not coincide, the dynamic unbalance torque will appear on the TAP. To clarify the influence of product of inertia, we give the TAP's parameters.

The inertia matrix is given by

$$A = \begin{bmatrix} 18.03 & 0.379 & -0.48 \\ 0.379 & 18.59 & 0.126 \\ -0.48 & 0.126 & 26.35 \end{bmatrix} \qquad (18)$$

The initial angle of the ξ-axis is 5.7^o, the initial angle of the η-axis is 0 and the initial angle of the ζ-axis is 0. The CM is $r_\xi = 2\times10^{-6}, r_\eta = 2\times10^{-6}, r_\zeta = -2\times10^{-4}$.

Then the simulation result is shown in Fig. 4,
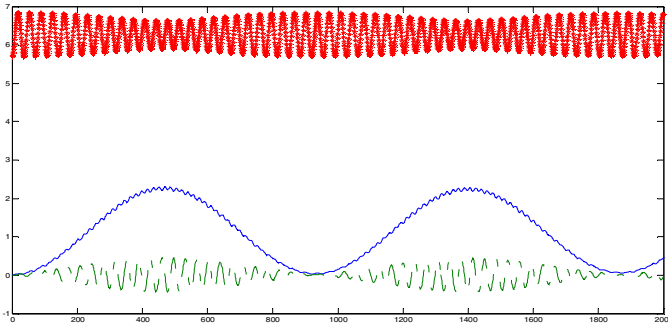
Fig. 4. Changes in the angles under the influence of product of inertia

where " $*$ " line is the angle of ξ -axis, the dotted line is the angle of η -axis and the solid line is the angle of ζ -axis.

According to Fig. 4, we know that product of inertia causes the energy coupling among each axis. In order to eliminate the dynamic unbalance torque, we need to eliminate the product of inertia. In theory, we can eliminate product of inertia by adding or removing two mass blocks. However removing mass from TAP would not be feasible. So William E. Lang introduces an equivalent method [4].

Apart from the major disturbance torque mentioned above, there are also some kinds of disturbance torque: airflow disturbance torque and electromagnetic interference torque. And shielding airflow disturbance and electromagnetic interference is a good method.

4   Experiments

In order to verify the accuracy of the above methods, a lot of experiments have been done. The structure of our TAP is shown in Fig. 5. The experiment system includes a PC104，an ART1020 motion control card, 4 step motors as actuator, a stabilized voltage supply as power unit, IMU as the sensor, a monitor and a TAP.

Firstly, we do the manual adjustment to ensure the CM is under CS, and the swing period is longer than 50s. Secondly, we adjust the balance further according to compound pendulum model and inverse dynamic model. Thirdly, we need to consider of all the disturbance torque, find out the dominant disturbance torque and eliminate it with a suitable method. We repeat the above steps until the unbalance torque is less than the required value.

After only ten adjustments, the unbalance torque becomes smaller than $10^{-3} N \cdot m$ .It is obvious that with the increase of adjustment times, both of the offset between CM and CS and the unbalance torque are getting smaller. After sufficient adjustments, the unbalance torque will be close to the set value.
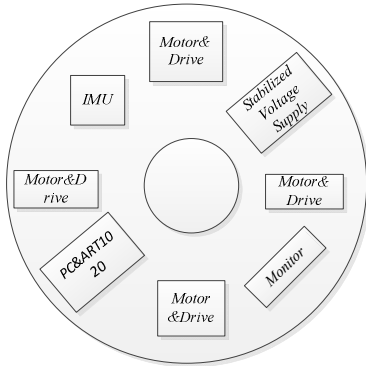
Fig. 5. The structure of our TAP

5   Conclusions

A complete balancing method is established in this paper. We introduce a compound pendulum model and an inverse dynamic model to describe the TAP, and provide a basic balancing method including manual adjustment, rough automatic adjustment and precise automatic adjustment. Then we discuss many kinds of disturbance torque and introduce the corresponding adjustment method. Finally, the effectiveness of the proposed method is validated through experiments.

References

1. Prado, J., Blslacchl, G., Reyes, L., Vlcente, E., Contreras, F., Meslnas, M., Juarez, A.: Three-Axis Air-Bearing Based Platform for Small Attitude Determination and Control Simulation. J. Applied Research and Technology 3, 222–237 (2005)
2. Kim, B., Velenis, E., Kriengsiri, P., Tsiotras, P.: A Spacecraft Simulator for Research and Education. In: Proceedings of the AAS/AIAA Astodynamics Conference. IEEE Press, Canada (2001)
3. Kennel, H. F., Golley, P.T.: Anisoelasticity Torque Compensator Design for Satellite Motion Simulators. Technical report, Marshall Space Flight Center (1962)
4. Lang, W.E.: A Study of Spacecraft Balance with Respect to Multiple Axes. Technical report, Goddard Space Flight Center (1964)

A Non-linearity Correction Method for Calibration of Optical Sensor at Low Level Light

Zilu Wang[1], Bin Wu[1,*], and Tima Sergienko[2]

[1] Beijing Aeronautical Science & Technology Research Institute(BASTR),
COMAC, Beijing 102211, China
`binwu@comac.cc`
[2] Swedish Institute of Space Physics, Kiruna, SE 98128, Sweden

Abstract. This paper describes a methodology developed for calibrating optical detector for light engineering, especially for devices used at low level light, including auroral imager, star sensor, astronomical camera and similar optical instruments. In order to know the physical meaning of optical sensor output, calibration is the first and most important process in a complete analysis of observed data. It is found that optical sensors, like CCDs, are not perfectly linear systems as they were assumed. After bias frame subtraction, the number of ADU counts is not exactly proportional to the number of incident photons. A key component of this paper is non-linearity correction. One of current applications using this method is auroral imager which is used for measuring aurora, high-altitude clouds, and other atmospheric optical objects light intensity, which is the first step to complete an optical object tomography simulation.

Keywords: Calibration, Low Light, Optical Sensor.

1 Introduction

The test data used in this paper is collected by ALIS ground-based optical system which is used for measuring aurora, high-altitude clouds, and other atmospheric optical objects. ALIS has been built in the north part of Scandinavia and consists of a grid of observation with approximately 50km separation and it is highly configurable and many configuration options exist to enable different operational modes [1, 3]. Each imager has a filter wheel with space for up to six narrow band interference filters. One of the main tasks of ALIS is to determine the 3-D spatial structures of aurora. The procedures to determine what physical value each pixel represents, so called camera absolute calibration, is the first and the most important task before starting analysis of ALIS data. This paper presents detailed procedures about this round of absolute calibration by using laboratory data and night sky images. As an example, only testing results from two cameras are listed in this paper.

---

* Corresponding author.

2   Determination of Optical Parameters

It is physically meaningful to know the number of photons which fall into the field-of-view of a pixel and which hit its effective area to produce a count. With knowledge of the number of photons per count, radiance in quantum unit can be derived from raw counts, defined as photons per solid angle per square meter per second, by using formula [2]:

$$L_\gamma = \frac{N}{d\Omega A_L d\lambda t} \quad \left[\frac{photons}{m^2 \cdot sr \cdot s} \right]$$

(1)

Where, N : Number of photons detected by pixel with field-of-view, $d\Omega$: Solid angle of pixel field-of-view, A_L : Pixel effective area, $d\lambda$: Wavelength band of photons, t : Exposure time.

   As the regards of the interpretation in terms of physical process should be convenient, in astronomy, Hunten proposed a photometric measurements of aurora and airglow reported in terms of $4\pi L_\gamma$, rather than radiance L_γ, further it is has been given the name Rayleigh[2], for ALIS absolute calibration, the task is to find the pixel sensitivity value, [Rayleighs]/[Count].

   Further, one of the reasons cause slightly variations of pixel-to-pixel sensitivity is the nonuniform field-of-view and effective area of CCD pixels. Thus, correction should be carried out during the procedure of absolute calibration. Gustavsson described a new mode with a compensate term in radial direction[2], to avoid the radial error for the pin-hole camera model:

$$u = f_u \left((1-\alpha) \cdot \tan(\theta) + \alpha \cdot \theta \right) \cdot \sin(\phi) + u_0$$
$$v = f_v \left((1-\alpha) \cdot \tan(\theta) + \alpha \cdot \theta \right) \cdot \cos(\phi) + v_0$$

(2)

Where: u : Horizontal coordinates in the image plane, v : Vertical coordinates in the image plane, f_u : Horizontal width, f_v : Vertical width, u_0 : Horizontal projection of optical axis, v_0 : Vertical projection of optical axis, θ : Polar angle relate to optical axis. ϕ : Azimuth angle around optical axis, and they are so called optical parameters. Star positions selected in the Yale Brightness Star catalogue (YBS) and identified star positions found in image are used as interest points to calculate optical parameters by using optimization tools.

   With knowledge of optical parameters, the pixel field-of-view is calculated as [2]:

$$d\Omega(u,v) = \sin\theta d\phi d\theta = \sin\theta \left| \frac{\partial(\phi,\theta)}{\partial(u,v)} \right| dudv = \sin\theta \left| \frac{\partial(u,v)}{\partial(\phi,\theta)} \right|^{-1} dudv$$

(3)

Where, Ω is Pixel field-of-view. $\left|\dfrac{\partial(\phi,\theta)}{\partial(u,v)}\right|$ is absolute value of the determinant of

the Jacobian of the optical transfer function. The variation of the effective area with angle θ relative to the optical axis is:

$$A(\theta) = A_L \cos(\theta) \tag{4}$$

Here A_L is the maximum area of the lens of optics, θ is polar angle.

Figure 1 shows identified stars plotted in a night sky image (red dots) and projected positions from a star catalogue (yellow circles) for each camera. Optical parameters are calculated by using an optimization method to minimize the error between the positions of stars in the image and star positions located in a sky map.
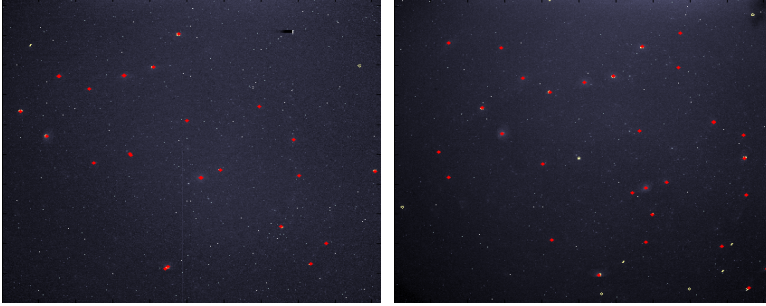
A. Camera2   B.Camera3

Fig. 1. Identified stars in camera field-of-view

Table 1 summarized the optical parameters for 2 ALIS imagers.

Here $(\tilde{u}, \tilde{v})$ are projection positions from the star catalogue and (u', v') are the identified image positions of stars [2]. Optimization methodology is used to find the minimum value of

$$\sum_{i=1}^{N}\left(\left(u'_i - \tilde{u}_i\right)^2 + \left(v'_i - \tilde{v}_i\right)^2\right)^{\frac{1}{2}} \tag{5}$$

Where N is the number of stars identified in an image.

Table 1. Optical parameters for 5 ALIS imagers

Camera	f_u	f_v	alpha	Beta	gamma	dx	dy	Alfa
2	-0.9945	1.0094	-26.680	-7.6313	-5.3789	-0.0042	-0.0507	-0.1788
3	-1.0465	1.0731	-10.691	9.2450	-1.8353	0.1594	-0.1588	1.1761

3   Absolute Calibration

Like astronomical CCD image processing [3], in order to absolute calibrate ALIS CCDs, data reduction should be carried out firstly. The process of removing the CCD signature consists of removing the bias, thermal contribution, non-linearity correction and dividing the resultant image by the flat field in order to standardize the response of each image pixel.

$$Signature\_removed = f\left(Raw - Bias - Overscan\right)\big/Flat \qquad (6)$$

The bias frame, over-scan-strip and nonlinearity should be removed firstly from the raw data, and then divide the data by flat field. See function, where $f\left(\cdot\right)$ is the non-linearity correction function.

For practical work, the linearity is checked by varying the integration time and the CCDs are illuminated by a stable standard laboratory source. For laboratory data, there are four different exposures for each operational mode. ALIS cameras are equipped with shutters that open and close rapidly, in the order of magnitude of 10ms. For laboratory data taken from Sept. 2007, exposure times varying from 2s to 200s, the relative error in pixel exposures changes from 0.5% to 0.005%. Usually we consider the accuracy is good enough if exposure times lager than 2s. For ALIS imagers linearity testing, images identified by 'Filt'=5, illuminated by source IRF, UJO 920B (4278 Å), are selected for a linearity test because the error caused by shutter characteristics can be ignored by its longest exposure times $\left(t_{int} > 2.5s\right)$.

For a camera with complete linearity, the ideal ratio of counts from same pixel between two images should equal the ratio of exposure time.

a. Ratio map   b. Line plot of ratio map at row 100
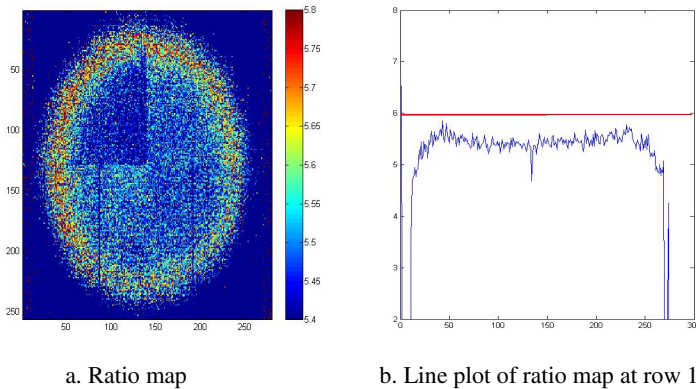
Fig. 2. The linearity checking by using exposure time 15s and 2.5s images with IRF, UJO 920B source for camera 2. The ratio map of camera 2 with average value 5.6035 (for central pixels), error between with idea ratio is 6.6%. Laboratory calibration data are collected from Sept. 2007.

However, it is clear to see from Figure 2, for the camera 2, linearity of the central pixels is worse, caused by its higher intensity than those at the boundary. Thus a non-linearity correction method is desired to improve the linearity.

A correction function should fulfill the following requirements:

$$DN'_{(int\_1)i,j} = f\left(DN_{(int\_1)i,j}\right)$$

$$DN'_{(int\_2)i,j} = f\left(DN_{(int\_2)i,j}\right) \quad \text{With constraints,}$$

$$DN'_{(int\_3)i,j} = f\left(DN_{(int\_3)i,j}\right)$$

$$DN'_{(int\_4)i,j} = f\left(DN_{(int\_4)i,j}\right)$$

$$\frac{DN'_{(int\_4)i,j}}{DN'_{(int\_1)i,j}} = \frac{int\_4}{int\_1}, \frac{DN'_{(int\_3)i,j}}{DN'_{(int\_1)i,j}} = \frac{int\_3}{int\_1}$$

$$\frac{DN'_{(int\_2)i,j}}{DN'_{(int\_1)i,j}} = \frac{int\_2}{int\_1}, \frac{DN'_{(int\_3)i,j}}{DN'_{(int\_2)i,j}} = \frac{int\_3}{int\_2} \tag{7}$$

$$\frac{DN'_{(int\_4)i,j}}{DN'_{(int\_2)i,j}} = \frac{int\_4}{int\_2}, \frac{DN'_{(int\_4)i,j}}{DN'_{(int\_3)i,j}} = \frac{int\_4}{int\_3}$$

Where,

$DN'_{(int\_k)i,j}$: Corrected digital counts for pixel in image coordinates at i row, j column with integration times (exposure times) k.

$DN_{(int\_k)i,j}$: Raw digital counts for pixel in image coordinates at i row, j column with integration times (exposure times) k.

After a set of functions were tested and evaluated by using laboratory data, we found that quadratic functions are the most reasonable in the following form. And it should be unique and independent to the exposure times,

$$DN'_{(int\_1)i,j} = DN_{(int\_1)i,j} \cdot \left(1 + \left(b \cdot DN_{(int\_1)i,j} + c\right)\right)$$

$$DN'_{(int\_2)i,j} = DN_{(int\_2)i,j} \cdot \left(1 + \left(b \cdot DN_{(int\_2)i,j} + c\right)\right)$$

$$DN'_{(int\_3)i,j} = DN_{(int\_3)i,j} \cdot \left(1 + \left(b \cdot DN_{(int\_3)i,j} + c\right)\right) \tag{8}$$

$$DN'_{(int\_4)i,j} = DN_{(int\_4)i,j} \cdot \left(1 + \left(b \cdot DN_{(int\_4)i,j} + c\right)\right)$$

Optimization tools which best minimizes the value of objective function is selected to find the b, c parameters for the correction function. Note that, the notation $s_{k_1 k_2}$ means square root of sum of squared residuals between real ratio and ideal ratio. With the knowledge b, c should be very close to zero, i.e. a minimum solution should exist around the point (0, 0, 0). Thus, a Trust Region algorithm is selected for this task because it can converge to the nearest optimized value from an initial point very quickly.

$$s_{k_1 k_2} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{DN'_{(\text{int}\_k_1)i,j}}{DN'_{(\text{int}\_k_2)i,j}} - \frac{\text{int}\_k_1}{\text{int}\_k_2} \right)^2} \tag{9}$$

Then, six groups of parameters are calculated by minimizing $s_{k_1 k_2}$, $s_{k_1 k_3}$, $s_{k_1 k_4}$, $s_{k_2 k_3}$, $s_{k_2 k_4}$ and $s_{k_3 k_4}$ respectively. Calculation results from laboratory data of camera 2, gain 20, bin 4×4. Fitting curves are shown in Figure 3.
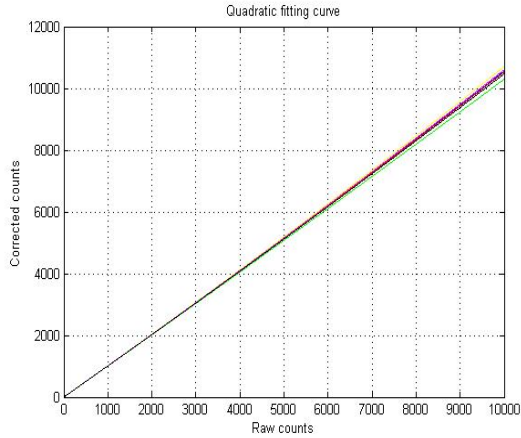
Fig. 3. Curve fitting of non-linearity correction function by using quadratic function, Legend: curve for $S_{k_1 k_2}$ is 'green', $S_{k_1 k_2}$ 'red', $S_{k_1 k_2}$ 'yellow', $S_{k_1 k_2}$,'pink', $S_{k_1 k_2}$,'black' and $S_{k_1 k_2}$, 'purple'
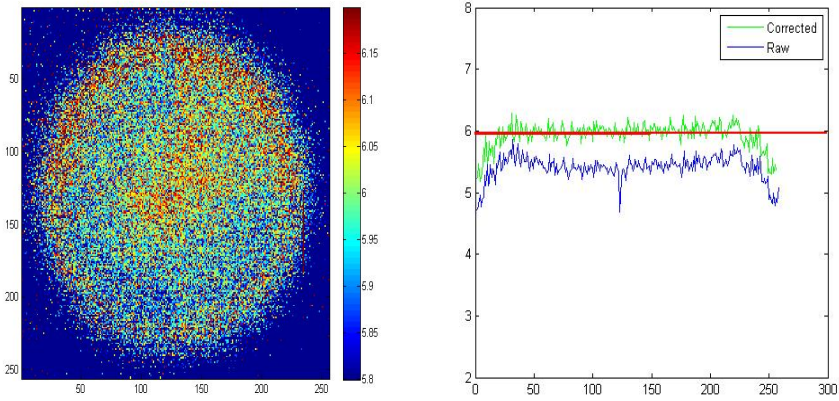
A. Ratio map of camera 2 after correction B. Line plot of ratio map

Fig. 4. The ratio of counts between exposure times 15s and 2.5s with IRF, UJO 920B source (after linearity correction).A. the corrected ratio map of camera 2 with average value (for central pixels) 6.0034, error between with idea ratio is 0.0006%. B. a line plot of the corrected ratio map at row 100 for camera 2.

One hundred and eighty ($5\,imagers \times 4\,quadrants \times 3$gain modes$\times 3$bin modes) groups of b and c have been calculated (see Appendix B). Correction results of camera 2 and camera 5(in gain 20, bin 4×4) is shown in Figure 4. It can be clearly seen that the error decreases rapidly after non-linearity correction.

Currently, no suitable flat field correction has been done for ALIS imagers, although several attempts have been tried by using cloudy sky, white screen and diffuse lamp cover, etc. However, none of them obtained a sufficient quality flat field image [3].

Instead, for ALIS imagers calibration, the variation caused by different between solid angle between pixels has been modeled according to the following procedure:

1. Calculate the product value of field-of-view and effective area for each pixel [2].
2. Find the maximum product value in the pixel matrix.
3. Normalizing correction matrix by dividing the maximum product value. (flat field matrix)

Inter-calibration is carried out to provide a comparable scale of intensities of aurora or airglow emission. Three radioactive ^{14}C light-standards are inter-calibrated against other light standards during European calibration workshops held at regular intervals. Results of the most resent ^{14}C-sources that are used for ALIS calibration (phosphors: 920B, Lauche lamp and Y275) (*Lauche and Widell, 2000, Brändström*, 2003 *and Henricson, 2008*)[4, 5] and the column emission rate is plotted against wavelength in Figure 5.
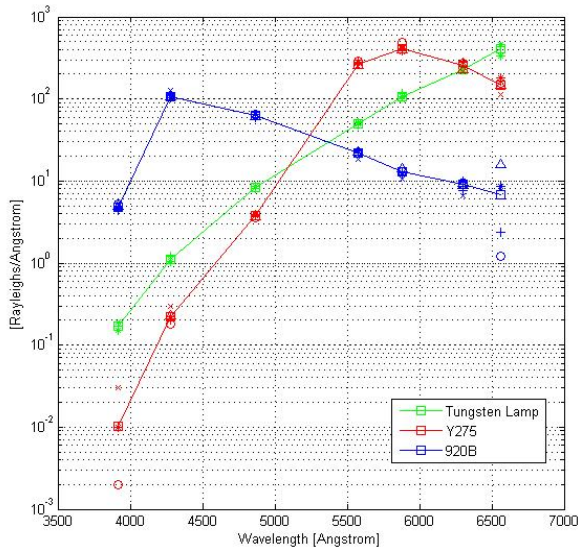
Fig. 5. Column emission rates [R/Å] against wavelength [Å] for three radioactive ^{14}C light-standards with phosphor: 920B (blue), Lauche Lamp (green), and Y275 (red). Legend: intercalibration sessions of 1981 'x' 1985'+', 1999'*', 2000 'o', 2006'^', and 2007'□'. Lines connect the results from the latest calibration workshop, held in Andoya Rocket Range, Norway, 2007.

Because the light-emitting surface of the light source does not cover the entire field-of-view of the imager, only the pixels located in the central part are used to find $\overline{DN^{'}}$. [3]

$$C_{abs\lambda c} = \frac{I_{cal} \cdot t_{int}}{DN^{'}} \left[\frac{Rayleighs}{Count} \right] \tag{10}$$

Where, $C_{abs\lambda c}$ indicates an absolute calibration constant. The column emission rate, I_{cal}, which is obtained by integral of light standards line over the filter pass band $\Delta\lambda = \lambda_2 - \lambda_1$ (for blue source, 920B, $\Delta\lambda = 50$ Å, for red and green source, Lauche and Y275, $\Delta\lambda = 40$ Å), is given by[3],

$$I_{cal} = \int_{\lambda 1}^{\lambda 2} I_{ls}(\lambda) d\lambda [R] \tag{11}$$

And $DN_{i,j}^{'}$ is calculated by [3],

$$DN_{i,j}^{'} = \frac{f\left(DN_{i,j} - B_{\overline{z}i,j} - B_{overj}\right)}{F_{i,j}} [Counts] \tag{12}$$

Where $DN_{i,j}$ is raw digit output of pixel at row i, column j, $F_{i,j}$ indicates flat field, and $f(\cdot)$ is non-linearity correction function. Calibration results for camera 2, bin 4by4 mode are included in Table 2

Table 2. Calibration results for Camera2 in 2011, by using laboratory data from Sept.2007

Camera2, 4278 Å, Bin: 4×4				Pixels selected
gain	13	20	31	
lamp	920B	920B	920B	X:105:135
2.5s	2.49	2.48	2.50	Y:120:150
5s	2.47	2.46	2.51	
10s	2.49	2.48	2.50	
15s	2.46	2.46	2.48	
Average	2.48	2.47	2.50	2.48

4   Conclusions

This paper presents detailed absolute calibration procedures for low light detector includes a key component of non-linearity correction. Former absolute calibration for ALIS assumed 6 imagers with complete linearity. However, after a linearity test for

each camera by using laboratory data in Sept. 2007, it was found that cameras exhibit non-linearity as pixel intensity over a certain value, especially for camera 2. Thus, non-linearity correction is introduced in the camera signature removing procedure.

References

1. Steen, A., Brändström, U.: A multi-station ground-based imaging system at high latitudes. STEP Int. Newsl. 3, 11–14 (1993)
2. Gustavsson, B.: Three Dimensional Imaging of Aurora and Airglow, PhD thesis (2000)
3. Brändström, U.: IRF Scientific Report 179 (2003)
4. Lauche, H., Widell, O.: Intercalibration of low light level sources. In: Proc. of 27th Annual European Meeting on Atmospheric Studies by Optical Methods. Meteorological institution, Stockholm university, Stockholm, Sweden (2000b)
5. Widell, O., Henricson, H.: Intercalibration of low light level sources. In: Proc. of 28th Annual European Meeting on Atmospheric Studies by Optical Methods, Oulu, Finland, Sodankyla Geophysical Observatory (2001)

Analyzing Effects of Ankle-Foot Parameters on Passive Bipeds Based on Dynamic Walking Modeling

Jingeng Mai[1,2], Yue Gao[2], Yan Huang[2], Qining Wang[2], and Lin Zhang[1]

[1] School of Automation Science and Electrical Engineering,
Beihang University, Beijing 100191, China
[2] Intelligent Control Laboratory, College of Engineering,
Peking University, Beijing 100871, China

Abstract. Though people's usual gaits tend to be natural and simple, the theoretical modeling and analysis are complicated based on the remarkable fact that the walking motion is a complex dynamic phenomenon. In this paper, we build a passive dynamic bipedal walking model with flat feet and compliant ankle joints. The bipedal walker travels on a slope actuated by the gravity. We analyze effects of ankle-foot parameters on walking characteristics based on dynamic walking modeling. Simulation results demonstrate that the model can perform stable walking cycle. The effects of ankle-foot parameters on motion characteristics with different ankle stiffness are shown in detail.

Keywords: Passive dynamic walking, bipedal walking modeling, ankle stiffness, ankle-foot parameters.

1 Introduction

Different from nonhuman primates, human beings can achieve stable and efficient dynamic bipedal walking on various different terrains without much effort. Since bipedal walkers with flat feet are more close to human beings in the structure and walking gaits, increasing studies of bipedal walking have focused on the effects of flat feet and ankle stiffness on the walking performance. Kwan *et al.* investigated the influence of introducing flat feet to bipedal walking [1]. The effects of foot geometric parameters are studied. Wang *et al.* found that there is an optimal foot-ankle ratio of the passivity-based bipedal walker [2]. Tlalolini *et al.* indicated the walking performance of different gaits of flat-foot bipeds. In addition, the significance of ankle stiffness is also explored by lots of studies [3]. Ker *et al.* showed that when the compliant ankle is actuated, extra energy can be inserted in the toe-off phase [4]. Hobbelen *et al.* studied the influence of ankle stiffness on walking velocity and energy efficiency both in simulation and in the physical prototype [5]. Several literatures reported that ankle stiffness plays an important role in gait selection and gait transition of the bipedal walking model with flat feet [6,7]. Investigation on bipedal walking with flat feet and ankle

stiffness can help us better understand the principle of real human walking and build advanced bipedal walking robots.

Most industrial bipedal robots were realized based on the trajectory-control approach. By controlling joint angles precisely, the zero moment point (ZMP) is kept within the convex hull of the supporting area [8]. The actively controlled robots can interact with human beings and complete different tasks. However, this kind of bipedal robots often walk with low energetic efficiency and unnatural gaits. Different from the actively controlled walking, passivity-based dynamic bipedal walking may not achieve equilibrium at each moment during motion, but can realize stable cyclic locomotion. As an example, passive dynamic walking [9] has been presented as a possible explanation for the efficiency of the human gait, which showed that a mechanism with two legs can be constructed so as to descend a gentle slope with no actuation and no active control. Compared with the actively controlled walking, passivity-based walking achieves higher efficiency and performs more natural walking gait, which shows a remarkable resemblance to human normal walking [10]. Recently, several studies focused on passivity-based bipedal walking with flat feet [5,11,12]. However, these works did not study the effects of foot geometric parameters with various ankle stiffness.

To study the optimal foot-ankle ratio of the passivity-based walker with adaptable ankle stiffness, in this study, we propose a passive dynamic walking model with flat feet and compliant ankle joints. The ankle stiffness is modeled as a torsional spring. The effects of foot geometric parameters and ankle stiffness on walking performance are studied. The results can be used to explore further understanding of bipedal walking, and help the design of future passivity-based robot prototypes towards more practical uses.

The rest of the paper is organized as follows. The detailed modeling of the walker is described in Section 2. In Section 3, we describe the simulation method and environment. Section 4 shows the experimental results and the relevant discussion. We conclude in Section 5.

2   Model

2.1   Bipedal Walking Model with Flat Feet and Ankle Stiffness

In this study, we propose a passive dynamic bipedal walking model with flat feet and compliant ankle joints. To guarantee lateral stability, both the legs are constructed in pairs. The two-dimensional model consists of two pairs of rigid legs interconnected individually through a hinge at the hip. The foot is mounted on the ankle with a torsional spring. The mass of each stick is averagely distributed, thus the center of mass (CoM) of each part is at the middle of the corresponding stick. The model travels on a slope actuated by gravity. Fig. 1 shows the bipedal walking model. To simplify the motion, we have several assumptions, including: 1) each stick suffers no flexible deformation; 2) there is no damping or friction in the hip joint and ankle joints; 3) the model is considered to fall down if the hip joint is lower than a critical position, then the motion is stop.
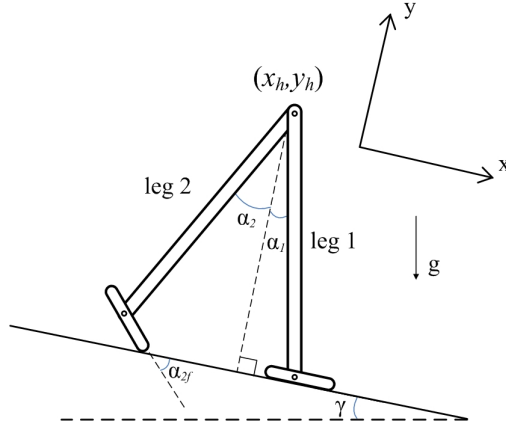
Fig. 1. Passive dynamic bipedal walking model with flat feet and compliant ankle joints

We suppose that the x-axis is along the ground while the y-axis is vertical to the ground upwards. The configuration of the walker is defined by the coordinates of the hip joint and several angles, including the angles between the y-axis and each leg and the angles between the x-axis and each foot (see Fig. 1 for details), which can be arranged in a generalized vector $q = (x_h, y_h, \theta_1, \theta_2, \theta_{1f}, \theta_{2f})'$. The superscript $'$ means the transposed matrix (the same in the following paragraphs). The positive directions of all the angles are counter-clockwise.

2.2 Walking Dynamics

The model can be defined by the rectangular coordinates r, which can be described by the x-coordinate and y-coordinate of the CoM of each stick. The walker can also be described by the generalized coordinates q as mentioned before (suppose leg 1 is the stance leg):

$$q = (x_h, y_h, \theta_1, \theta_2, \theta_{1f}, \theta_{2f})' \tag{1}$$

We define matrix J as follows:

$$J = dr/dq \tag{2}$$

Thus J transfers the independent generalized coordinates $\dot{q}$ into the velocities of the rectangular coordinates $\dot{r}$.

The mass matrix in rectangular coordinate r is defined as:

$$M = diag(m_l, m_l, m_l, m_l, m_f, m_f, m_f, m_f), \tag{3}$$

where m_l is the leg mass, while m_f is the foot mass.

We denote F as the active external force vector in rectangular coordinates. The constraint function is marked as $\xi(q)$, which is used to maintain foot contact

with ground. Each component of $\xi(q)$ should keep zero to satisfy the contact condition.

We can obtain the EoM by Lagrange's equation of the first kind:

$$M_q \ddot{q} = F_q + \Phi' F_c \tag{4}$$

$$\xi(q) = 0 \tag{5}$$

where $\Phi = \frac{\partial \xi}{\partial q}$. F_c is the generalized constraint force vector. M_q is the mass matrix in the generalized coordinates q:

$$M_q = J' M J \tag{6}$$

F_q is the active external force in the generalized coordinates:

$$F_q = J' F - J' M \frac{\partial J}{\partial q} \dot{q} \dot{q} \tag{7}$$

Equation (5) can be transformed to the followed equation:

$$\Phi \ddot{q} = -\frac{\partial(\Phi \dot{q})}{\partial q} \dot{q} \tag{8}$$

Then the EoM in matrix format can be obtained from Equation (4) and Equation (8):

$$\begin{bmatrix} M_q & -\Phi' \\ \Phi & 0 \end{bmatrix} \begin{bmatrix} \ddot{q} \\ F_c \end{bmatrix} = \begin{bmatrix} F_q \\ -\frac{\partial(\Phi \dot{q})}{\partial q} \dot{q} \end{bmatrix} \tag{9}$$

The equation of strike moment can be obtained by integration of Equation (4):

$$M_q \dot{q}^+ = M_q \dot{q}^- + \Phi' I_c \tag{10}$$

where $\dot{q}^+$ and $\dot{q}^-$ are the velocities of generalized coordinates just after and just before the strike, respectively. Here, I_c is the impulsive acted on the walker which is defined as follows:

$$I_c = \lim_{t^- \to t^+} \int_{t^-}^{t^+} F_c dt \tag{11}$$

Since the strike is modeled as a fully inelastic impact, the walker satisfies the constraint function $\xi(q)$. Thus the motion is constrained by the followed equation after the strike:

$$\frac{\partial \xi}{\partial q} \dot{q}^+ = 0 \tag{12}$$

Then the equation of strike in matrix format can be derived from Equation (10) and Equation (12):

$$\begin{bmatrix} M_q & -\Phi' \\ \Phi & 0 \end{bmatrix} \begin{bmatrix} \dot{q}^+ \\ I_c \end{bmatrix} = \begin{bmatrix} M_q \dot{q}^- \\ 0 \end{bmatrix} \tag{13}$$

3   System Simulations

All simulations and data processing were performed using Automatic Dynamic Analysis of Mechanical Systems (ADAMS). Based on the EoMs mentioned above, we analyzed the walking motion of the proposed bipedal model. The walking characteristics of the passive walking model with flat feet and ankle stiffness are studied in the simulation experiments to reveal the effects of foot-ankle ratio and ankle joint stiffness. Fig. 2 shows the model build in ADAMS. The coefficient of static friction and coefficient of sliding friction between the feet and the ground are set to be 0.5 and 0.4, respectively. The parameter values used in the analysis are obtained from Table 1. In this study, foot ratio indicates the ratio of the distance from heel to ankle joint to the distance from ankle joint to toe.

Passive bipedal walking with flat feet is more complicated than that with round feet or point feet [7]. When the flat foot strikes the ground, there are two impulses, "heel-strike" and " foot-strike", representative of the initial impact of the heel and the following impact as the whole foot strikes the ground [2].

Stable periodic walking of the proposed model is found with proper parameters. Fig. 3 shows a representative stable walking cycle. The passive walker successfully performs periodic locomotion. The ankle stiffness is $35Nm/rad$, and the foot ratio is 1.

Fig. 4 shows the leg trajectory of the bipedal walking model with flat feet and compliant ankle joints. These simulations show that the proposed model can achieve stable passive dynamic walking.
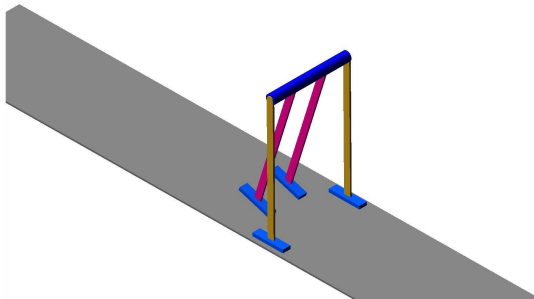
Fig. 2. Passive dynamic bipedal walking model with flat feet and compliant ankle joints

Table 1. Parameter values in simulations

Parameter	Value
leg mass	$4.0kg$
foot mass	$1.0kg$
leg length	$0.8m$
foot length	$0.20m$
slope angle	$0.005rad$
gravitational acceleration	$9.81m/s^{-2}$
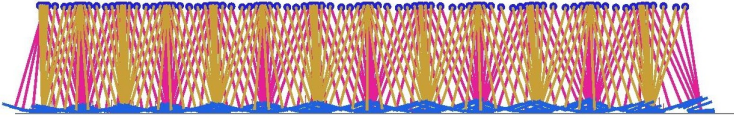
Fig. 3. Stable periodic locomotion of the passive dynamic model with flat feet and compliant ankle joint. The stick diagram is obtained every 9 frames during a continuous walking.
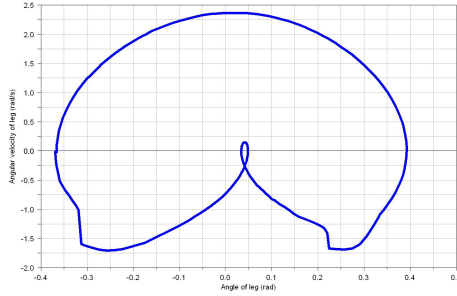
Fig. 4. Leg trajectory of passive dynamic walking with flat feet and compliant ankle joints. Stable steps form the same cycles on the plane.

4   Results

Previous studies revealed that an optimal foot-ankle ratio exists in passivity-based bipedal walking with flat feet [1,12]. However, in these studies, the ankle joint is completely rigid or with constant compliance. In this section, we focus on the optimal foot ratio in terms of walking velocity and step length with different ankle stiffness.

Energy efficiency is an important gait characteristic of passive dynamic walking. Since the cost of transport (energy per unit distance, per unit body weight) is constant for a given slope, efficiency is characterized by velocity in this study, which is similar to [1]. Thus walking velocity and step length are used as the measurements of the proposed model. In the experiments of this subsection, the effects of foot ratio on walking velocity and step length are studied under three ankle stiffness values: $20Nm/rad$, $35Nm/rad$ and $50Nm/rad$. Fig. 5 shows the motion characteristics with various foot ratios at the relatively low ankle stiffness (the spring constant at the ankle is $20Nm/rad$). The results indicate that the bipedal model achieves the largest walking velocity when the foot ratio is 0.4, similar to the foot structure of human beings. The largest velocity is $0.39m/s$. In general, the step length decreases with increasing foot ratio. An exceptional case appears when the foot ratio is 0.1.
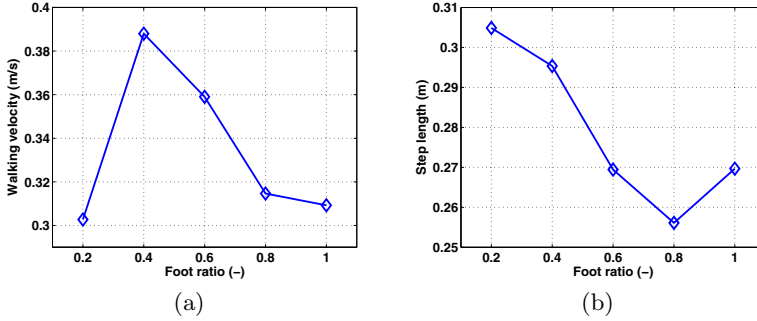
Fig. 5. The effects of foot ration on walking performance of the proposed bipedal walking model with flat feet and compliant ankle joints. The ankle spring constant is $20Nm/rad$. (a): the relationship between foot ratio and walking velocity. (b): the relationship between foot ration and step length.

For the moderate ankle stiffness (the constant of the ankle torsional spring is $35Nm/rad$), the largest walking velocity is also obtained when the foot ratio is 0.4, as shown in Fig. 6 (a). In this case, the model can travel on the slope at 0.45m/s. When the foot ratio is less than 0.4, the velocity is quite small, while the walking speed is close to the largest value with foot ratio larger than 0.4. The step increases monotonously with increasing foot ratio (see Fig. 6 (b)).

When the ankle stiffness increases to a relatively large value (the constant of the ankle torsional spring is $50Nm/rad$), the optimal foot ratio has an offset. The optimal foot ratio changes to 0.8 for both walking velocity and step length (see Fig. 7). The walking velocity of the proposed model can reach to $0.58m/s$ with the optimal foot ratio.
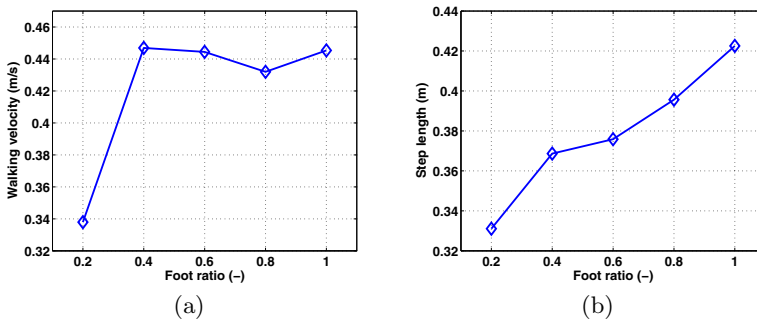
Fig. 6. The effects of foot ratio on walking performance of the proposed bipedal walking model with flat feet and compliant ankle joints. The ankle spring constant is $35Nm/rad$. (a): the relationship between foot ratio and walking velocity. (b): the relationship between foot ration and step length.

From the results above, one can find that larger ankle stiffness results in faster walking with larger step length. In passive dynamic walking, ankle stiffness performs as the tendinous tissues of the skeletal muscle in real human walking [13]. Larger ankle stiffness means stronger push-off effect and more energy stored and released by the elastic elements in one step, which results in larger walking velocity and step length.

For the walking velocity, the effects of foot ratio are different with different ankle stiffness. The optimal foot ratio is similar to that of human beings at the low and the moderate ankle stiffness. This result is also a possible explanation for the evolution of human structure. The optimal foot ratio of walking with high ankle stiffness changes to 0.8. Thus ankle stiffness influences the optimal foot geometric parameters of passive dynamic walking.

(a)   (b)
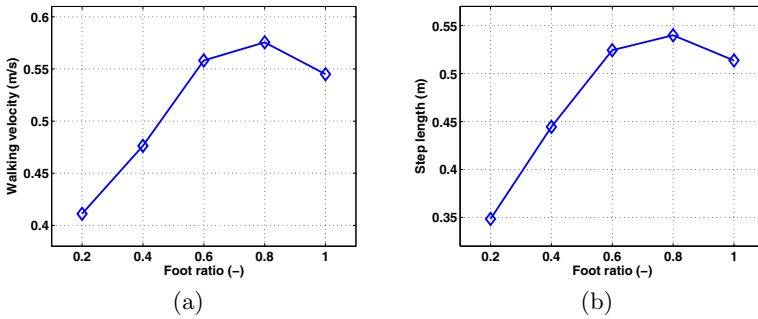
Fig. 7. The effects of foot ration on walking performance of the proposed bipedal walking model with flat feet and compliant ankle joints. The ankle spring constant is $50Nm/rad$. (a): the relationship between foot ratio and walking velocity. (b): the relationship between foot ration and step length.

5   Conclusion

In this study, we have proposed a passive dynamic walking biped with flat feet and compliant ankle joints. The bipedal model achieves stable periodic walking on a slope with a small angle. In simulation, we studied the effects of foot geometric parameters on walking velocity and step length for variable ankle stiffness. The results show that the optimal foot ratio is different for the walkers with different ankle stiffness.

There are several ways to extend this work. More foot structure parameters (for example, foot length) can be studied in the future. Adding compliant knee joints and the upper body will make the model more similar to human beings. Study on dynamic walking on the level ground with actuation is also meaningful improvement.

References

1. Kwan, M., Hubbard, M.: Optimal foot shape for a passive dynamic biped. J. Theor. Biol. 248, 331–339 (2007)
2. Wang, Q., Huang, Y., Wang, L.: Passive dynamic walking with flat feet and ankle compliance. Robotica 28(3), 413–425 (2010)
3. Tlalolini, D., Chevallereau, C., Aoustin, Y.: Comparison of different gaits with rotation of the feet for a planar biped. Robot. Auton. Syst. 57, 371–383 (2008)
4. Ker, R.F., Alexander, R.M., Bennett, M.B.: Why are mammalian tendons so thick? J. Zool. London 216, 309–324 (1988)
5. Hoboelen, D.G.E., Wisse, M.: Ankle actuation for limit cycle walkers. Int. J. Robot. Res. 27(6), 709–735 (2008)
6. Owaki, D., Osuka, K., Ishiguro, A.: Gait transition between passive dynamic walking and running by changing the body elasticity. In: Proc. of the SICE Annual Conference, pp. 2513–2518 (2008)
7. Huang, Y., Wang, Q., Chen, B., Xie, G., Wang, L.: Modeling and gait selection of passivity-based seven-link bipeds with dynamic series of walking phases. Robotica 30, 39–51 (2012)
8. Vukobratovic, M., Frank, A., Juricic, D.: On the stability of biped locomotion. IEEE Trans. Biomedical Eng. 17(1), 25–36 (1970)
9. McGeer, T.: Passive dynamic walking. Int. J. Robot. Res. 9, 68–82 (1990)
10. Collins, S., Ruina, A., Tedrake, R., Wisse, M.: Efficient bipedal robots based on passive-dynamic walkers. Science 307, 1082–1085 (2005)
11. Wisse, M., Feliksdal, G., van Frankenhuyzen, J., Moyer, B.: Passive based walking robot - Denise, a simple, efficient, and lightweight biped. IEEE Robot. Autom. Mag. 14(2), 52–62 (2007)
12. Wang, Q., Huang, Y., Zhu, J., Wang, L., Lv, D.: Effects of foot shape on energetic efficiency and dynamic stability of passive dynamic biped with upper body. Int. J. Hum. Robot. 7(2), 295–313 (2010)
13. Ishikawa, M., Komi, P.V., Grey, M.J., Lepola, V., Bruggemann, G.: Muscle-tendon interaction and elastic energy usage in human walking. J. Appl. Physiology 99, 603–608 (2005)

Investigation of Eddy Diffusivity in a Reactive Plane Jet by Using Direct Numerical Simulation

Tomoaki Watanabe[1], Yasuhiko Sakai[1], Kouji Nagata[1], Osamu Terashima[1], Yasumasa Ito[1], Hiroki Suzuki[2], and Toshiyuki Hayase[3]

[1] Nagoya University, Nagoya, Japan
[2] Nagoya Institute of Technology, Nagoya, Japan
[3] Tohoku University, Sendai, Japan
watanabe.tomoaki@c.nagoya-u.jp

Abstract. Eddy diffusivity in a planar turbulent jet with a second-order chemical reaction ($A + B \rightarrow R$) is investigated by using direct numerical simulation. Reactant A and B are premixed into the jet flow and the main flow, respectively. The eddy diffusivity of species A on the jet centerline is calculated from the mean concentration of species A and the streamwise turbulent mass flux of the species A. The results show that the chemical reaction makes the eddy diffusivity of species A small, and the effects of the chemical reaction on the eddy diffusivity depend on the Damköhler number. From these results, we can conclude that the gradient diffusion model should be used with considerations of the effects of chemical reactions on eddy diffusivity.

Keywords: Direct Numerical Simulation, Eddy Diffusivity, Chemical Reaction, Turbulent Flow.

1 Introduction

Turbulent flows with chemical reactions can be seen in various fields. Combustions in engines and formations of pollutants in atmosphere are typical examples of turbulent reactive flows. Therefore, numerical methods to predict turbulent reactive flows are widely needed. Direct numerical simulation (DNS) provides the most precise predictions of turbulent reactive flows because DNS requires no turbulent models. However, because of high computational costs of DNS, DNS can be only applied to simple flow geometries and simple chemical reactions, and it is not feasible to apply DNS to reactive flows with complex chemical reactions at high Reynolds number. Reynolds-averaged approach is practically used for reactive flows. In this approach, models for a turbulent mass flux and chemical source term are required to solve Reynolds-averaged scalar transport equations. One of the simplest models for the turbulent mass flux $\langle u_i \gamma_\alpha \rangle$ is a gradient diffusion model [1]. The gradient diffusion model gives $\langle u_i \gamma_\alpha \rangle$ by

$$\langle u_i \gamma_\alpha \rangle = -D_{t\alpha,i} \frac{\partial \langle \Gamma_\alpha \rangle}{\partial x_i}. \tag{1}$$
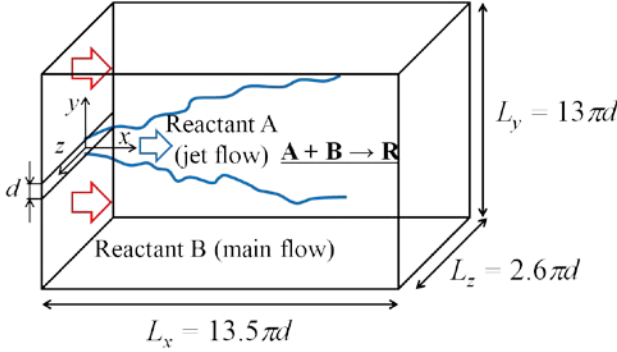
Fig. 1. Schematic diagram of the reactive planar jet

Here, u_i and γ_α are fluctuation components of i direction velocity, U_i, and concentration of species α, Γ_α, respectively, and $\langle\,\rangle$ denotes an ensemble average. $D_{t\alpha,i}$ is the eddy diffusivity of species α, corresponding to the i direction. $D_{t\alpha,i}$ is the key parameter in the Reynolds-averaged approach using the gradient diffusion model for the turbulent mass flux. Bilger et al. [2] and Komori et al. [3] have experimentally measured the eddy diffusivity of reactant species in a scalar mixing layer, and they have shown that the chemical reaction changes the eddy diffusivity. Because experimental measurements of the eddy diffusivity in turbulent reactive flows are difficult, the effects of chemical reactions on the eddy diffusivity have not been investigated sufficiently.

In this study, DNS of a planar turbulent jet with a second-order reaction is performed, and we directly calculate the eddy diffusivity of the reactive species. The effects of the chemical reaction on $D_{t\alpha,i}$ are investigated in detail.

2   Direct Numerical Simulation

Figure 1 shows the schematic diagram of the planar jet with the second-order reaction $A + B \rightarrow R$. The reactant A is premixed into the jet, and the other reactant B is premixed into the main flow. The planar jet is issued through a rectangular slit with width of d. Species R is produced by the chemical reaction, which is isothermal and irreversible. Concentrations of reactive species are assumed to behave as passive scalars. Therefore, the governing equations for this problem are incompressible Navier – Stokes equations, the continuity equation, and the passive scalar transport equations in their instantaneous, local form:

$$\frac{\partial U_i}{\partial x_i} = 0, \tag{2}$$

$$\frac{\partial U_i}{\partial t} + U_j \frac{\partial U_i}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \nu \frac{\partial^2 U_i}{\partial x_j \partial x_j},$$ (3)

$$\frac{\partial \Gamma_\alpha}{\partial t} + U_j \frac{\partial \Gamma_\alpha}{\partial x_j} = D_\alpha \frac{\partial^2 \Gamma_\alpha}{\partial x_j \partial x_j} + S_\alpha.$$ (4)

Here, ν is the kinetic viscosity, D_α is the molecular diffusivity of species α, and S_α is the production rate of Γ_α by the chemical reaction. D_α is assumed to be constant for all reactive species. For the second-order reaction $A + B \rightarrow R$, S_α is given by

$$S_R = -S_A = -S_B = k\Gamma_A \Gamma_B.$$ (5)

Here, k is the reaction rate constant. Equation (4) is solved for the product species R. We define the mixture fraction ξ as follows,

$$\xi = \frac{\Gamma_A - \Gamma_B + \Gamma_{B0}}{\Gamma_{A0} + \Gamma_{B0}},$$ (6)

where, $\Gamma_{\alpha 0}$ is the initial concentration of species α. Transport equation for ξ can be obtained by Eqs. (4) for species A and B. Because $S_A = S_B$, the chemical reaction does not change ξ. We solve the transport equation for ξ instead of solving Eqs. (4) for reactant species A and B, and Γ_A and Γ_B are calculated from Γ_R and ξ by using the mass conservation law written as follows [2];

$$\Gamma_A = \Gamma_{A0}\xi - \Gamma_R,$$ (7)

$$\Gamma_B = \Gamma_{B0}(1-\xi) - \Gamma_R.$$ (8)

The computational domain is resolved with $475 \times 425 \times 65$ grid points. The grid is equidistant in the x and z directions. In the y direction, the grid is stretched near the lateral boundaries, and fine grid is used near the jet centerline. The spatial resolution is the same order as the Kolmogorov scale. The fractional step method is used to solve the governing equations which are spatially discretized by using central differences, and the Poisson equation is solved by conjugate gradient method. The variables are stored on a staggered grid arrangement. The third-order Runge – Kutta method and the Crank – Nicolson method are used for temporal discretization.

At the outflow boundary, the convective outflow condition is applied. The velocity data at inflow plane is generated by superimposing random fluctuations on the mean velocity profiles. The velocity data at inflow plane is determined by the mean velocity profile and the variance of x direction velocity fluctuation which are measured at the jet exit in our experiments [4]. The ratio of the streamwise mean velocity in the main flow, U_M, to the streamwise velocity averaged in the jet exit, U_J, is $U_M/U_J = 0.056$. The traction-free condition is applied at the lateral boundary, and the periodic boundary condition is applied to the homogeneous (z) direction. The Reynolds
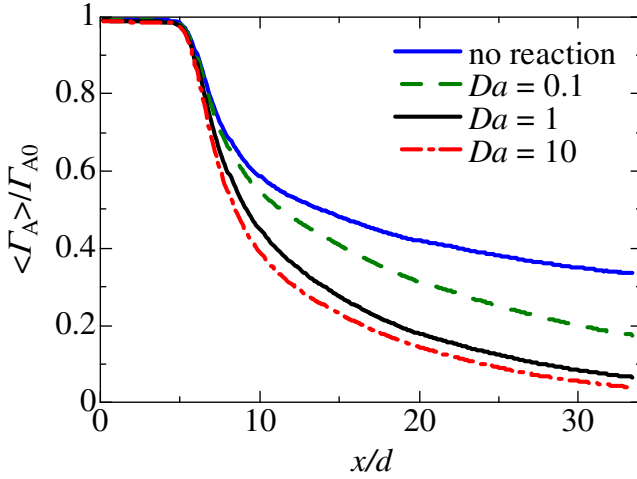
Fig. 2. Axial profiles of mean concentration of reactant A
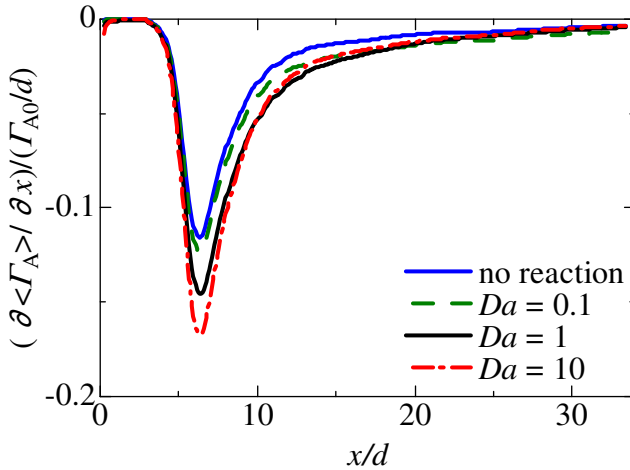
Fig. 3. Axial profiles of x gradient of mean concentration reactant A

number based on U_J and d is set to 2,000. The concentrations of reactive species are imposed to satisfy $\Gamma_A = \Gamma_{A0} = 0.4$ mol/m^3 and $\Gamma_B = \Gamma_R = 0$ in the jet exit, and $\Gamma_B = \Gamma_{B0} = 0.2$ mol/m^3 and $\Gamma_A = \Gamma_R = 0$ in the main flow. Therefore, from Eq. (6), $\xi = 1$ in the jet exit and $\xi = 0$ in the main flow. Schmidt number ($Sc = \nu/D_\alpha$) is set to 1.

Damköhler number is the ratio of a time scale of the flow to a time scale of the chemical reaction, and is defined by $Da = k(\Gamma_{A0} + \Gamma_{B0})d/U_J$. The reactive planar jet is simulated with three Damköhler numbers, namely, $Da = 0.1$, 1, and 10, by adjusting the reaction rate constant k.
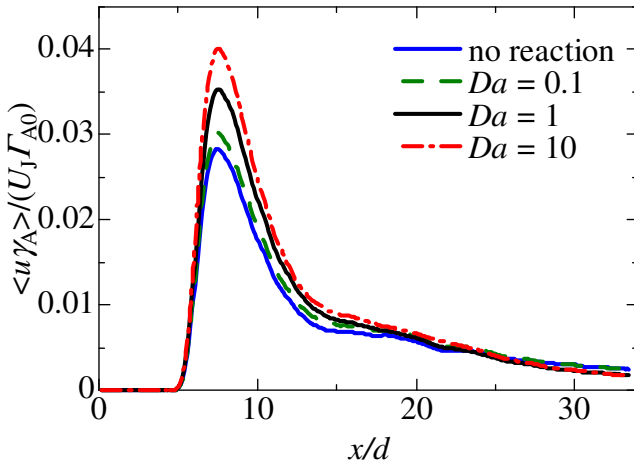
Fig. 4. Axial profiles of streamwise turbulent mass flux of reactant A
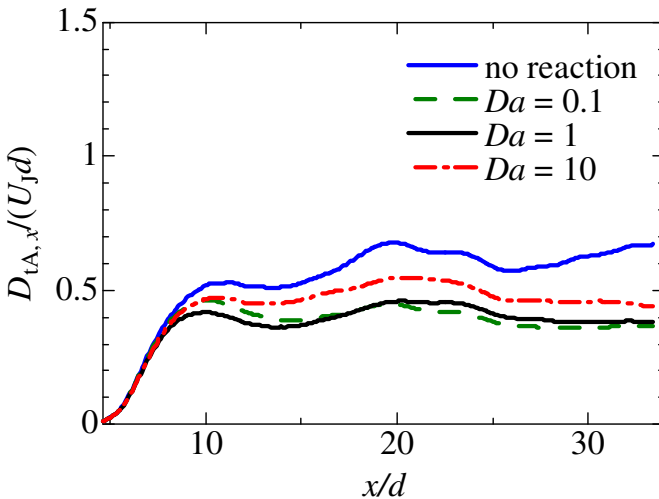
Fig. 5. Axial profiles of axial eddy diffusivity of reactant A

3   Results and Discussion

Figure 2 shows the axial profiles of mean concentration of reactant A, $<\Gamma_A>$, on the jet centerline for Da = 0.1, 1, and 10 and for the non reactive case. $<\Gamma_A>$ monotonically decreases in the downstream direction, and the chemical reaction makes $<\Gamma_A>$ small because the chemical reaction consumes the reactant species A. It is also found that as Da increases, $<\Gamma_A>$ becomes small because large Da means that the chemical reaction proceeds rapidly.

Figure 3 shows $\partial \Gamma_A / \partial x$ on the jet centerline. In Fig.3, $\partial \Gamma_A / \partial x$ is normalized by d and Γ_{A0}. It is found that the chemical reaction makes $\partial \Gamma_A / \partial x$ large in the region of $5 < x/d < 25$. Especially, the change of $\partial \Gamma_A / \partial x$ by the chemical reaction is large at $x/d \approx 7$. In the downstream region ($25 < x/d$), the chemical reaction hardly changes $\partial \Gamma_A / \partial x$, and the magnitude of $\partial \Gamma_A / \partial x$ is very small.

Figure 4 shows the streamwise turbulent mass flux of reactant species A, $<u\gamma_A>$, on the jet centerline. In Fig. 4, $<u\gamma_A>$ is normalized by U_J and Γ_{A0}. Figure 4 shows that the chemical reaction makes $<u\gamma_A>$ large in the region of $x/d < 25$. However, $<u\gamma_A>$ for $Da = 1$ and 10 is smaller than $<u\gamma_A>$ for the non reactive case in the region of $25 < x/d$.

Figure 5 shows the axial eddy diffusivity of reactant species A, $D_{tA,x}$. In Fig. 5, $D_{tA,x}$ is normalized by d and U_J. Figure 5 shows that the chemical reaction makes $D_{tA,x}$ small, and the effect of the chemical reaction on $D_{tA,x}$ for $Da = 0.1$ and 1 is larger than that for $Da = 10$. It is found that the chemical reaction affects the eddy diffusivity of reactant species, and the effects of the chemical reaction on the eddy diffusivity have the Damköhler number dependency. These results imply that the gradient diffusion model for the turbulent mass flux should be used with considerations for the chemical reaction effects on the eddy diffusivity.

4   Conclusions

The eddy diffusivity of reactant species A is investigated in the planar jet with the second-order chemical reaction ($A + B \rightarrow R$) by using direct numerical simulation. Simulations are performed at different Damköhler numbers, $Da = 0.1$, 1, and 10. The results show that the mean concentration of species A, $<\Gamma_A>$, becomes small due to the chemical reaction. It is also shown that the chemical reaction makes the streamwise turbulent mass flux of reactant species A, $<u\gamma_A>$, large in the region of $x/d < 25$, whereas $<u\gamma_A>$ for $Da = 1$ and 10 becomes small due to the chemical reaction in the region of $25 < x/d$. The axial eddy diffusivity of species A, $D_{tA,x}$, is calculated from $\partial \Gamma_A / \partial x$ and $<u\gamma_A>$. It is found that the chemical reaction makes $D_{tA,x}$ small, and the effects of the chemical reaction on $D_{tA,x}$ depend on the Damköhler number. Thus, it is concluded that the Reynolds-averaged approach using the gradient diffusion model should be applied to turbulent reactive flow with considerations of the effects of the chemical reaction on the eddy diffusivity.

References

1. Combest, D.P., Ramachandran, P.A., Dudukovic, M.P.: On the Gradient Diffusion Hypothesis and Passive Scalar Transport in Turbulent Flows. Industrial Engineering Chemistry Research 50, 8817–8823 (2011)
2. Bilger, R.W., Saetran, L.R., Krishnamoorthy, L.V.: Reaction in a Scalar Mixing Layer. Journal of Fluid Mechanics 233, 211–242 (1991)
3. Komori, S., Nagata, K., Kanzaki, T., Murakami, Y.: Measurements of Mass Flux in a Turbulent Liquid Flow with a Chemical Reaction. AIChE Journal 39, 1611–1620 (1993)
4. Watanabe, T., Sakai, Y., Yasuhara, H., Nagata, K., Terashima, O., Kubo, T.: Simultaneous Measurements of Velocity and Reactive Scalar in Plane Liquid Jet with the Second-Order Chemical Reaction. Trans. JSME 78-788 B, 769–782 (2012)

A Design Method of Derivative State Constrained H_2 Integral Servo Controller for Suppressing under Damping of Oscillatory System

Noriyuki Komine[1], Masakatsu Nishigaki[1], Tadanori Mizuno[2], and Kunihiro Yamada[3]

[1] Graduate School of Science and Technology, Shizuoka University
3-5-1, Johoku, Naka-ku, Hamamatsu City, 432-8011, Japan
`komine@keyaki.cc.u-tokai.ac.jp, nisigaki@inf.shizuoka.ac.jp`
[2] Faculty of Information Science, Aichi Institute of Technology
1247 Yachigusa, Yakusa-cho, Toyota, 470-0392, Japan
`mizuno@mizulab.net`
[3] Professional Graduate School of Embedded Technology, Tokai University
2-2-12, Takanawa, Minato-ku, Tokyo, 108-0074, Japan
`yamada@kunighiroi.com`

Abstract. A design method in this paper concerns the optimal H_2 integral servo problems for two-inertia model via the constraints of the derivatives of state variables added to the standard constraints. It is shown in the paper that the derivative state constrained optimal H_2 integral servo problems can be reduced to the standard optimal H_2 control problem. The main subject of the paper is to derive the derivative the theorem for state constrained H_2 integral servo. The effect of our proposed controller with respect to reduce an under damping for two-inertia model system is also verified.

Keywords: Optimal H_2 controller, Integral servo problem, Oscillatory system.

1 Introduction

The problem of how to control an under damping response of controlled system has been one of the fundamental problem in practical control engineering. The optimal servo control method minimizing a given performance index is known as a method for tracking desired speed of motor system with constant disturbance of its load. However, it is difficult to select the weighting matrices of performance index to mitigate an under damping response of oscillatory system. The integral servo problem is initiated by H. W. Smith and E. J. Davison [1], in which they proposed dual approaches, prototype affine and differential transformations, and gave some suggestions on measurement feedback schemes. However, successive researches are restricted only on the affine transformation approach that introduces integrators deductively and employs exclusively the state feedback [2],[3]. The optimal H_2 integral control yields a zero steady-state tracking error for a disturbance which equals to both the dimension of the disturbance input and the dimension of the reference output. The regulator problem is formulated as the optimal control for

oscillatory system such as minimizing a performance criterion involving time derivatives of state vector as well as usual system two-norm [4],[5].

In this paper, we derive the optimal H_2 integral servo controller which stabilizes an oscillatory system such that the optimal control law is more effective to control an under damping steady-state tracking error by H_2 control framework. The main subject of the paper is to apply the results of derivative state constrained H_2 integral servo theorem in the oscillatory system.

2   H$_2$ Integral Servo Problem

Let us consider a system as

$$\frac{d}{dt} x(t) = Ax(t) + B_2 u(t) + d \quad x(0) = x_0$$
$$y(t) = C_2 x(t) \tag{1}$$

where $x(t) \in R^n$ is the state, $u(t) \in R^m$ is the input, d is the constant disturbance and $y(t) \in R^p$ is the output. The matrices A, B_2 and C_2 are denoted the state matrix, input matrix and output matrix, respectively. Define the tracking error $e(t)$ and its integration is represented as;

$$e(t) = r(t) - y(t) \tag{2}$$

$$x_I(t) = \int_0^t e(\tau)d\tau \quad or \quad \frac{d}{dt} x_I(t) = r(t) - y(t) \tag{3}$$

where $x_I(t) \in R^l$. In order to zero the steady state tracking error $\lim_{t \to \infty} e(t)$ should be vanished, the derivative augmented state vector defined as $\left[\frac{dx(t)}{dt} \quad \frac{dx_I(t)}{dt} \right]^T$ which should be vanished for approaching infinity of t. The derivative augmented system is given by combining of the derivative state equation of Eq.(1) and the derivative state equation of (3) as

$$P_1(s) \begin{cases} \frac{d}{dt}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \begin{bmatrix} B_1 & 0 & | & 0 & 0 \\ 0 & -B_{1I} & | & I & 0 \end{bmatrix}\dot{w}(t) + \begin{bmatrix} B_2 \\ 0 \end{bmatrix}\dot{u}(t) \\ \dot{z}(t) = \begin{bmatrix} C_1 & 0 \\ 0 & C_{1I} \\ \hline 0 & 0 \end{bmatrix}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \hline D_{12} \end{bmatrix}\dot{u}(t)) \\ \dot{y}_I(t) = \begin{bmatrix} 0 & I \end{bmatrix}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & | & D_{21I} \end{bmatrix}\dot{w}(t) \end{cases} \tag{4}$$

where, the disturbance $w(t) = \begin{bmatrix} w_1^T(t) & w_2^T(t) & r^T(t) & | & w_3^T(t) \end{bmatrix}^T$ is continuously differentiable with respect to t and the matrices $B_1, B_{1I}, C_1, C_{1I}, D_{12}$ and D_{21I} are denoted the design parameter matrices to obtain the integral servo controller.

Statment of H_2 integral servo problem

Let $r(t)$ denote the step reference vector. Optimal H_2 integral servo problem is to find an admissible optimal integral controller such that the controlled plants with augmented integrator is stabilized and the output $y(t)$ tracks the constant reference signal $r(t)$ while minimizing the H_2 norm of the closed-loop transfer function with controlled plant from the $L[\dot{w}(t)]$ to $L[\dot{z}(t)]$ of $P_1(s)$ where $L[\cdot]$ denote the notation of Laplace transformation.

Solution of the statement of H_2 integral servo problem

The optimal H_2 integral servo controller for the controlled plant $P_1(s)$ is given as

$$\frac{d}{dt}\begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\hat{x}}_I(t) \end{bmatrix} = \hat{A}_2 \begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\hat{x}}_I(t) \end{bmatrix} - L_2 \dot{y}_I(t)$$

$$\dot{u}(t) = \begin{bmatrix} F_2 & F_{2I} \end{bmatrix} \begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\hat{x}}_I(t) \end{bmatrix}$$

(5)

where,

$$\hat{A}_2 = \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \begin{bmatrix} B_2 \\ 0 \end{bmatrix} F_2 + L_2 \begin{bmatrix} 0 & I \end{bmatrix}, F_2 = -\begin{bmatrix} B_2^T & 0 \end{bmatrix} X_2, L_2 = -Y_2 \begin{bmatrix} 0 & I \end{bmatrix}^T$$

and X_2, L_2 are given by the positive semi-definite solution of the Riccati equations derived from $X_2 = Ric(H_2) \geq 0$, $Y_2 = Ric(J_2) \geq 0$, where

$$H_2 = \begin{bmatrix} \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} & -\begin{bmatrix} B_2 \\ 0 \end{bmatrix}\begin{bmatrix} B_2^T & 0 \end{bmatrix} \\ -\begin{bmatrix} C_1^T & 0 & 0 \\ 0 & C_{1I}^T & 0 \end{bmatrix}\begin{bmatrix} C_1 & 0 \\ 0 & C_{1I} \\ 0 & 0 \end{bmatrix} & -\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}^T \end{bmatrix}$$

$$J_2 = \begin{bmatrix} \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}^T & -\begin{bmatrix} 0 \\ I^T \end{bmatrix}\begin{bmatrix} I & 0 \end{bmatrix} \\ -\begin{bmatrix} B_1 & 0 & 0 \\ 0 & -D_{1I} & 0 \end{bmatrix}\begin{bmatrix} B_1^T & 0 \\ 0 & -D_{1I}^T \\ 0 & 0 \end{bmatrix} & -\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} \end{bmatrix}$$

under the assumptions

(A1)' $\left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}, \begin{bmatrix} B_2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & I \end{bmatrix} \right)$ is stabilizable and detectable.

(A2)' D_{12} has full column rank, D_{21I} has full row rank.

(A3)' $\left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} - j\omega I \quad \begin{bmatrix} B_2 \\ 0 \end{bmatrix} \\ \begin{bmatrix} C & 0 \\ 0 & C_{1I} \\ 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 0 \\ D_{12} \end{bmatrix} \right)$ has full column rank for all ω.

(A4)' $\left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \\ [0 & I] \end{bmatrix} - j\omega I \quad \begin{bmatrix} B_1 & 0 & 0 & 0 \\ 0 & -B_{1I} & I & 0 \\ 0 & 0 & 0 & D_{21I} \end{bmatrix}\right)$ has full row rank for all ω.

3   Derivative State Constrained H_2 Integral Servo Problem

In order to suppress an unwanted oscillation of output in servo problem, the standard plant with derivative state is introduced as in the following equations:

$$P_1(s): \begin{cases} \dfrac{d}{dt}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \begin{bmatrix} B_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -B_{1I} & 0 & 0 & I & 0 \end{bmatrix}\dot{w}(t) + \begin{bmatrix} B_2 \\ 0 \end{bmatrix}\dot{u}(t) \\[4mm] \dot{z}(t) = \begin{bmatrix} C_1 & 0 \\ 0 & C_{1I} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & D_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & D_{11I} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\dot{w}(t) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ D_{12} \end{bmatrix}\dot{u}(t)) \\[4mm] \dot{y}_I(t) = \begin{bmatrix} 0 & I \end{bmatrix}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & D_{21I} \end{bmatrix}\dot{w}(t) \end{cases} \tag{6}$$

where the matrices D_{11} and D_{11I} are denoted the design parameter matrices to obtain the derivative constrained integral servo controller.   The disturbance $w(t) = \begin{bmatrix} w_1^T(t) & w_2^T(t) & \dot{x}^T(t) & \dot{x}_I^T(t) & r^T(t) & w_3^T(t) \end{bmatrix}^T$ is continuously differentiable in time. By definition of the optimal H_2 integral servo problem, the augmented general plant is given by

$$\hat{P}_2(s): \begin{cases} \dfrac{d}{dt}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \breve{B}_1\dot{w}(t) + \begin{bmatrix} B_2 \\ 0 \end{bmatrix}\dot{u}(t) \\[4mm] \dot{z}(t) = \breve{C}_1\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \breve{D}_{12}\dot{u}(t)) \\[4mm] \dot{y}_I(t) = \begin{bmatrix} 0 & I \end{bmatrix}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \breve{D}_{21}\dot{w}(t) \end{cases} \tag{7}$$

where

$$\breve{B}_1 = \begin{bmatrix} B_1 & 0 & AD_{11} & 0 & 0 & 0 \\ 0 & B_{1I} & -C_2D_{11I} & 0 & I & 0 \end{bmatrix}$$

$$\breve{C}_1 = \begin{bmatrix} C_1 & 0 \\ 0 & C_{1I} \\ D_{11}A & 0 \\ -D_{1I}C_2 & 0 \\ 0 & 0 \end{bmatrix}, \quad \breve{D}_{12} = \begin{bmatrix} 0 \\ 0 \\ D_{11}B_2 \\ 0 \\ D_{12} \end{bmatrix}$$

$$\breve{D}_{21} = \begin{bmatrix} 0 & 0 & 0 & D_{11I} & 0 & D_{21} \end{bmatrix}.$$

Statement of Derivative State Constrained H_2 integral servo problem

Let $r(t)$ denote the step reference vector. Derivative State Constrained Optimal H_2 servo integral problem is to find an admissible optimal integral controller such that the controlled plants with augmented integrator is stabilized and the output $y(t)$ tracks the constant reference signal $r(t)$ while minimizing the H_2 norm of the closed-loop transfer function with controlled plant from $L[\dot{w}(t)]$ to $L[\dot{z}(t)]$ of $\hat{P}_2(s)$.

4   Solution of the Statement of Derivative State Constrained H_2 Integral Servo Problem

The correct   The solution to the derivative state constrained H_2 optimal control defined above is given by the following procedure.

i) Singular value decomposition

There always exist unitary matrices $V_j, U_j, j = 1,2$ for the singular value decomposition of $\breve{D}_{12}$ and $\breve{D}_{21}$;

$$\breve{D}_{12} = U_1 \begin{bmatrix} 0 \\ \Sigma_1 \end{bmatrix} V_1, \Sigma_1 = \begin{bmatrix} \sigma_{11} & & \\ & \ddots & \\ & & \sigma_{1r} \end{bmatrix}, r = \dim(u) \tag{8}$$

$$\breve{D}_{21} = U_2 [0 \quad \Sigma_1] V_{12}, \Sigma_2 = \begin{bmatrix} \sigma_{11} & & \\ & \ddots & \\ & & \sigma_{1l} \end{bmatrix}, l = \dim(\begin{bmatrix} y(t) \\ y_I(t) \end{bmatrix}) \tag{9}$$

where $\Sigma_i, i = 1,2$ are the diagonal singular value matrices. Using the results obtained above, input and output vectors and accordingly the generalized plant are transformed as follows.

ii) Variable transformation

The generalized plant can be obtained by using the following variable transformations defined by

$$\dot{w}(t) = V_2 \dot{\hat{w}}(t) \tag{10}$$

$$\dot{\hat{z}}(t) = U_1^T \dot{z}(t) \tag{11}$$

$$\dot{u}(t) = V_1 \Sigma_1^{-1} \dot{\hat{u}}(t) \tag{12}$$

$$\begin{bmatrix} \dot{\hat{y}}(t) \\ \dot{\hat{y}}_I(t) \end{bmatrix} = \Sigma_2^{-1} U_2^T \begin{bmatrix} \dot{y}(t) \\ \dot{y}_I(t) \end{bmatrix}. \tag{13}$$

Substituting (10) through (13) into (7), then the generalized plant is obtained as

$$
\hat{P}_3(s)\begin{cases}
\dfrac{d}{dt}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \hat{B}_1\dot{\hat{w}}(t) + \hat{B}_2\dot{\hat{u}}(t) \\[4mm]
\dot{\hat{z}}(t) = \hat{C}_1\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \hat{D}_{12}\dot{u}(t) \\[4mm]
\begin{bmatrix} \dot{y}(t) \\ \dot{y}_I(t) \end{bmatrix} = \hat{C}_2\begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \hat{D}_{21}\dot{\hat{w}}(t)
\end{cases}
\tag{14}
$$

where the coefficient matrices are given as follows:

$$
\hat{B}_2 = \begin{bmatrix} B_2 \\ 0 \end{bmatrix}V_1\Sigma_1^{-1}, \quad \hat{C}_2 = \Sigma_2^{-1}U_2^T\begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix}, \hat{B}_1 = \breve{B}_1 V_2, \quad \hat{C}_1 = U_1^T\breve{C}_1,
$$

$$
\hat{D}_{12} = U_1^T\breve{D}_{12}V_1\Sigma_1^{-1} = \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad \hat{D}_{21} = \Sigma_2^{-1}U_2^T\breve{D}_{12}V_2 = \begin{bmatrix} 0 & I \end{bmatrix}
\tag{15}
$$

The generalized plant is now reduced to the standard form of the H_2 control problem. Suppose that the transformed generalized plant parameter matrices (14) satisfy the following relations:

(A1) $\left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}, \ \hat{B}_2, \ \hat{C}_2\right)$ is stabilizable and detectable.

(A2) $\hat{D}_{12}$ and $\hat{D}_{21}$ have full rank.

(A3) $\left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} \quad \hat{B}_2 \\ \hat{C}_1 \quad \hat{D}_{12}\right)$ has full column rank for all ω.

(A4) $\left(\begin{bmatrix} A_2 & 0 \\ -C_2 & 0 \end{bmatrix} \quad \hat{B}_1 \\ \begin{pmatrix} C_2 & 0 \\ 0 & I \end{pmatrix} \quad \hat{D}_{21}\right)$ has full row rank for all ω.

Under the assumptions (A1)-(A4), the optimal H_2 controller to the generalized plant (14) is given by the transfer function representation as

$$
\hat{K}_{H_2}(s) = \left[\begin{array}{c|c} \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \hat{B}_2\hat{F}_2 + \hat{L}_2\hat{C}_2 & -\hat{L}_2 \\ \hline \hat{F}_2 & 0 \end{array}\right]
\tag{16}
$$

where

$$
\begin{aligned}
\hat{F}_2 &= -(\hat{B}^T{}_2 X_2 + \hat{D}_{12}\hat{C}_1) \\
\hat{L}_2 &= -(Y_2\hat{C}_2^T + \hat{B}_1\hat{D}_{21}
\end{aligned}
\tag{17.}
$$

A couple of the unique positive-definite solution X_2 and Y_2 in (17) can be expressed in the matrix Riccati equation forms as;

$$X_2 = Ric(H_2) > 0, \ Y_2 = Ric(J_2) > 0 \tag{18}$$

where a couple of matrices H_2 and J_2 are the Hamiltonian matrices as

$$H_2 = \begin{bmatrix} \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} - \hat{B}_2 \hat{D}_{12}^T \hat{C}_1 & -\hat{B}_2 \hat{B}_2^T \\ -\hat{C}_1^T \hat{C}_1 + \hat{C}_1^T \hat{D}_{12} \hat{D}_{12}^T \hat{C} & -\left\{ \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} - \hat{B}_2 \hat{D}_{12}^T \hat{C}_1 \right\}^T \end{bmatrix}$$

$$J_2 = \begin{bmatrix} \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}^T - \hat{C}_2^T \hat{D}_{21} \hat{B}_1^T & -\hat{C}_2^T \hat{C}_2 \\ -\hat{B}_1 \hat{B}_1^T + \hat{B}_1 \hat{D}_{12} \hat{D}_{21}^T \hat{B}_1^T & -\left\{ \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}^T - \hat{C}_2^T \hat{D}_{21} \hat{B}_1^T \right\}^T \end{bmatrix}$$
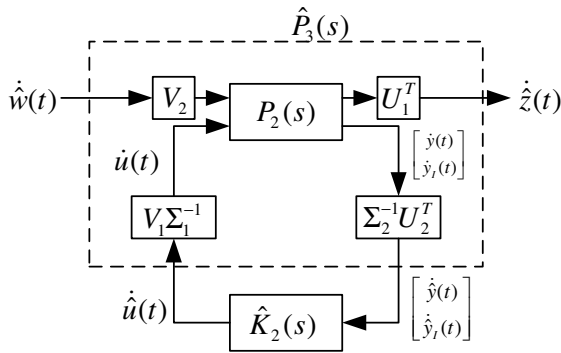
Fig. 1. General configuration

A general control formulation with the derivative state constrained optimal H_2 integral servo controller $\hat{K}_2(s)$ is given by the general configuration shown in Fig.1. Consequently, the assumptions supposed (A1) - (A4) can be reduced to the following expressions.

Lemma: Suppose the system parameter matrix in equation (14) satisfy the assumptions (A1) - (A4), then following assumptions hold;

(A1)' $\left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}, \begin{bmatrix} B_2 \\ 0 \end{bmatrix}, \begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix} \right)$ is stabilizable and detectable.

(A2)' D_{12} and $\begin{bmatrix} D_{21} & 0 \\ 0 & D_{21I} \end{bmatrix}$ have full rank.

(A3)' $\left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} - j\omega I \quad \begin{bmatrix} B_2 \\ 0 \end{bmatrix} \\ \begin{bmatrix} C_1 & 0 \\ 0 & C_{1I} \end{bmatrix} \qquad D_{12} \right)$ has full column rank for all ω.

(A4)' $\left(\begin{bmatrix} A & 0 \\ -C_p & 0 \end{bmatrix} - j\omega I \quad \begin{bmatrix} B_1 & 0 \\ 0 & B_{1I} \end{bmatrix} \\ \begin{bmatrix} C_2 & 0 \\ 0 & C_I \end{bmatrix} \qquad \begin{bmatrix} D_{21} & 0 \\ 0 & D_{21I} \end{bmatrix} \right)$ has full row rank for all ω.

Proof of Lemma:It is clearly shown that the optimal solution for the generalized plant (14) can be obtained under the assumptions (A1)'- (A3)', as the facts of the rank properties.

Thus, the optimal H_2 servo control solution for the system (7) is given as in (16) under the assumptions (A1)'-(A4)'. We have the following main result.

5   Main Results

The theorem for the derivative state constrained H_2 integral servo problem is given as follows;

Theorem (Derivative State Constrained Optimal H_2 Integral Servo)
The derivative state constrained H_2 integral servo controller for the controlled plant (1)-(3) is given by the transfer function representation as

$$K_{H_2}(s) = \left[\begin{array}{c|c} \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \begin{bmatrix} B_2 \\ 0 \end{bmatrix} F_2 + L_2 \begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix}_2 & -L_2 \\ \hline F_2 & 0 \end{array} \right] \qquad (19)$$

or its dynamical equation form from (19) as

$$\frac{d}{dt} \begin{bmatrix} \hat{x}(t) \\ \hat{x}_I(t) \end{bmatrix} = \hat{A} \begin{bmatrix} \hat{x}(t) \\ \hat{x}_I(t) \end{bmatrix} - L_2 \int_0^t e(\tau) d\tau$$

$$u(t) = F_2 \begin{bmatrix} \hat{x}(t) \\ \hat{x}_I(t) \end{bmatrix} \qquad (20)$$

where

$$\hat{A}_2 = \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \begin{bmatrix} B_2 \\ 0 \end{bmatrix} \left[V_1\Sigma_1 F_2 + L_2\Sigma_2^{-1}U_2^T \begin{bmatrix} 0 & I \end{bmatrix} \right]$$

$$F_2 = V_1\Sigma_1^{-1}F_2 = -V_1\Sigma_1^{-1}V_1^T \left\{ \begin{bmatrix} B_2^T & 0 \end{bmatrix} X_2 + D_{12}^T \breve{C}_1 \right\}$$

$$L_2 = L_2\Sigma_2^{-1}U_2^T = -\left\{ Y_2 \begin{bmatrix} 0 \\ I^T \end{bmatrix} + \breve{B}_1\breve{D}_{21}^T \right\} U_2\Sigma_2^{-1}U_2^T.$$

The X_2 and Y_2 are given by the positive semi-definite solution of the Riccati equations derived from

$$X_2 = Ric(H_2) \ge 0$$

$$Y_2 = Ric(J_2) \ge 0$$

$$H_2 = \begin{bmatrix} \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + B_2 D_{12}^T C_1 & -B_2 B_2^T \\ -C_1^T C_1 + C_1^T D_{12} D_{12}^T C_1 & -\left\{ \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} - B_2 D_{12} C_1 \right\}^T \end{bmatrix}$$

$$J_2 = \begin{bmatrix} \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}^T - C_2^T D_{12}^T B_1^T & -C_2^T C_2 \\ -B_1 B_1^T + B_1 D_{21} D_{21}^T B_1^T & -\left\{ \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + C_2^T D_{21} B_1^T \right\}^T \end{bmatrix}$$

under the assumptions (A1)'-(A4)'.

Proof of the theorem: As the facts of the rank properties of the lemma, this immediately shows that the theorem imply the optimal solution to the generalized plant (4) under the assumptions of (A1)' - (A4)'. This concludes the proof of the theorem. □

6   Simulation Results of the Proposed Method

A torsional vibration is occurred to the speed of motor by connecting flexible shaft. The vibration is an impediment to improve the characteristics of the two-inertia system. The simulation and experimental results of the speed control of the two-inertia system using the proposed controller will be shown in this section. A structure model of a two-inertia system is shown in Fig.2.

The linear dynamic equation of the two-inertia resonant system with constant disturbance T_L is represented by

$$J_m \frac{d\omega_m}{dt} + F_m\omega_m = T_d + \tau_d$$

$$J_L \frac{d\omega_L}{dt} + F_L\omega_L = \tau_d - T_L \tag{21}$$

$$\frac{d\tau_d}{dt} = K_s(\omega_m - \omega_L)$$

where J_m, J_L, F_m, F_L and K_s are the inertia of motor, the inertia of load, the friction of motor, friction of load and spring constant of the shaft, respectively. The integral $x_I(t)$ of the error vector $e(t)$ between the reference input $r(t)$ and controlled output $\omega_m(t)$ is defined as

$$\frac{dx_I}{dt} = e(t) = r(t) - \omega(t) \tag{22}$$

The state equation with Eq. (21) is then given by

$$\frac{d}{dt}\dot{x}(t) = \begin{bmatrix} F_m/J_m & 0 & 1/J_m & 0 \\ 0 & -F_L/J_L & 1/J_L & 0 \\ K_s & -K_s & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \dot{x}(t) + \begin{bmatrix} 1/J_m \\ 0 \\ 0 \\ 0 \end{bmatrix} \dot{u}(t)$$

$$\begin{bmatrix} \dot{\omega}_m(t) \\ \dot{x}_I(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \dot{x}(t) \tag{23}$$

where, $x(t) = [\omega_m(t) \quad \omega_L(t) \quad \tau_d(t) \quad x_I(t)]^T$, $\omega_m(t)$ denotes the speed of motor at time t, $\omega_L(t)$ denotes the speed of load at time t and $\tau_d(t)$ represents the torque of shaft. The numerical values of J_m, J_L, K_s are shown in Table 1. In the case of the numerical values, the friction of motor, friction of load and spring constant of the shaft are neglected, respectively.
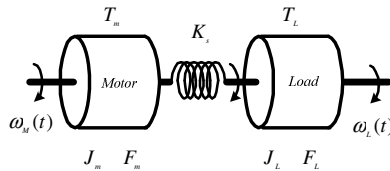
Fig. 2. Two-inertia model

Table 1. Numerical values of two-inertia system

$J_m[Kg \cdot m^2]$	$J_L[Kg \cdot m^2]$	$K_s[N/m]$
0.0866	0.0866	80

The designing parameters in the generalized plant B_1, C_1, D_{11}, D_{12}, D_{21} and D_{21I} are chosen as:

$$C_1 = B_1^T = \begin{bmatrix} diag\left[\sqrt{10^{qi}} & \sqrt{10^{qi}} & \sqrt{10^{qi}} & 20000 \right] \\ \left[0_{4\times 4} \right] \\ \left[0_{4\times 4} \right] \end{bmatrix}$$

$$D_{11} = \left[diag\left[\sqrt{e^{ni}} \quad \sqrt{e^{ni}} \quad \sqrt{e^{ni}} \quad \sqrt{100} \right] \right], \quad D_{12} = \sqrt{1},$$

$$\begin{bmatrix} D_{21} & 0 \\ 0 & D_{21I} \end{bmatrix} = \begin{bmatrix} \sqrt{0.01} & 0 \\ 0 & \sqrt{0.01} \end{bmatrix}.$$

The variation of closed-loop poles when $\sqrt{e^{ni}}$ varying from $ni = -9$ to $ni = -2$ is shown in Fig.3. It is seen that the original poles of the open-loop system locate on the imaginary axis. It verifies that the pair of poles with imaginary part approach to the real axis when the parameter $\sqrt{e^{ni}}$ becomes large.
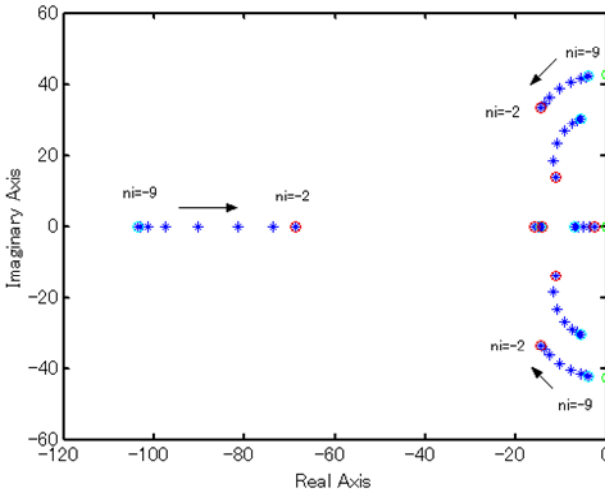
Fig. 3. Closed-loop poles location for $\sqrt{e^{ni}}$ varying from $ni = -9$ to $ni = -2$

The simulation results for step responses of the speed of motor with step disturbance shown in Fig.4 clearly explain the effectiveness of the proposed controller when the reference speed of the two-inertia system is assigned to be 2000 rpm. Significantly, the torsional resonance of two-inertia system is removed when the
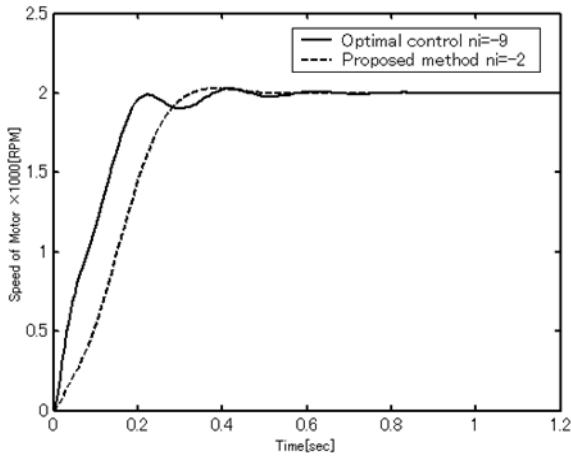
Fig. 4 Step responses for $ni = -2$ and $ni = -9$

designing parameter $\sqrt{e^{ni}}$ is equal to $ni=-2$ as shown by the dotted line. It is also seen that the torsional resonance of two-inertia system cannot be rejected by $ni = -9$.

7   Conclusions

The derivative state constrained H_2 integral servo controller for oscillatory system has been proposed in this paper. It has been shown that the parameter ni could reduce the imaginary part of closed-loop poles than a popular parameter qi in term of performance cost as H_2 optimization in the general framework. It is recognized that the servo problem can be applied to the systems whose reference inputs as well as disturbances are all given by step functions. It has been shown in an illustrative example that the proposed schemes has applied to reduce the under damping for two-inertia system.

References

1. Smith, H.W., Davison, E.J.: Design of Industrial Regulators. Proc. of the IEEE 119(8), 1210–1216 (1972)
2. Takeda, T., Kitamori, T.: A Design Method of Linear Multi-Input-Output Optimal Tracking Systems. Trans. SICE 14(4), 359–364 (1978)
3. Ikeda, M., Suda, N.: Synthesis of Optimal Servosystems. Trans. SICE 24(1), 40–46 (1988)
4. Trisuwannawat, T., Komine, N., Iida, M.: Optimal H2 Control of Oscillations via Derivative State Constraints. In: Proc. of 1999 American Control Conference (ACC), San Diego, California, USA, pp. 2305–2309 (1999)
5. Ichiyama, W., Komine, N., Benjanarasuth, T., Yoshida, M.: Optimal H2 Integral Controller with Derivative State Constraints for Suppressing under Damping of Oscillatory System. In: JSST 2011, Tokyo, Japan (2011)

A Temporal Coherence Based Framework for Visualizing Time-Varying Unstructured Volume with PBVR

Kun Zhao[1], Naohisa Sakamoto[2], and Koji Koyamada[2]

[1] Graduate School of Engineering, Kyoto University, Japan
[2] Institute for the Promotion of Excellence in Higher Education, Kyoto University, Japan
{zhao.kun,naohisas}@viz.media.kyoto-u.ac.jp,
koyamada.koji.3w@kyoto-u.ac.jp

Abstract. Time-varying visualization is always an important topic in the field of scientific visualization. Due to the large data size and complex grid, the unstructured time-varyingvisualization is still a hard problem. In this paper, we propose atemporal coherence based framework for visualizing large-scale time-varying unstructured volumewith the Particle-based volume rendering (PBVR).PBVR is a visibility sorting free rendering method, which can render a large-scale unstructured grid dataset efficiently. However, the pre-process of PBVR, which is to generate particles from the original volume, can always cost too much time. To overcome this problem, we utilize the temporal coherence between consecutive time steps to decrease the particle generation time.After the particle generation, we also compress the particle data before it is stored in hard disk. In account of the compressed data size, the loading time of the particle data during the rendering process is decreased obviously.And after the loading of the particle data, our system can render the data as an animation very smoothly by utilizing the LOD control of PBVR.We applied our system to rendering of 465 time steps of 3,676,828 hexahedral cell grid dataset to confirm the effectiveness.

1   Introduction

Time-dependent simulations and time-varying data can be found in almost every major scientific discipline. The time-varying data usually involves a certain phenomenon that grows, persists, and declines in several distinct stages. The rate of change at each stage could vary dramatically in space and time.As a result, the volume data from these simulations always have a highly complex and large-scale structure. Andan effective visualization for these time-varying volume data is extremely difficult due to the complexity of illustrating multiple time-steps and clearly showing changes and variations over time. This problem becomes even more conspicuous for the unstructured volume. Therefore, a good visualization method is also needed to render the unstructured dataset smoothly and effectively.

In the paper, we propose a temporal coherence based framework for effective time-varying unstructured volume visualization with PBVR.PBVR [1] is an efficient rendering method for the unstructured volume. For a given unstructuredvolume

data,PBVR first generate particles from the original unstructured volume dataand thenproject the particles onto the image plane. As a result, the visibility sorting is not needed so that a high frame rate can be promised for the rendering result.However, since the particle generation process of PBVR needs to be done for each time step, the generation could always cost too much time. To overcome this problem, we first analyze the temporal coherence between the consecutive time steps and then generate particles depending on the temporal coherence. With this approach, the particle generation time can be decreased greatly for a strong temporal coherence. Moreover,since the large-scale time-varying volume data can always be generated as a large particle data size, these generated particle data will need to be stored in the hard disk before rendering. This would lead to a long loading time when the particles are loaded for the rendering process. Hence, we compress the particle data before it is stored to hard disk so that these data can be loaded faster during rendering process.After the loading of the particle data, our system can render the particle as an animation smoothly by taking advantage of the level of detail (LOD) control of PBVR.In this paper, we apply our system to large-scale simulation results of oral airflow data composed of 465 time steps and 3,676,828 hexahedral cells for each step to confirm its effectiveness.

The remainder of this paper is organized as follows: Section 2 introducesthe related work of the time-varying unstructured volume visualization. Section 3introduces the rendering method of PBVR and section 4 introduces the utilizing of the temporal coherence in our system. Andour approach for visualizing the time-varying unstructured volume based on these methods is detailed in Section 5. The effectiveness of the proposed approach is presented in Section 6, followed by the discussion and conclusion of this paper in Section 7 and 8.

2   Related Work

In the last decades, there already exists some visualization method for the unstructured volume data [2, 3, 4, 9,10].Bernardon et al. [2] use the technique of HAVS (Hardware Assisted Visibility Sorting) to render the time-varying unstructured grid dataset. Since they have to store the volume data to the memory of GPU, it works well when there are just hundreds of thousands grids. However, when the dataset have millions of grids, this framework just doesn't work since there is not enough GPU memory.Furthermore,HAVS only supports data grid consisting of tetrahedral cell. For the data grid consisting of hexahedral cell such as the data grid used in this paper, the HAVS system does not work. Andfor the other rendering methods, a visibility sorting is always needed, which can increase the rendering cost for a large-scale volume data. And also, the hexahedra cell type is always not supported in these rendering systems.

On the other hand, the utilizing of the temporal coherence for efficient rendering also has been a central focus for many research efforts [6, 7, 8]. Due to the good performance, the temporal coherence can also be used in our PBVR system to improve the processing speed.

3   Particle-Based Volume Rendering

In our rendering system, we use PBVR to render the time-varying unstructured dataset.PBVR is based on the density emitter model of Sabella's theory [5], which uses grainier particles than the splatting algorithms and assumes that the particles are fully opaque. With this proposed particle model, PBVR requires neither alpha blending nor visibility ordering when rendering.

PBVR can be generally divided into three steps: particle generation, particle projection and ensemble averaging (Figure1).

At first step, PBVR generate particles from the volume data with the given transfer function(Figure 1.A). The generation is donecell by cell and itis done multiple times for every cell by using different random number. This process is performed by CPU as a pre-process.After this, the generated particles are projected to the image plane, which is called the particle projection process (Figure 1.B). This projection process is performed by GPU. At the same time, the calculating of the particle size and shadow processing is also performed by using normal vector.The third step ensemble averaging process means that, for the group of particles generated by using different random numbers, particle projection process is performed multiple times, and the projectedimages superimposeeach other averagely (Figure1.C).Here, the multiple times is called repeat level. As a result, a high-quality rendering results can be obtained by setting a high repeat number.However, with the increasing of the particle number, the particle data size is also increasing which would exert pressure on thesystem memory and may lead toa deceleration of the rendering speed.Hence, PBVR provide a LOD control which allows user to control the level of detail dynamicallyby setting different repeat level.
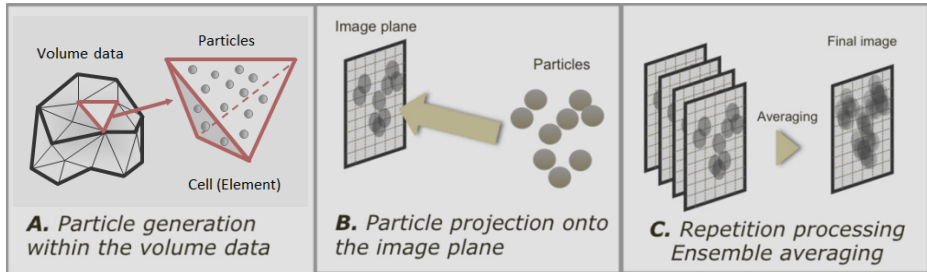
Fig. 1. The Three Steps of PBVR: Particle Generation, Particle Projection, Ensemble Averaging

4   Temporal Coherence Cells Table

Since the time-varying volume data always contains much temporal coherence, some part of the volume may have a strong similarity with the same part of consecutive steps volume. Hence, the generated particles of these coherence parts also have a strong similarity. That is to say, if we can find the coherence parts, the particles in these parts can be directly copied from the previous step. Generally, the particle

generation step can always cost much time, the direct copy from the previous step of some parts can decrease the generation time remarkably. Because PBVR generate particles in cell-by-cell manner (Figure 1.A), we propose a temporal coherence cells table to record the coherence property for every cell. Here, the coherence is calculated for every two continuous steps.Assume the two steps as the previous step and current step,our system calculate the coherence cells table as follows:

1. Calculate the difference of every vertexbetween these two steps into a vertices table (VT). Marking the difference for the   vertex as VT_i ,it is calculated as:

$$VT_i = \begin{cases} 0, & \left| V_i^{prev} - V_i^{prev} \right| < tolerance \\ 1, & else \end{cases} \quad (1)$$

Where V_i^{prev} represents the ith vertex value of the previous step, and V_i^{curr} represents the ith vertex value of the current step. User can set a tolerance rate ε , and our system will calculate the tolerance as:

$$tolerance = \varepsilon \times (V_{max}^{prev} - V_{min}^{prev}) \quad (2)$$

Here, V_{max}^{prev} and V_{min}^{prev} respectively means the maximum value and minimum value of the previous step.

2. Calculate the coherence for each cell into to a cells table (CT). Based on the VT calculated by step 1, the cells table value CT_i will be calculated as "1" if and only if the vertex values belonging to this cell are all calculated as "0" in VT. If not, CT_i will be marked as 0.

3. Delete VT, do the following process and then move to step 1.

$$\begin{cases} previous\ step = current\ step \\ current\ step = current\ step + 1 \end{cases} \quad (3)$$

Obviously, there is no CT for the first time step because it does not have the previous time step. For a hexahedral volume, the process of calculating the coherence cells table is shown in Figure 2.

Since the CT only need 1 bit (0 or 1) for each cell, the size of the CT for a time-varying volume dataset can be calculated as:

$$S_{CT} = (n-1) \times N_c \ (bits) \quad (4)$$

Where n means the number of time steps and N_c represents the number of cells for one time step volume. The building of CT needs to be accomplished before particle
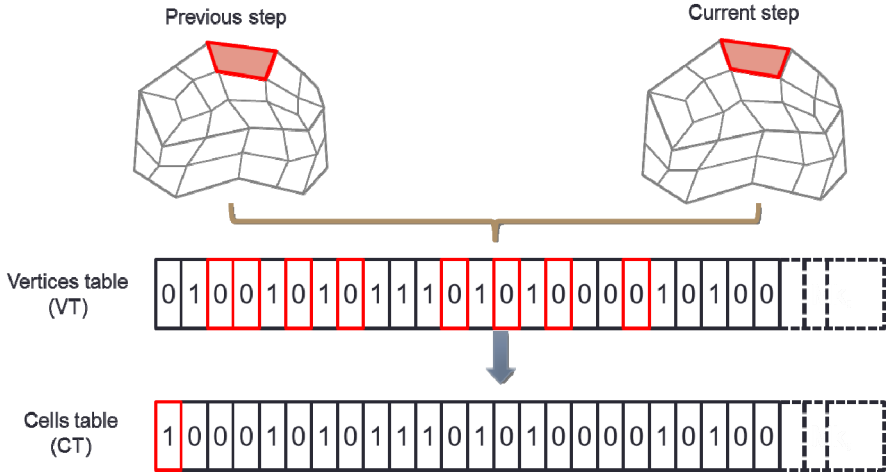
Fig. 2. The process of calculating the cells table for the time-varying hexahedral volume. Since a hexahedral cell has eight vertices, every value in the CT needs to check eight values on VT, which belongs to this cell. Here, the red frame shows the values belonging to one same cell.

generation so that it can be referred in the particle generation process. This will be detailed in section 5.

5   Proposed System

With the time-varying volume data, our system first build the coherence cells table for every step, then generate particles based on the cells table and store the compressed particles data into hard disk, at last our system load the compressed particle data render the data as ananimation.

5.1   Build Coherence Cells Table

The building of the cells table (CT) is done for every step and the CT tables for each step(see section 4) are stored before particle generation. The CT table is to be referred in particle generation process to determine whether the particles in one cell need to be generated or not. Since for every cell, the CT table only need one bit space to store "1" or "0", it would not cost so much space.

5.2   Generate Particles with CT

With the built CT table for every step of the original volume data, our system generate particles for every step based on the CT table. Assume the cells number of the volume as , the generation process is shown as following:

1. Load the CT table for the current time step and begin the particle generation process.
2. For the ith cell, check the value of .
3. If , copy particles from the previous time step. If , regenerate particles for this cell.
4. If , move to the next cell and back to step 2. If not, move to the next time step, back to the step 1.

Since the first time step does not have a CT, our system would first generated particles for all cells of this time step. And for the following step, our system only generated particles for the cell with the CT value equals 0, and copy the particles from the previous time step for the cell with the CT value equals 1.Moreover,to manage the particles, we also keep a particles table that storesthe number of particlesfor each cell. With this particles table, the copy process will be easy to find the particles needed to be copied by referring the particles table of the previous time step. In addition,the particles table is only used once for one step. And after the particle generation of the current time step is finished, the particles table of the previous time step will be deleted and a new particles table of the current time step is also built.

Furthermore, for every time step, particles aregenerated at a high repeat level and saved as one file (Figure 2).Since the generated particle dataset of the whole time steps are too large to be stored in the system memory, we store these data into the hard disk. During the rendering step, only the particles amounting to repeat level, which can be assigned by the user,are loaded and projected to the screen. Therefore, the user can control the number of particles projected to the image plane by changing the repeat level. A lower repeat level can provide a higher speed but a coarse image quality, and a higher repeat level can provide a higher image quality but a low processing speed.
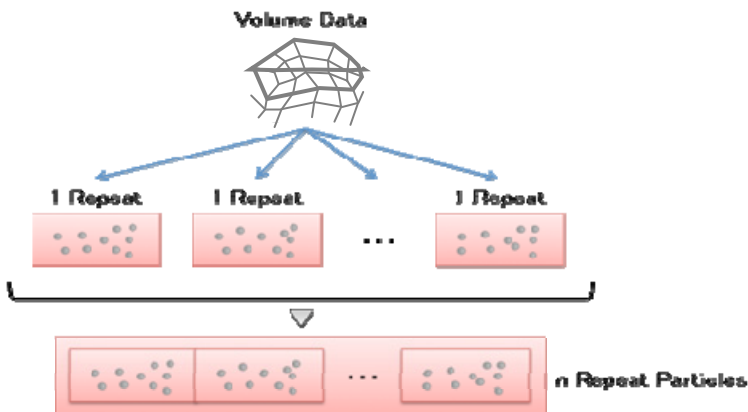
Fig. 3. The figure shows the process of repeatedly generation ofparticles. For one time step, the repeatedly generated particles are saved in one file. And during the rendering step, only the particles amountingto repeat level are projected to the screen.

Since the stored particles data need to be loaded to system memory before the rendering begins, the data loading could cost too much time for the large data size. To reduce the loading time, we also make an effort to compress the particle data before it is stored in hard disk. During the loading process,the compressed particle data can provide a faster hard disk accessing speed for the smaller data size. Then the memory-loaded compressed data will be expanded in the memory. And when the expanding is completed, the loaded compressed data will be deleted to save the memory space (Figure4). This process can provide a faster loading speed than using the original data (see section 6).
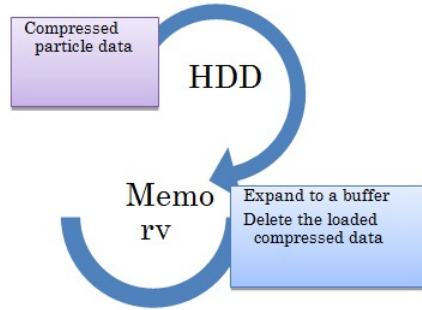
Fig. 4. The figure shows the loading process of the compressed particle data. For the smaller size, the loading of the compressed data can save much hard disk accessing time.

5.3   Particle Data Rendering

In the rendering process, we render the loaded time-varying particles as an animation.By utilizing the LOD control of PBVR, our system could provide a smooth animation and a high quality rendering image of any time step when the animation is stopped.

To get a high frame rate when the animation is flowing, we first utilize the low repeat level particle data. Before the animation is launched, we firstload the compressed particle data for a low repeat level from the hard disk for all time steps so that there is no need to access hard disk during the animation flows. In the rendering process, CPU transfers the loaded particle data to GPU step by step, and GPU projects the transferred particle data step by step so that the user can get a smooth rendering animation (Figure 5).Of course, the rendering is done with the ensemble averaging manner. When the particle data of step $t$is rendered by GPU, it will be deleted from GPU memory. As a result, our system does not consume so much GPU memory space.

And when we stop the animation at any time step, our system loads the compressed particle data for a high repeat level of this time step and renders it to the image plane. As a result, we can obtain a high-quality rendering image when the animation is stopped. In this case, the pre-loaded low repeat level particle data are still stored in the system memory; therefore the animation can be resumedat any time (Figure6). The process can make sure that the user could obtain a high quality image of any interested time step.
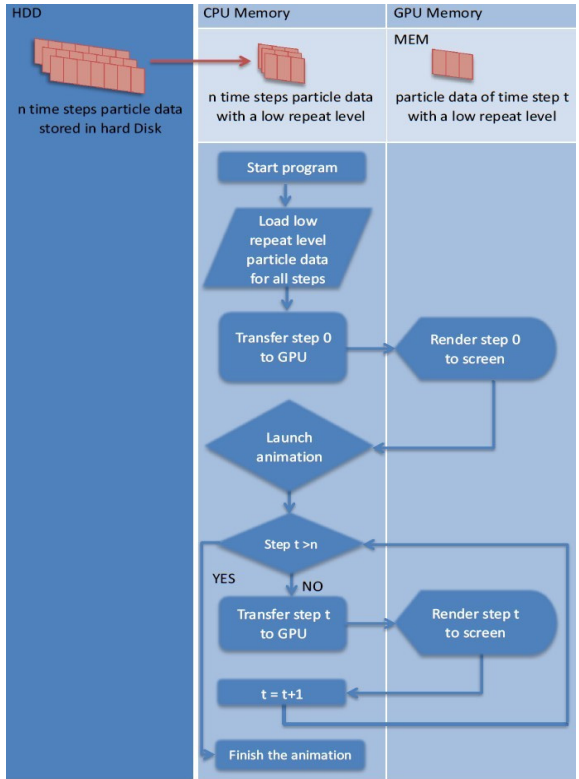
Fig. 5. Process Flow Diagram of the Animation
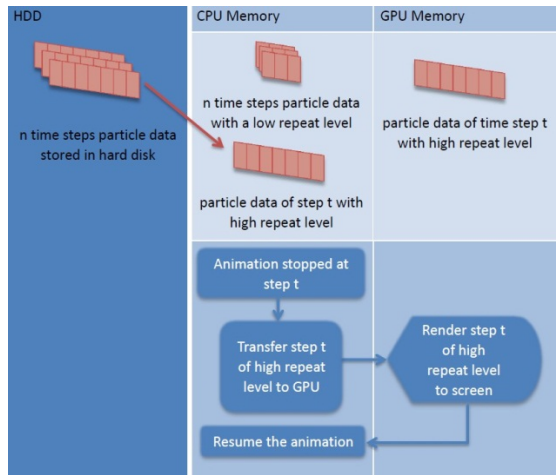
Fig. 6. Process flow diagram when the animation is stopped

As the user interface, our system allows user to stop or resume the animation by pressing keys on the keyboard. Moreover, our system also provides a time slider that shows whichtime step that is currently displayed.User is able to quickly view the interested time step by moving the slider using the mouse.

6   Experiment and Results

In this paper, we apply our system to a large-scale simulation of oral airflow data (composed of 465 time steps). This volume data contains 3,769,092 vertexes and 3,676,828 hexahedral grids. And it is composed in float type.With the time step varying, the grid of the volume data not changes, and only the value of the vertex is changing. The data contains several variables such as velocity, pressure, and so on. And in our experiment, we use the scalar value of velocity vector of the airflow in the rendering. The transfer function used in the experiment is shown in Figure 7. We conduct our experiment with the environment of aQuad-Core AMD Opteron Processer 2350 CPU (2.0 GHz), anNVIDIA QuadroFX 4700 1GB GPU, and 32GB system memory. The Operation System used is SUSE Linux Enterprise Server 10 SP2.

Fig. 7. Transfer Function Used in the Particle Generation

As the pre-process, the whole time step particle data are generated from the volume data. Table 1 shows the result of the particle generation. Here, we use the tolerance rate of 0.1% to build the coherence cells table. And to make a comparison, the particle generation time without using the coherence cells table is also shown in table 1. From the result we can see, our temporal coherence based framework can save much particle generation time. After the particle generation, we compress the particle data before they are stored to the hard disk. In our system, we use the zlib as the compression method because it can provide an in-memory compression and decompression, which could fast processing speed. The average compression ratio for all time steps is 50.36%.

Table 1. Theresult of particle generation for all time steps (, repeat level =144)

Volume Data Loading Time	189.2 sec
Coherence Cells Table(CT) Building Time	10.8 sec
Particle Generation Time	882.7 sec
Particle Generation Time without CT	2,188.4 sec
Average Particle Numberfor One Time Step	3,528,544

After the pre-process, our system will begin the rendering process. Before the animation is launched, we first load the low repeat level compressed data for all time steps. The loading of the particle data is shown in Table 2.As we can see, the compressed data can provide a faster loading time than the original particle data.

And then, after the animation is launched, the rendering images are shown in Figure 8. When the animation is stopped, the rendering imagesare shown in Figure 9.Also, the rendering speed when the animation is running and stopped is shown in Table 3. From these results we can see, the low repeat level particle data can provide a high rendering speed but a coarse image quality. And the high repeat level particle data can cost more time on loading and rendering but it can provide a good image quality.

Table 2. The loading timeof low repeat level particle data. For the compressed data, the loading time includes the hard disk accessing time and decompression time.

LoadingTime of CompressedLow Repeat Level Data for All Time Steps	27.89sec
LoadingTime ofOriginal Low Repeat Level Data for All Time Step	45.80 sec

Table 3. Animation rendering speed

Average Rendering Time for One Time Step in Animation	96.83msec (10.33FPS)
Loading and Rendering Timefor One Time Step when Stopped	1.09 sec (0.92 FPS)
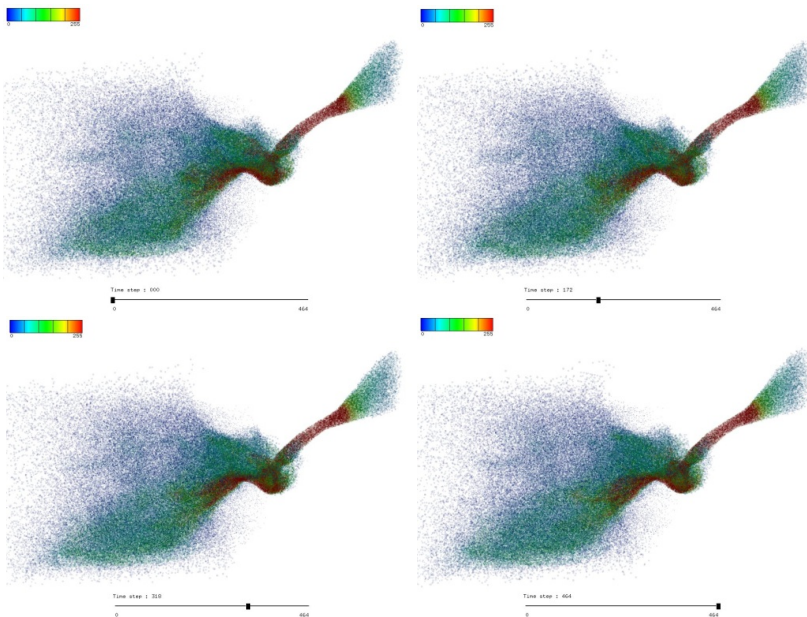
Fig. 8. The renderingimage during the animation. The figure on the upper left represents the image of time step 0, the upper right shows the time step 172, the lower left shows the time step 318 and the last one shows the image of time step 464.
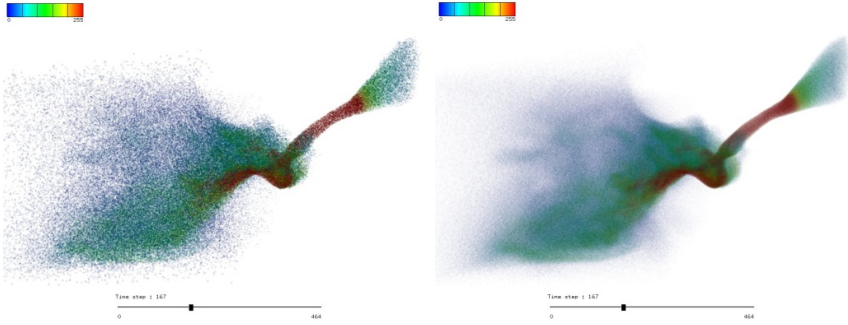
Fig. 9. The contrast rendering image of time step 167 during the animation flows (left figure) and stopped (right figure)

7   Discussion

From the rendering speed (Table 3) we can see, our system can provide a 10.33FPS rendering animation for a large-scale time-varying unstructured volume dataset. This result shows the large-scale time-varying unstructured volume dataset can be rendered smoothly with our system. Moreover, from Table 1 we can see, by utilizing the temporal coherence, our system can decrease the particle generation time obviously. Table 2 also shows the compression process could decrease the particle data loading time remarkably in our system. From these results, we can see our system is efficient on the rendering of the volume data with a large number of cells and time steps.

In our system, we copy the particles during the particle generation process by utilizing the coherence cells table. In fact, this particles copy process can also be moved to the rendering process. That is to say, we only need to store the generated particles without the copied particles for each step. Since the particles in the coherence cell do not need to be stored into the hard disk,the particle data size can be decreased a lot. However, this process would need to store a particles table for each time step to manage the generated particles so that the rendering process can know which part of particles would be needed to be copied. The particles table might also consume a lot of space, but we think there should be a trade-off, which can provide a better performance for our framework. This will be done in our future work.

8   Conclusions

In the present paper, we proposed a temporal coherence based framework for visualizing the time-varying unstructured volume with PBVR. With our framework, the particle generation time can be decreased a lot based on the temporal coherence. And then the generated particle data is compressed before stored to hard disk. Since the compressed data has a smaller data size, the accessing time for hard disk is also decreased when the particle data is loaded to system memory. After the loading of the particle data, our system can render the time-varying data at a high frame rate. And

also, our system can provide a good quality rendering image when the animation is stopped by the user so that the user can get a detail image for the interested time step.Moreover, to confirm the effectiveness, we apply our system to large-scale simulation results of oral airflow data composed of 465 time steps and 3,676,828 hexahedral cells for each step.From the experimental results we can see, our framework can provide an efficient performance for the large-scale time-varying unstructured volume data.

References

1. Kawamura, T., Sakamoto, N., Koyamada, K.: A Level-of-Detail Rendering of a Large-Scale Irregular Volume Dataset Using Particles. Journal of Computer Science and Technology 25(5), 905–915 (2010)
2. Bernardon, F.F., Callahan, S.P., Comba, J.L.D., Silva, C.T.: An Adaptive Framework for Visualizing Unstructured Grids with Time-varying Scalar Fields. Parallel Computing 33(6), 391–405 (2007)
3. Lum, E.B., Ma, K., Clune, J.: A Hardware-Assisted Scalable Solution for Interactive Volume Rendering of Time-Varying Data. IEEE Transactions on Visualization and Computer Graphics, 286–301 (2002)
4. Maximo, A., Marroquim, R., Farias, R.: Hardware-Assisted Projected Tetrahedra. Commputer Graphics Forum 29(3), 903–912 (2010)
5. Sabella, P.: A Rendering Algorithm for Visualizing 3D Scalar Fields. Computer Graphics 22(4), 51–58 (1988)
6. Guthe, S., Straßer, W.: Real-time decompression and visualization of animated volume data. In: Proceedings of IEEE Visualization Conference, pp. 349–356 (2001)
7. Lum, E.B., Ma, K.-L., Clyne, J.: A hardware-assisted scalable solution for interactive volume rendering of time-varying data. IEEE Transactions on Visualization and Computer Graphics 8(3), 286–301 (2002)
8. Shen, H.-W., Johnson, C.R.: Differential volume rendering: A fast volume visualization technique for flow animation. In: Proceedings of IEEE Visualization Conference, pp. 180–187 (1994)
9. Roettger, S., Ertl, T.: A two-step approach for interactive pre-integrated volume rendering of unstructured grids. In: VolVis, pp. 23–28 (2002)
10. Moreland, K., Angel, E.: A fast high accuracy volume renderer for unstructured data. In: 2004 IEEE Symposium on Volume Visualization and Graphics, pp. 9–16 (2004)

Detection of Linear Relationship among Dimensions in Multivariate Data by Parallel Coordinates

Chi Zhang, Naohisa Sakamoto and Koji Koyamada

Graduate School of Engineering, Kyoto University, Japan
`zhang.chi.24e@st.kyoto-u.ac.jp`

Abstract. When visualizing multivariate data as parallel coordinates, although it is convenient to view the relationship between two adjacent dimensions, the exploring of the relationships among three or more nonadjacent dimensions is relatively difficult. In this paper, we propose a way to solve this problem by using axis-translation method and axis-comparison method, which can keep the original positions of axes and exploring the relationships among three or more nonadjacent dimensions.

Keywords: parallel coordinates, linear relationship, axis-translation method, axis-comparison method.

1 Introduction

Parallel coordinates plot is widely used for visualizing and analyzing high dimensional data. In parallel coordinates, users can find the relationship between the data in two adjacent dimensions by viewing the slope of the lines between the two axes. However, the exploring of the relationships among three or more nonadjacent dimensions is relatively difficult. One of the solutions to view the relationship between two nonadjacent axes is to move these axes and locate them next to each other, which has the disadvantage that it changes the original positions of axes and it cannot show the relationship among three or more dimensions. In order to keep the original positions of axes in parallel coordinates and explore the relationship among dimensions, we apply axis-translation method and axis-comparison method to parallel coordinates to analyze the linear relationship among dimensions. By using axis-translation method, an additional axis is added to parallel coordinates and shifted in corresponding region. According to the intersections of the additional axis and the lines connected to the related axes, it is possible to find the linear relationship through the comparison of their positions and the positions of the relative points on the target axis. In the axis-comparison method, we add a new dimension with the linear relationship with the dimensions to be verified next to the target dimension. By changing the parameter of the new dimension, users are able to explore the lines between the new axis and the target axis so as to find the coefficient that can evaluate the linear relationship.

2   Related Works

Parallel coordinates were invented by Maurice d'Ocagne in 1885,[1] and were independently re-discovered and popularised by Al Inselberg[2] in 1959 and systematically developed as a coordinate system starting from 1977. Several methods and extensions about parallel coordinates used for analyzing hyperdimensional data were introduced by Wegman in 1990[3].

3   Proposed Method

3.1   Axis-Translation Method

Suppose the input data has the linear relationship among three dimensions which are labelled as x, y and z. $z = ax + by + c$ where a, b and c are constants. When utilize the axis-translation method, the first step is to normalize the maximal values and the minimal values of all dimensions and plot it as parallel coordinates. Here we use the maximal value and minimal value of the whole dataset as the maximal value and minimal value of all axes. Then an additional axis, or we can simply consider it as a line as well, is added to the present parallel coordinates plot. Shift this additional axis and connect the intersections of the axis and the original lines in the parallel coordinates plot to the data points on the z axis with lines. As the axis moves, the lengths and slopes of these lines are also changed. When the axis reaches some position where these lines meet in one point or they are parallel to each other, the distances between x axis, y axis and the additional axis and their proportions are calculated. From the proportions of these distances we can get the proportional coefficient of the original linear relationship.

3.2   Axis-Comparison Method

The axis-comparison method add a new dimension to the present parallel coordinates plot. Similarly, we suppose that the input data has the linear relationship among three dimensions which are labelled as x, y and z. $z = ax + by + c$ where a, b and c are constants. The additional dimension has its own data which is linear related to the dimension x and y. Let the new axis be labelled as α and we have the linear relationship $\alpha = x + sy$ or $\alpha = sx + y$. Then the data with the four dimensions are plotted as parallel coordinates. Here we locate the α axis between y axis and z axis. As the value of s is changed from -1 to 1, the slopes of the lines between α axis and z axis will be changed as well. If the maximal values and the minimal values are not normalized, in other words, the maximal value and the minimal value of each axis are the maximal value and the minimal value of that dimension, when all of the lines between α axis and z axis become horizontal, the value s is the proportional coefficient we want.

4   Experimental Results and Discussion

For simplicity and without loss of generality, two sample datasets are prepared with six data points and three linearly dependent dimensions in each dataset. One of the datasets is positive linear correlation in the three dimensions while the other one is negative linear correlation. If the three dimensions are labelled as x, y and z respectively, the two datasets can be listed as Table 1a and Table 1b. The linear relationship in the two datasets can be described as $z = 0.2x + 0.3y + 10$ and $z = 0.8x - 0.4y - 2$ respectively.

Table 1. Sample Data

(a) $z = 0.2x + 0.3y + 10$				(b) $z = 0.8x - 0.4y - 2$		
x	y	z		x	y	z
0.5	19.5	15.95		0.1	9.3	-5.64
2.2	14.6	14.82		3.6	4.2	-0.8
3.7	9.9	13.71		7.4	2.8	2.8
5.1	1.1	11.35		9.0	8.0	2
7.4	4.8	12.92		11.3	6.9	4.28
9.3	11.0	15.16		14.8	1.1	9.4

The results are illustrated in Figure 1 and Figure 2. Figure 1a and Figure 1b show the results by using the axis-translation method. Notice that the regions where the additional axis will be located depends on the sign of the proportional coefficient. If the dataset is positive linear correlation, the position will be found between x axis and y axis, which is shown in Figure 1a. If the dataset is negative linear correlation, the position will be found to the left of x axis, as illustrated in Fugure 1b. If we let the distance between two axes in the parallel coordinates plot be d, in Figure 1a, when all of the blue lines meet in one point, the distance between the additional axis and x axis, y axis are $0.6d$ and $0.4d$ respectively. Therefore, we can figure out the proportional coefficient of $\frac{x}{y}$ as $\frac{0.4d}{0.6d} = 0.67$. Figure 2a and Figure 2b show the results by using the axis-comparison method. In Figure 2a, an additional dimension is added to the input data and plotted as the α axis next to the left of z axis with the linear relationship of $\alpha = sx + y$. When $s = 0.67$, the lines between α axis and z axis become horizontal, indicating that in the input data, the proportional coefficient $\frac{x}{y} = s = 0.67$. Similarly, in Figure 2b, an additional dimension is added to the input data and plotted as the α axis next to the left of z axis with the linear relationship of $\alpha = x + sy$. When $s = -0.5$, the lines between α axis and z axis become horizontal, indicating that in the input data, the proportional coefficient $\frac{x}{y} = \frac{1}{s} = -2$.
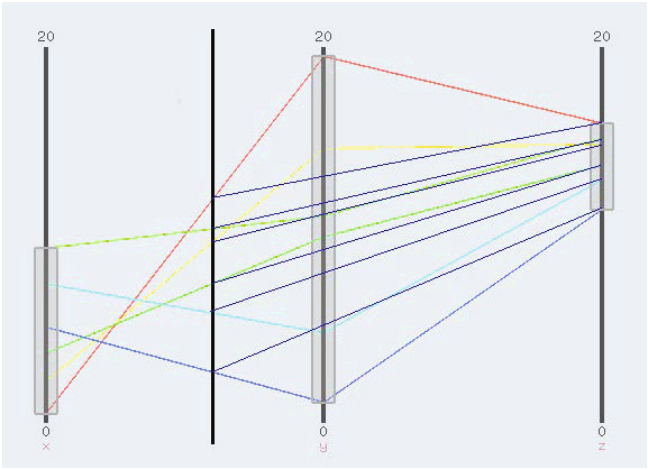
(a) $z = 0.2x + 0.3y + 10$
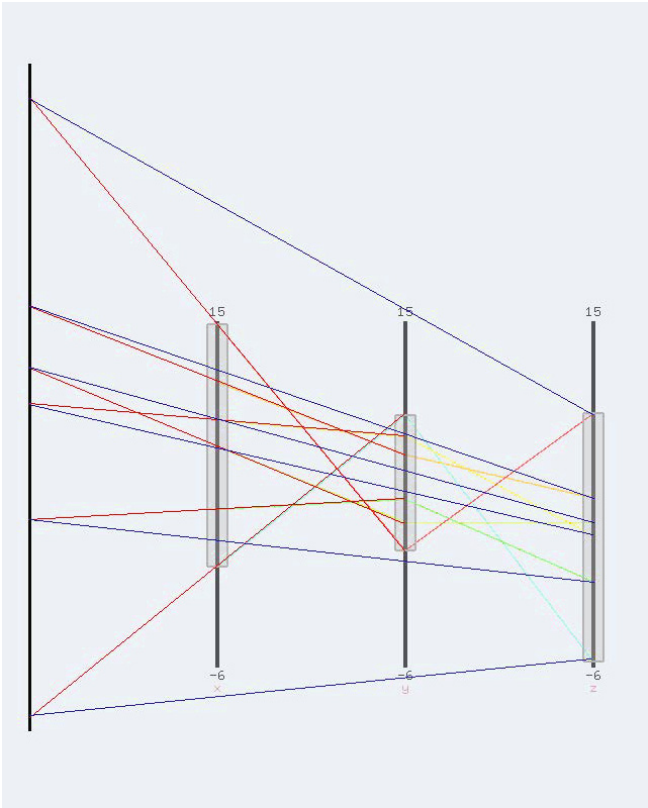
(b) $z = 0.8x - 0.4y - 2$

Fig. 1. Axis-Translation Method
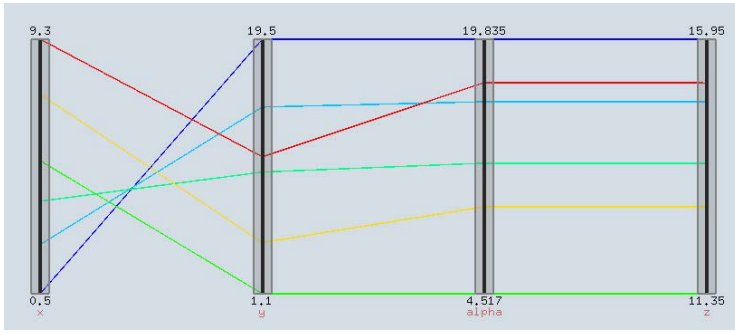
(a) $z = 0.2x + 0.3y + 10$
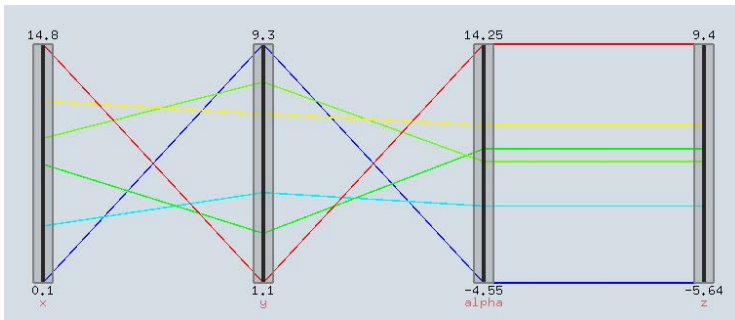
(b) $z = 0.8x - 0.4y - 2$

Fig. 2. Axis-Comparison Method

5   Mathematical Proof

First of all, we consider the case as Figure 3. There are two axes labelled as x axis and y axis. The distance between the two axes is d. Create the Cartesian coordinates system and let the horizontal axis be p, vertical axis be q. Suppose the positions of a data point on x and y are $U(n, u)$ and $V(n + d, v)$ respectively. Assume that the relationship between the two dimensions is $y = ax + b$, where a and b are constants. Therefore the data point also satisfies the equation $v = au + b$. The equation of the line UV is

$$q = \frac{v - u}{d}p + u - \frac{v - u}{d}n = \frac{(a - 1)u + b}{d}p + u - \frac{(a - 1)u}{d}n. \qquad (1)$$

If there is another data point with its positions on x axis and y axis are $U'(n, u')$ and $V'(n + d, v')$, the relationship between u' and v' will be $v' = au' + b$. The equation of the line $U'V'$ is

$$q = \frac{v' - u'}{d}p + u' - \frac{v' - u'}{d}n = \frac{(a - 1)u' + b}{d}p + u' - \frac{(a - 1)u'}{d}n. \qquad (2)$$

If $a = 1$, the slopes of the two lines are $\frac{b}{d}$, which is a constant, indicating that the two lines are parallel to each other. If $a \neq 1$, the intersection of the two lines (p_1, q_1) can be figured out as

$$
\begin{cases}
p_1 = n - \dfrac{d}{a - 1} \\
q_1 = \dfrac{bn}{d} - \dfrac{b}{a - 1}
\end{cases}
\tag{3}
$$

According to the result, the values of p_1 and q_1 do not depend on the values of u, u', v and v'. In other words, the position of the intersection point is constant. Therefore, we have proven that when two dimensions are linearly correlation with a constant proportional coefficient, the lines between them in parallel coordinates meet in one point or parallel to each other.

As illustrated in Figure 4, assume that there are three parallel axes x, y, and z in the parallel coordinates plot with the distance between two adjacent axes d. Create the Cartesian coordinates system and let the horizontal axis be p, vertical axis be q, which is the same as x axis. Suppose that there is a data point whose positions on x, y and z axes are $U(0, u)$, $V(d, v)$ and $W(2d, w)$ respectively. Let the linear relationship in the input data be $z = ax + by + c$, where a, b and c are constants, a and b are positive. Then add a new axis α to the present plot between x and y axes with the distance between x and α axes sd. Let the intersection of α and the line UV be $M(sd, m)$. When $s = \frac{b}{a+b}$, the value of m can be figured out as

$$
m = (v - u)s + u = \frac{a}{a + b}u + \frac{b}{a + b}v.
\tag{4}
$$

Therefore, the relationship between M and W is $w = (a + b)m + c$. According to the conclusion above, if $a + b = 1$, the lines between α axis and z axis which satisfy this equation will parallel to each other. Otherwise, they will meet in one point.
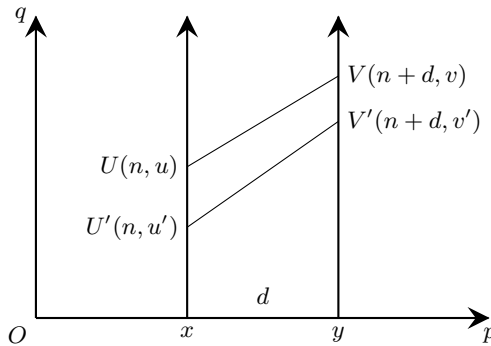
Fig. 3. The Case of Two Axes

Similarly, we can prove the conclusion when a or b is not positive. In this case, the value of s will be different and the α axis is located in the region to the left of x axis or to the right of y axis.
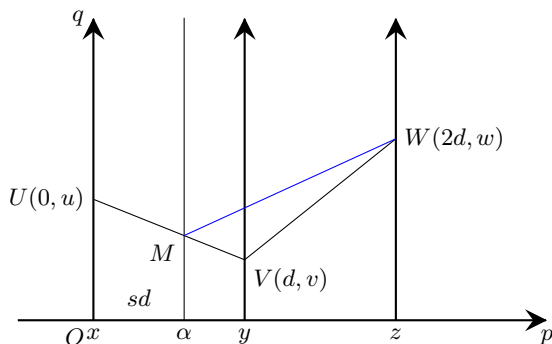
Fig. 4. Parallel Coordinates in Cartesian Coordinates

6   Conclusions and Future Work

According to the experimental results, the axis-translation method and the axis-comparison method are managed to detect the proportional coefficient when we have linear relationship among dimensions. In the future, more experiments will be done to verify and improve the methods in other cases. We will try to develop more methods to detect more types of relationships among dimensions.

References

1. d'Ocagne, M.: Coordonnées Parallèles et Axiales: Mthode de transformatio ngéométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèlles. Gauthier-Villars, Paris (1885)
2. Inselberg, A.: The Plane with Parallel Coordinates. Visual Computer 1(4), 69–91 (1985)
3. Wegman, E.J.: Hyperdimensional Data Analysis Using Parallel Coordinates. Journal of the American Statistical Association 85(411), 664–675 (1990)

VNSP: A Virtual Network Based Simulation Platform

Dawei Li, Shaojie Mao, and Lixin Zhu

Science and Technology on Information Systems Engineering Laboratory, Nanjing 210007
lidw1981@163.com

Abstract. Network based simulation is playing an important role in information technology research fields. However, some simulation activity such as information security test may cause unexpectable breakage to the network infrastructure. In order to overcome these drawbacks and provide more flexible simulation networks, a virtual network based simulation platform named VNSP is designed and implemented. Based on system virtualization technology and software routing methods, the platform has the function of real network infrastructure, can build the target network according to simulation task rapidly. This paper discussed the virtual granularity, system design and system implement of the simulation platform. In the end, a serious of experiments was carrying out to exam the performance of the platform.

Keywords: virtual router, virtual network, network simulation.

1 Introduction

Network based simulation become more and more important in information technology research since the network applications becoming popular nowadays. In order to verify the performance of newly designed protocols or methods in network environment, many experiments must be lunched in network infrastructure. However, such experiments may cause damages to the network which they running in. For example, network security experiments always do harm to network nodes and links. It is known that large scale network is difficult to configure and maintenance, so that simulations in real networks have the disadvantages of reduce the efficiency and waste a lot of resources. Build the virtual simulation platform has the remarkable significance [1].

In order to obtain the same credible simulation result in virtual networks as in real networks, the target networks built by the simulation platform must satisfy the required of network experiments in both network scale and transmission capability [2-3]. Because of the significance of virtual network simulation platform, many research institution and organization carried out many projects of the field. The successful projects include GENI[4],VIOLIN[5],VINI[6] and 4WARD[7] etc. GENI realized the virtualization during slicing up the resource,which is an opening simulation platform for large scale networks. It can create custom network simulation environment, offer resources to different users and make sure the networks' security and controllable. VIOLIN is a virtual networks environment developed by Purdue University. Users can build custom topologies according to the different tasks. The system can form the

isolatable networks environment with multiple virtual machines. VINI project assembles several soft routing modules and realizes an instance of Overlay network, the instance can interoperation with VIOLIN virtual networks mentioned above. 4WARD virtual framework allows user to build several virtual networks in a common infrastructure, and supports networks with different architectures and protocols build trusted links and interoperation with each other according to the task.

There are several virtual network test platform projects are researched in China, such as CIVIC[8] of Beijing University of Aeronautics and Astronautics, which combined many virtualization technologies and offered single isolation running environment and integrated control view to every user. The platform can shield the differences of hardware and offer common services to upper layers transparently.

However, the virtual network simulation platforms mentioned above are mostly emphasize particularly on networks running environment supporting, they are short of network dynamic design and configure mechanism, it hard to satisfy the requests of complex large scale network simulation projects. As the increasing of demands of the simulation tasks, the virtual network simulation platform should improve the functions in aspects of protocol, heterogeneous network, and virtual granularity and so on. So the research of virtual network platform enforced with large scale and high fidelity has important meanings in both theoretic and practical.

2   Network Virtualization

Virtualization technology offering separated running environment to upper level users by adding an abstract middle layer between corresponding layers. For upper layer, virtualization can shield the randomcity, heterogeneous and distribution of different system resources, offer the unified interface and services. For lower layer, virtualization can drive various platforms or hardware to response upper layers request. Virtualization platform also offers an integrated viewer to mange different virtual resources. In the same way, network virtualization forming different isolated logical networks on a same physical network using virtual technologies in order to build different network simulation environments and improve the resource utilization and network flexible [9].

Network virtualization technology can be divided into horizontal model and vertical model[10], the difference between two models are shown in Fig. 1. Horizontal model use overlay networks build virtual networks, the virtual networks connected by tunnel through real networks. The typical examples are VLAN, VPN, MPLS etc. While vertical model based to system virtualization theory, built virtual network devices (e.g. virtual router), connected and configured them to form virtual networks.

Generally, interconnections of autonomy domains through public networks are fit for horizontal model, in this situation, different heterogeneous domains connected by tunnels, is a form of N:1 virtualization. Compared with horizontal model, vertical model is a kind of 1:N virtualization. Because of the advantages of both model, almost all virtual network projects combined both of them to obtain better performance. In this paper, vertical model and soft router technology are mainly used to build the platform.
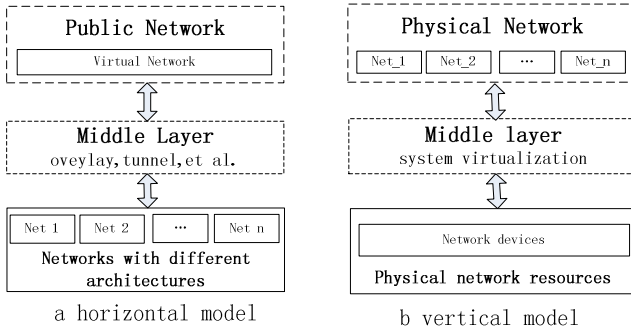
Fig. 1. Comparison of two network virtualization models

3   VNSP System Design

VNSP is designed to support all network resources' configure, rebuild and evaluate. During opening interface, modules of VNSP cooperated with each other to improve the fidelity of virtual networks. In order to satisfy the requirement of the different simulation tasks and improve the whole performance of the platform, VNSP has the ability to modulate the granularity of virtualization.

From vertical point, VNSP can be divided into 3 parts, which are physical layer, virtual layer and application layer. Physical layer consist of high performance computers and network devices that support upper level's computing and transmission request. Since all the system function will be translate to CPU instruction of the physical layer, so that physical layer usually become the bottleneck of the system. Virtual layer transfers resources of physical layer into common resources with serious granularity and offers agile configure and scheme methods. In the end of the virtual resources' life cycle, the virtual layer releases the resource to the system at the same time. Application layer call the virtual resources of virtual layer according to the simulation request and build the network environment. Applications in upper layer can not say the difference between virtual network and real network, i.e. the lower layer is transparent to its upper layer.

From horizontal point, VNSP includes hardware platform and six subsystems (Fig.2), namely, simulation design subsystem, simulation control subsystem, user control subsystem, virtual resources scheduling subsystem, display & report subsystem and simulation result analysis subsystem. All the subsystems are connected to high speed physical networks and interoperated with each other to accomplish the simulation tasks.

Hardware Platform. Hardware platform consists of a set of servers with high performance, network devices with high throughput and mass storage system. If the virtual networks scale becomes larger, the cost of system resources will increase correspondingly, so that the performance of hardware platform should be equipped with enough resources.

Simulation Design Subsystem. Simulation design subsystem offers a unique user interface to manage, control and configure all kinds of virtual resources. Simulation designer can design the needed network environment conveniently. The virtual resources which the platform offered include node type, link type and template type. Node includes computing node, network node (e.g. router, switch.) and servers etc. Link includes wire links, wireless links, interface binding and other links defined by different protocols. Template helps simulation designer to build networks with special architecture rapidly, for example, networks with some kind of topology and protocols. By defining different templates user can change typical environment to template type for later simulation environment's build. By doing this, user can create network environment rapidly, improve the efficiency of simulation.

Simulation Control Subsystem. Simulations control subsystem schedules the simulation process and monitors the status of all simulation resources, includes the cost of physical and virtual resources, the network traffic of links and so on. The system uses software tools to manage resources and realized system level control, such as reboot and initialization some dead nodes, shutdown network of abnormity and so on. In a word, virtual resources scheduling subsystem's duty is ensure the simulation task running in line and control the system on the top level.
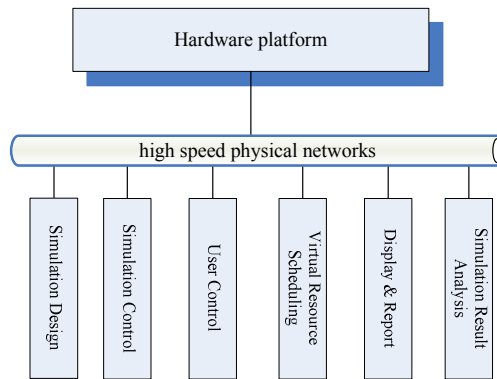
Fig. 2. The architecture of VNSP

User Control Subsystem. User control subsystem contains a GUI to simulation user to running simulation tasks. Different users have different rights according to their roles. By the help of user control subsystem, simulation users accomplish the configuration conveniently. It easy to change the topology of networks, the scale of nodes, the bandwidth and delay of links and other features of virtual networks. User control subsystem offers an interface between the simulation users and the platform, what's more, it take the behaviors of user's as a kind of input of the platform, this is useful in simulation of counterwork situations.

Virtual Resources Scheduling Subsystem. Virtual resources scheduling subsystem takes charge to backup, motion the mirror images of the virtual networks, load balancing of the networks and servers. The main function of the subsystem is distributing the

simulation resources and improves the utilization ratio of the platform. On the one hand, the subsystem analysis the resources cost and optimize the perforce according to the settings of the configuration file; On the other hand, system administrator can adjust the load balance of networks according to the requirement of the simulation to control the resources.

Display & Report Subsystem. Display & report subsystem displays the status and running information includes network topology, traffic and so on. In network based simulations, the subsystem shows the data flow, network situation, events evolvement in an intuitionistic way. The system also records the logs of the simulation, rebuild and reply the simulation process. Simulation user can use the information to analysis performance, adjust parameter and record the simulation process.

Simulation Result Analysis Subsystem. Simulation result analysis subsystem is designed to collect the simulation data and help the user to analysis the simulation data. The subsystem maintains a database and a set of statistic tools of simulation process. Simulation user can use the tools to analysis and form several kind of reports of the simulation.

4   Performance Analyses

The performance of the virtual network is mainly in aspects of ability of nodes and network transmission. Compared with real network which provided with exclusive hardware resources, the stability is the critical aspect considered in virtual network designing. Although the agility of the topology of the networks is improved, the performance of the networks is reducing at a large extent. In this section, we will discuss the performance of the virtual network via aspects of CPU/RAM costs and network traffics.

The computing and memory resources of virtual networks are divided by the physical CPU and RAM. Let $VM = \{v_1, v_2, \cdots, v_n\}$ as the set of running virtual nodes, $CPU(v_i)$, $RAM(v_i)$ ($v_i \in VM$) as the number and the size which was distributed to virtual node v_i respectively. The total resources of the system is CPU and RAM, then we have $\sum_{v \in VM} RAM(v) \equiv RAM$ and $\sum_{v \in VM} CPU(v) \equiv CPU$.

The granularity of the virtualization also affects the performance and resource costs. Under the same resources, light virtualization has strong ability to form large scale of virtual networks.

As the same reason of CPU costs, the performance of networks also depends on the ability of physical devices. Besides, the virtualization platform has extra costs to realize the different network protocols, also affect the performance of total virtual system.

Another factor makes an impact on virtual network's throughput is the distribution of the virtual nodes. As shown in Fig.3, links between node A and node B, node C and node D have perfect channel because they are in the same host. However, link between node E and node F is affected by the physical channel, the real throughput is decreased by about 50% than theoretic value. In the simulation tasks, the virtual network designer should reserve the resources in advance to reduce this affection.
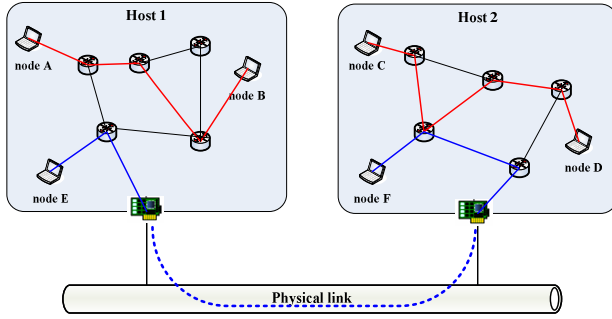
Fig. 3. Links in virtual network

In order to exam the performance of the system, we implement a prototype system of VNSP in laboratory. The computer with Intel core 2 E7500 CPU and installed 2GB DDR2 RAM, the operation system is Window server 2003 and Windows XP sp3. The virtual platform is VMware and software routers. The virtual network configured RIP protocol to route packages.

Under the situation of increasing the number of the virtual routers, we use the FTP server and client to exam the bandwidth of the network and exam the CPU and RAM cost at the same time. All the date collected is the average of 10 times.

The host's CPU utilization is increasing with the growing of the number of the virtual routers. As shown in Fig.5(a), the CPU utilization curve is similarity with a conic because all the virtual nodes share the CPU of the host, when the number of the virtual nodes increasing, additional CPU costs are needed to manage the network.
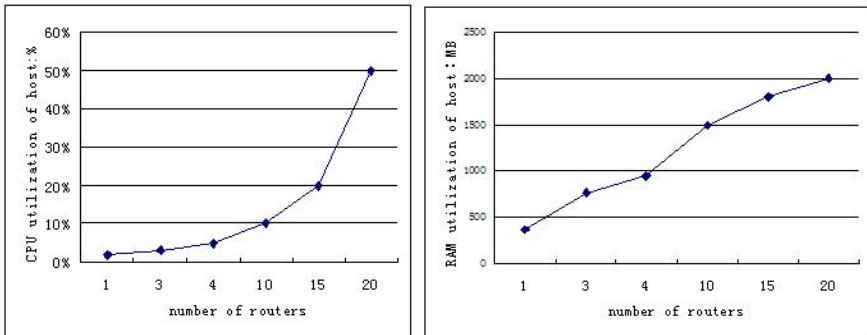
(a) CPU utilization   (b) RAM utilization

Fig. 4. Curve of resources utilization

Fig.5(b) illustrates the RAM utilization curve when the number of routers increasing. It is not hard to see that the RAM utilization is directly proportional to the number of virtual routers. This is because the platform allocates the same RAM size to same type virtual routers.
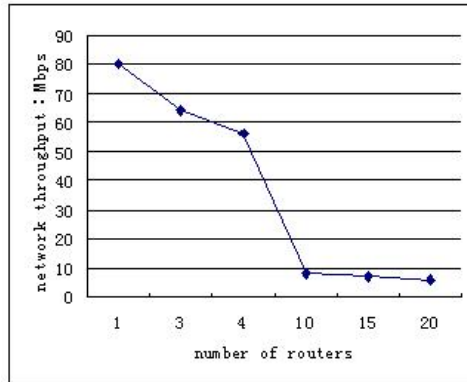
Fig. 5. Virtual network's throughput curve

We use FTP protocol to test the performance of virtual networks by putting the FTP server and client to two sides of virtual link. The results are shown in Fig.6. From the result, when the hops of virtual link below 4, the throughput is high, but when the hops increase to 10, the throughput reduced rapidly. This is because the throughput of virtual networks are affected by many facts such as distribution of nodes, transmission protocols, nodes' virtual levels, network traffic and so on. In order to obtain better network performance, we should make sure the plenty of resources in physical infrastructure.

5   Conclusions

Virtual network simulation platform offers a useful environment to network based simulation which is important in today's simulation research. The virtual platform converges physical resources of different type and distribution to uniform and isolated resources for simulation task, improved the agility and extensible of the network [11-12]. In this paper, we discussed the structure and key technologies of building virtual network simulation platform VNSP, designed and implemented the prototype system. In the end, the performances are tested by a serious of examinations.

References

1. Wang, X., Zhu, Y.: The Analysis Based on Large-scale Network Experiment Technology. Computer Knowledge and Technology 11(7), 4213–4215 (2011)
2. Qi, N., Wang, B., Wang, B., et al.: Research on Balanced Construction Algorithm of Virtual Network. Journal of Electronics & Information Technology 33(6), 1301–1306 (2011)

3. Chen, W., Xu, M., Yang, Y.: Virtual Network with High Performance: VegaNet. Chinese Journal of Computers 33(1), 63–73 (2010)
4. Wroclawski, J.: GENI: Global Environment for Network Innovations (January 08, 2011). Using the Component and Aggregate Abstractions in the GENIArchitecture, GDD-06-42 (December 19, 2006), `http://www.geni.net`
5. Ruth, P., Jiang, X.X., Xu, D.Y., Goasguen, S.: Virtual distributed environments in a shared infrastructure. IEEE Computer 38(5), 39–47 (2005)
6. About VINI: What is VINI (January 08, 2011), `http://www.vini-veritas.net/`
7. 4ward Project (2008), `http://www.4ward2project.eu/`
8. Huai, J., Li, Q., Hu, C.: Research and Design on Hypervisor Based Virtual Computing Environment. Journal of Software 8(18), 2016–2026 (2007)
9. Jiang, M., Wang, B., Wu, C., et al.: Research on Network Virtualization and Virtual Network Mapping Algorithm. Acta Electronica Sinica 39(6), 1315–1320 (2011)
10. Zhu, W., Sun, M.: Research on Network Virtualization. Telecommunications Network Technology 9, 19–23 (2010)
11. Feng, Y.: Construction of Enterprise Information Management and Control System. Command Information System and Technology 3(2), 38–43 (2012)
12. Wang, W., Wang, D., Yue, C., et al.: Information Collection Technology for Simulation Training. Command Information System and Technology 3(2), 67–71 (2012)

A Multi-resolution Display Method for Digital Archives of Cultural Properties Using Polygon Reconsruction

Megumi Okumoto[1], Yuri Iwakata[1], Asuka Komeda[1], Tomoko Yonezawa[2], Masami Takata[1], and Kazuki Joe[1]

[1] Nara Women's University, Nara, Nara, Japan
[2] Kansai University, Takatsuki, Osaka, Japan

Abstract. In this paper, we propose a multi-resolution display method for digital archives of cultural properties using polygon reconstruction. When digital archives of cultural properties are displayed in VR systems for research purpose, mesh resolution had to be changeable on the demand of user. Additionally, for user's comfortability, polygon reconstruction needs to be executed in real-time. History of preliminary polygon reductions is adopted for the proposed method so that polygon reconstruction is executed in real-time. To validate the VR system using the polygon reconstruction based multi-resolution display method, some experiments are performed. As the result of the experiments, the proposed method is performed within 1.0 second, which is considered as "real-time" by the definition of typical user-interface.

1 Introduction

In recent years, digital archives of cultural properties are performed aiming at cultural property protection [1]. To display special purpose 3D objects on Virtual Reality (VR) systems, when they can be analyzed intuitively and easily, it is possible for those 3D objects to be applied to cultural property research. It is sometimes difficult for users who are not familiar with VR to control objects displayed in VR according to their intention because they must understand the method for the operations. Almost users who are interested in the research of cultural property are sometimes very inexperienced in VR or even computer operations. Therefore, to control VR objects for such users, operations of VR systems should be intuitive and very easy.

Motions are one of intuitive human's operations. Kinect [2] is a device for capturing human's motions. Kinect can detect human's motions in real-time. In addition, Kinect is cheaper than existing VR interface devices such as data gloves [3], which capture the motions of human's hands. In this paper, we develop a VR system using Kinect as a VR interface device.

When a 3D object is displayed in a VR system, it is represented using a polygon model, which consists of a lot of triangle meshes. To faithfully describe the detail of a cultural property, its digital archive becomes extremely complex

and huge, and the number of meshes of the polygon model is too many to be efficiently displayed in VR systems. As the number of the meshes increases, the operability of VR systems deteriorates. To avoid the problem, polygon reduction is applied to such polygon models to be displayed in VR.

By decreasing the number of polygons, some information of the cultural property may be lost in detail. Using a reduced polygon model, details may not be observed by the zoom of the 3D object since some important features of the original polygon model may be deformed to be a flat surface. When users analyze the 3D object of a cultural property, the polygon model should consist of the large enough number of meshes. To use digital archives of cultural properties for research, which are displayed in VR, it is required that the number of meshes is changeable on demand from user. Consequently, in this paper, we propose a real-time polygon reconstruction method. To apply this method to a VR system with Kinect, the VR system is intuitive and easy for user.

The rest of the paper is organized as follows. In Sect. 2, polygon reduction is introduced. In Sect. 3, a VR system using Kinect is described. In Sect. 4, a real-time polygon reconstruction method is proposed. In Sect. 5, to evaluate the performance of the proposed method , some experiments are explained.

2   Quadric Error Metrics

Quadric Error Metrics (QEM) [4] is a polygon reduction algorithm. Let a polygon model consist of a vertex set V ($V \ni V_i : i = 1, 2, 3 \dots$). Then the flow of polygon reduction to V by using QEM is shown as follows:

Step a-1   Compute a 4×4 symmetric matrix Q for each initial vertex.
Step a-2   Select a pair of vertices (V_{j1}, V_{j2}) that constructs a ridge line.
Step a-3   Compute the optimal contraction target V_k for each valid pair.
Step a-4   Compute the cost for each vertex V_k.
Step a-5   Sort all the pairs by cost in ascending order.
Step a-6   Delete the pair (V_{j1}, V_{j2}) with the smallest cost and create V_k.
Step a-7   When the target number of vertices is obtained, terminate Step a.
Step a-8   Chane the pair information relevant to the vertex V_{j1} and V_{j2} to the new vertex V_k.
Step a-9   For all pairs relevant to the vertex V_k, compute each optimal contraction target and cost.
Step a-10 Return to Step a-5.

Figure 1 shows the basic concept of QEM.

In Step a-1, a symmetric matrix Q for each vertex is derived from the below equation (1) of the planes $planes(V_{j1})$ which meet at the vertex.

$$
Q = \begin{bmatrix} q_{1,1} & q_{1,2} & q_{1,3} & q_{1,4} \\ q_{1,2} & q_{2,2} & q_{2,3} & q_{2,4} \\ q_{1,3} & q_{2,3} & q_{3,3} & q_{3,4} \\ q_{1,4} & q_{2,4} & q_{3,4} & q_{4,4} \end{bmatrix}
$$

$$
= \sum_{planes(V)} K_p. \tag{1}
$$
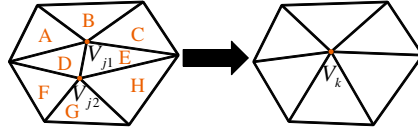
Fig. 1. The over view of QEM

The plane is defined as the following equation (2).

$$ax + by + cz + d = 0 \quad \left(a^2 + b^2 + c^2 = 1\right).$$

(2)

Then, the matrix K_p is obtained as the below expression (3).

$$K_p = \begin{bmatrix} a^2 & ab & ac & ad \\ ab & b^2 & bc & bd \\ ac & bc & c^2 & cd \\ ad & bd & cd & d^2 \end{bmatrix}.$$

(3)

For example, let take a look at the vertex V_{j1} in Fig. 1. The plane $planes(V_{j1})$ which meet at that vertex V_{j1} include the planes A, B, C, D and E. V_{j1} and V_{j2} in Step a-2 are arbitrary vertices in the vertex set V. At a vertex $V_i = [v_x, v_y, v_z, 1]^\top$, the cost ΔV_i is defined as

$$\begin{aligned} \Delta V_i &= V_i^\top Q_i V_i \\ &= q_{1,1}x^2 + 2q_{1,2}xy + 2q_{1,3}xz + 2q_{1,4}x \\ &\quad + q_{2,2}y^2 + 2q_{2,3}yz + 2q_{2,4}y + q_{3,3}z^2 \\ &\quad + 2q_{3,4}z + q_{4,4}, \end{aligned}$$

(4)

where v_x, v_y and v_z are the x, y and z coordinate values of V_i, respectively. In ΔV_i, when V_i is unknown and Q_i is given, V_k is found by solving the following partial differential equation.

$$\frac{\partial \Delta}{\partial x} = \frac{\partial \Delta}{\partial y} = \frac{\partial \Delta}{\partial z} = 0.$$

(5)

$$\begin{bmatrix} q_{1,1} & q_{1,2} & q_{1,3} & q_{1,4} \\ q_{1,2} & q_{2,2} & q_{2,3} & q_{2,4} \\ q_{1,3} & q_{2,3} & q_{3,3} & q_{3,4} \\ 0 & 0 & 0 & 1 \end{bmatrix} V_k = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

(6)

The bottom row of the left-hand side matrix is empty because V_i is an homogeneous vector. By using the equation (6), an optimal vertex is found in Steps a-3. The symmetric matrix Q_k is expressed as the sum of two symmetric matrices Q_{j1} and Q_{j2}, which are for V_{j1} and V_{j2}, respectively.

$$Q_k = Q_{j1} + Q_{j2}.$$

(7)

In Step a-3, if the left-hand side matrix of the equation (6) is not invertible, an optimal vertex V_k is found along the curved segment $V_1 V_2$. In Step a-4, by using the equation (4), ΔV_k is computed. The cost ΔV_k is the error to merge V_k to the tangent. In Step a-5, the costs are sorted in ascending order, and the pair (V_{j1}, V_{j2}) with the minimum cost ΔV_k is removed to add a new vertex V_k the vertex set V in Step a-6. In Step a-7, computing the number of polygons, it proceeds to Step a-8 when the number of polygons does not reach to the target number. In Step a-8 and Step a-9, a new vertex V_k and the pairs of the vertex with which polygon is tangent to V_k are created.

3   A Virtual Reality System Using Display Control by Kinect

In this paper, we develop a VR system using Kinect as a VR interface device. Kinect can detect various motions of human's skeleton. So user can easily operate and control objects displayed in a VR system by using the recognition ability of Kinect. Figure 2 shows a conceptual image of the developing VR system with Kinect. The flow of the VR system operations using Kinect is shown as follows:

Step b-1   Display objects on VR.
Step b-2   Capture user's motion by Kinect.
Step b-3   Update the objects as control requests from the captured motions.
Step b-4   Return to Step b-1.
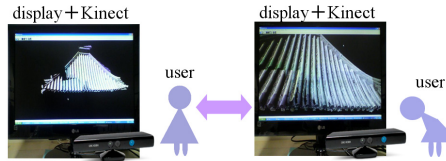
Fig. 2. Conceptual image of a VR system using Kinect

In Step b-1, objects are displayed on VR. OpenGL is used to render the target objects on VR. In Step b-2, user's motions are captured by Kinect. In this paper, user's motions are assumed as three patterns: upstanding, bowing and deep bowing. Figures in Tab. 1 show the three patterns. There is a depth sensor equipped with Kinect which measures the distance to the user. Using the depth sensor, the gradient angle of user's upper body is calculated from two distances to user's shoulder and waist. In Step b-3, the captured motion in Step b-2 is sent to the VR system so that the user request (e.g. changing the object size) is reflected to the displayed objects in VR. Table 1 shows the corresponding table between captured motions and object resolutions. In the case of upstanding, the object size and resolution is set to normal. In the case of bowing, objects are scaled up. By scaling up, objects are visible in detail. In the case of deep bowing, objects are furthermore scaled up. In this way, these motions are applied to the target object so that three kinds of resolution for the object are selectable.

Table 1. Captured motions and object resolutions

the inputted motion			
	Upstanding	Bowing	Deep bowing
changes of the size	standard	scale-up	more scale-up
changes of the number	standard	detail	more detail

4   Implementation of Real-Time Polygon Reconstruction

Sect. 3 describes that resolution in a VR system changes according to user's motions. Using the VR system, 3D objects with the polygon model that allows multiple resolutions are available. In the polygon reduction method described in Sect. 2, when the number of meshes in the polygon model is large, the polygon reduction method becomes compute intensive. Therefore, it is not effective that polygon reduction is performed every time the polygon model changes. In this paper, we propose a real-time polygon reconstruction method for digital archives of cultural properties. Using the reconstruction method, a digitalized cultural property is shown in a VR system. The number of meshes constructing the digital archive is changed according to the three patterns described in Sect. 3. In addition, polygon reconstruction needs to be executed in real-time so that user feels comfortable to operate the VR objects. In [5], the meaning of real-time for interactive systems is defined as less execution time than 1.0 second that is the time user can wait without so much stress.

The flow of real-time polygon reconstruction is shown as following steps c.

Step c-1   By using QSlim, create a history of polygon reductions which are to be used for the polygon reconstruction.
Step c-2   Display a 3D object with the current polygon model on VR.
Step c-3   Capture user's motion by Kinect.
Step c-4   According to the captured user's motion, change the polygon model to another model from the polygon reduction history.
Step c-5   Return to Step c-2.

To execute real-time polygon reconstruction, Step c-1 should be performed in advance because of the long calculation time for polygon reduction. Changing polygon models can be quickly done with the polygon reduction history, where QSlim is applied to each polygon reduction. QSlim is an application for polygon reduction based on QEM described in Sect. 2. Items saved in the polygon reduction history are shown as follows:

- The ID numbers of two vertices, which are both endpoints of a collapsed edge.
- Distance between the newly created vertex and the corresponding deleted vertex.
- The number of planes which meet at one of the both endpoints.
- The polygon numbers of collapsed edges.
- The number of planes affected by collapsing an edge.

In Step c-2, a polygon model is displayed on VR. The initial setting for the polygon model has the smallest number of meshes because the initial motion is upstanding. The mesh information of the polygon model is saved as a list structure. In Step c-4, user's motion is captured by Kinect. The captured motions in this paper are the same as described in Sect. 3: upstanding, bowing and deep bowing. In Step c-4, the polygon model is changed according to the obtained user's motion. The change of polygon models in the proposed system works at the same time with scaling, which is described in Sect. 3.

Table 1 shows the relationship between the inputted user's motion and the number of polygons. In the case of upstanding, the number of meshes is defined as the normal. In the case of bowing, the number of meshes increases. By increasing the number of meshes, a more detailed polygon model is adopted. In the case of deep bowing, the number of meshes furthermore increases and a much more detailed polygon model is employed. In other words, inputting three motions, three kinds of resolution in polygon models are available. Polygon reconstruction is based on the polygon reduction history. When decreasing meshes, meshes to be deleted are labeled $false$, and are removed from the list structure. Then, the position of a vertex V_k is calculated. As shown in Fig. 1, the position of the vertex V_k which is relevant to the deleting mesh is calculated by adding the position of the vertex V_{j1} to the distance retrieved from the history. The edges related to the vertex V_{j1} and the vertex V_{j2} is replaced by vertex V_k.

When increasing meshes, meshes to be added are labeled $true$, and are added to the list structure. Then, the positions of two vertices V_{j1} and V_{j2} which are relevant to meshes to be added are calculated. As shown in Fig. 1, the position of the vertex V_{j1} is calculated by subtracting the distance, which is retrieved from the history, from the position of the vertex V_k. The position of the vertex V_{j2} is obtained by retrieving from the history. The edges related to V_k are replaced by V_{j1} and V_{j2}.

The real-time polygon reconstruction procedure repeats the above steps until a pre-defined user's motion is detected.

5   Experiments

In this section, the results of two experiments are explained for validating the proposed real-time polygon reconstruction method.
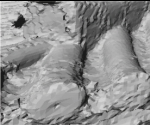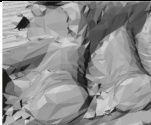
The experimental data is a part of 3D coordinate measuring data of the five-story small pagoda (national treasure of Japan) enshrined in Kairyuouji temple.

5.1   Computation Time for Polygon Reduction

As the first experiment, the calculation time of polygon reduction is measured. The polygon reduction is carried out in advance of polygon reconstruction. In polygon reduction, the number of meshes for a polygon model is reduced in two steps. In the first step, the number of meshes is reduced from 1,6553,199 meshes to 330,637 meshes. In the second step, the number of meshes is reduced from 330,637meshes to 33,064 meshes. These steps are defined Reduce-1 and Reduce-2, respectively.

Table 2 shows calculation time for Reduce-1 and Reduce-2, where Setup, Init, Run and Output represent file loading, command line parsing, pre-processing, polygon reducing and history file writing times, respectively. Three figures in Tab. 2 are polygon models consisted of 1,653,199, 330,637 and 33,064 meshes, respectively. When the number of meshes for a polygon model is small, the polygon model is coarse.

Table 2. Required time for polygon reduction [sec.]

	Recuce-1	Recuce-2
Setup	135.18	67.31
Init	29.45	14.91
Run	31.30	23.72
Output	10.98	9.61
Total	206.92	115.54

In Tab. 2, it takes 29.4 seconds and 14.9 seconds for Reduce-1 and Reduce-2 of Init, respectively. Sect. 2 describes Init includes Step a-1 to Step a-5. In Init, as pre-processing, the optimal vertex is selected and all the pairs of edges are sorted in descending order. Since the calculation time is not affected by the number of reduced polygons but the number of inputted polygons, the difference between Reduce-1 and Reduce-2 occurs. It takes 31.2 seconds and 23.7 seconds for Reduce-1 and Reduce-2 of Run, respectively. Sect. 2 describes Run includes Step a-6 to Step a-10. In Run, meshes of a polygon model are reduced. The optimal vertex for edges, which is related to reducing meshes, is computed. Then, all pairs of edges, which are related to the reducing meshes, are sorted in descending order by cost. Since the number of reductions in Reduce-1 is larger than in Reduce-2, it takes longer time for Reduce-1.

As the result of the first experiment, since polygon reduction takes long time, it is unsuitable to execute polygon reduction during changing polygon models. Therefore, in the proposed system, polygons should be restructured with a history of polygon reduction.

5.2   Elapsed Time for Polygon Reconstruction

In the second experiment, the elapsed times for file read and polygon model change in the developed polygon reconstruction method is measured. Figure 3 shows the scenery of the second experiment. Figure 3-(a), (b) and (c) show examinee's motion of upstanding, bowing and deep bowing, respectively. It is observed that the resolution of each polygon model is changed according to the motions. Additionally, examinee can intuitively operate the developed system by using Kinect. Therefore, it is easy for user to operate this system without any pre-training for the operation.

(a) Upstanding   (b) Bowing   (c) Deep bowing
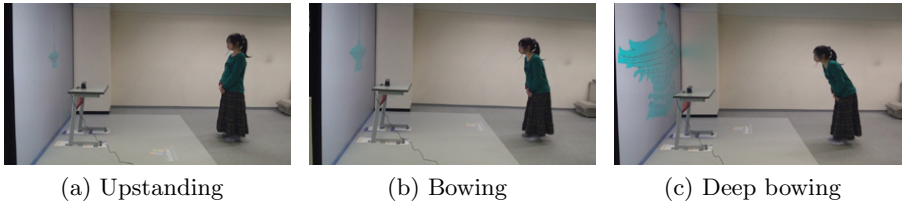
Fig. 3. The scenery of the second experiment

Table 3 shows various times in the second experiment, where Setup and Change indicate file loading and polygon model change for given examinee's motions, respectively. As Tab. 4 shows, any changes of meshes require less than 1.0 second. Since the definition of real-time in this paper is 1.0 second, the response time for the developed system is sufficient.

Table 3. Required Time for the Proposed System [sec.]

Setup		181.703		
		a	b	c
change	scale-up	0.813	0.781	
	scale-down	0.782	0.765	

6   Conclusions

In this paper, we proposed a polygon reconstruction method for digital archives of cultural properties. Presenting special purpose 3D objects on VR systems, they can be analyzed intuitively and easily. Such application is suitable for cultural property research. It is sometimes difficult for users who are not familiar with VR to control the objects displayed in VR according to their intention because they must understand the method for the operations. Therefore, operations of VR systems should be intuitive and very easy so that such users control VR objects. In this paper, we developed a VR system using Kinect as a VR interface device.

To use digital archives of cultural properties in VR for research purpose, it is required that mesh resolution is changeable on the demand of users. Additionally,

polygon reconstruction needs to be run in real-time so that users feel comfortable for their demands. In this paper, we proposed a real-time polygon reconstruction method using history of preliminary polygon reductions. To validate the proposed system, the polygon reconstruction method is evaluated. As the result of the experiment, the polygon reconstruction is performed within 1.0 second, which is considered as "real-time" by the definition of typical user-interface.

Our future work includes the implementation of view-dependent refinement. In the proposed method, all polygons are reconstructed. However, it is inefficient to reconstruct back-polygons and other polygons located outside of display. By implementing view-dependent refinement, polygon models can be displayed more rapidly. Additionally, to develop the more comfort system for users, the proposed method should be speeded up.

References

[1] Clark, J.T., Slator, B.M., Bergstrom, A., Larson, F., Frovarp, R., Landrum III, J.E., Perrizo, W.: Preservation and access of cultural heritage objects through a digital archive network for anthropology. In: Proceedings Seventh International Conference on Virtual Systems and Multimedia, pp. 28–38 (2001)

[2] Kinect-Xbox.com, `http://www.xbox.com/en-US/kinect/`

[3] Greenleaf, W.J.: Developing the tools for practical VR applications. Engineering in Medicine and Biology Magazine, 23–30 (March/April 1996)

[4] Garland, M., Heckbert, P.S.: Surface simplication using quadric error metrics. In: SIGGRAPH 1997 Proceedings, pp. 209–216 (1997)

[5] Nielsen, J., Pernice, K.: Eyetracking Web Usability. New Riders Press (December 2009)

A Prototype Architecture for Assembly-Oriented Cyber-Physical Systems

Jietao Dong, Tianyuan Xiao, and Linxuan Zhang

National CIMS Engineering Research Centre, Department of Automation,
Tsinghua University, Beijing, China
{dongjt10,xty-dau,lxzhang}@mails.tsinghua.edu.cn

Abstract. ACPS is a new generation intelligent system based on the embedded systems, sensing technology and networked connectivity. As a typical application of CPS, it has many special features different from the traditional systems. According to these features, a four-layer architecture is proposed which consists of Physical Layer, Network Layer, Co-processing Layer and Application Layer. Furthermore, some research directions and challenges of ACPSs are suggested, and the ultimate goal is to develop foundations and techniques for building safe and effective ACPSs.

Keywords: Assembly systems, Cyber-physical systems (CPSs), Four-layer Architecture.

1 Introduction

The assembly industry is undergoing huge changes, with the rapid development of embedded systems, sensing technology and networked connectivity. Embedded systems provide devices with intelligence while sensing equips them with perception and networked connectivity gives them the ability of communication [1]. Other than stand-alone devices that can be designed, certified and used independently of each other to accomplish special assembly steps, we will be faced in the near future with distributed systems that simultaneously monitor and control the whole assembly process, which make it more intelligent and effective. The combination of embedded software controlling the devices, networked systems of the distributed assembly devices and objects, and complicated physical information from the assembly process make modern assembly systems a distinct class of cyber-physical systems (CPSs), which we refer to as assembly-oriented cyber-physical systems (ACPSs).

To develop safe and effective ACPSs, new design, verification and validation techniques will be required, as the system size and complexity keep increasing. Moreover, model-based technology should play an important part in the ACPSs design. Models should cover assembly devices and objects as well as communications between them.

Additionally, the regulatory assembly procedures must be redefined so as to approve their use for assembly systems, because the traditional process-based regulatory regime has already been too lengthy and extremely expensive with the increased ACPSs complexity.

In this paper, some of the main challenges and research directions in building ACPSs are described. The ultimate goal is to develop foundations and techniques for building safe and effective ACPSs. Section 2 provides a conceptual overview of ACPSs and the principal challenges. Section 3 analyses the features of ACPS and describes a prototype architecture as well as its components. Finally, section 4 concludes this paper.

2   The Conceptual Overview of ACPS

ACPSs are safety-critical, interconnected, intelligent systems of assembly devices and objects. Traditional assembly scenarios can be viewed as closed-loop systems where workers are the controllers, assembly devices act as sensors and actuators, and assembly objects are "plants" [2]. Fig. 1 shows the conceptual overview of ACPSs. The devices used in ACPSs can be categorized into two large groups based on their primary functionality: sensing devices, such as GPS and RFID, which real-timely perceive the assembly process and provide information of assembly objects, such as the position and posture; and control devices, such as manipulator and robotic arm, which execute the operation order to promote the smooth implementation of assembly process. In APCSs, the sensing devices can transmit the collected information to a data management center or decision support, each of which plays a different, but complementary role. Overall, these devices are the interface between the physical assembly system and the cyber decision support system.
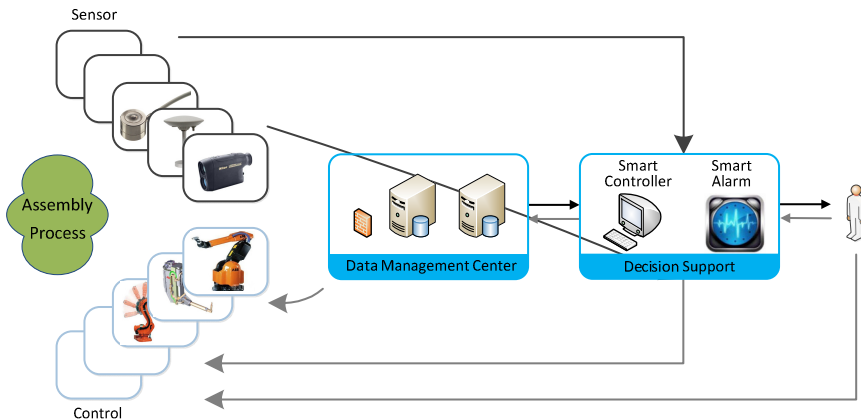
Fig. 1. The Conceptual Overview of Assembly-Oriented Cyber-Physical Systems

In ACPSs, Data Management Center can collect the data from the sensing devices and deliver it to Decision Support after some necessary preprocessing. Meanwhile, it will record the assembly steps, so it will be convenient to examine the assembly history.

Decision Support can process the data collected and generate operation orders according to certain rules. On the other hand, it can also generate alarms for emergencies, such as equipment failure or short supply of materials.

3   Architecture for ACPS

3.1   Features of ACPS

The goal of ACPSs is to deeply integrate the physical assembly and the cyber process [3], and then improve our safe, real-time, reliable monitor and control of the assembly process. ACPSs are a typical application of CPS on traditional manufacturing industry, they are different from desktop computing, traditional embedded/real-time systems, wireless sensor network (WSN), etc. and they have some special characteristics as follows [4-6].

- *Closely integrated.* The essence of ACPSs is the integration of computation and physical process.
- *Event driven.* The change of assembly conditions forms the ACPSs event, and then it finally change the assembly state through the closed loop of "event -> sensor -> decision -> control -> event".
- *Heterogeneous.* ACPS contain many subsystems of different functions and frameworks, they interact through wire or wireless communication. So ACPSs can be viewed as systems of systems.
- *Concurrency.* Physical processes are compositions of many things occurring at the same time, unlike software processes, which are deeply rooted in sequential steps.
- *Sensitivity to time.* Time of the physical world is irreversible so that ACPSs set higher requirements on timeliness, the delay may affect the users' estimate and decision. So, the time it takes to perform a task is an issue of correctness, not just performance like the general-purpose software.
- *High degrees of automation.* ACPSs can completely support self-sensing, self-determination and self-control, so they must be capable of dynamically self-reorganizing, self-reconfiguration, self-maintenance, self-optimization and self-protecting.
- *Control loops must be closed.* ACPSs use feedback to assure the accuracy of control, and sometimes the worker is a node in the control loops.
- *Operations must be dependable, and certified in some cases.* It's necessary for ACPSs, as the operations will influence the real assembly, a minor mistake may cause enormous loss.

3.2   A Prototype Architecture for ACPS

Based on the research of CPS, the structure is mainly of hierarchy and bus [7-9], here a four-layer architecture for ACPS is proposed, which fairly reflects the features of ACPS. As Fig. 2 shows, it contains four layers: Physical Layer, Network Layer, Co-processing Layer and Application Layer.
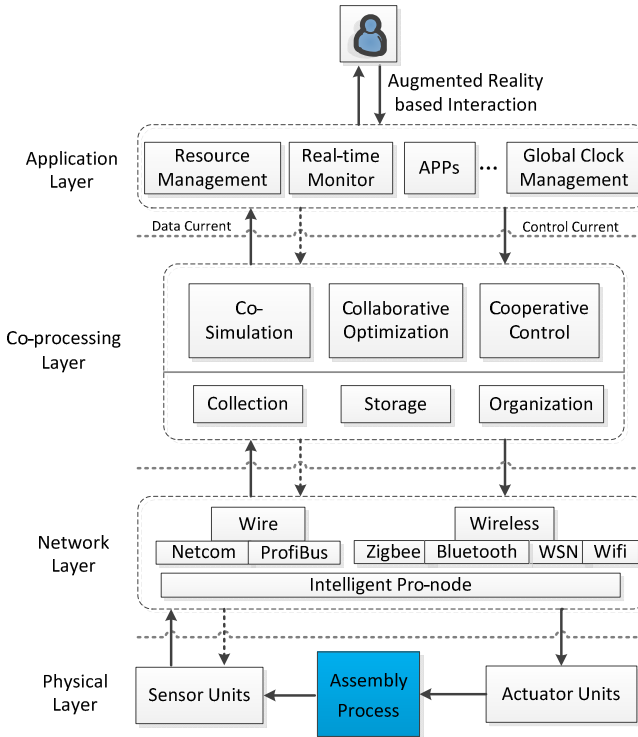
Fig. 2. A Prototype Architecture of Assembly-Oriented Cyber-Physical Systems

Physical Layer (PL) is the fundamental layer, which consists of distributed sensor units and actuator units as well as the assembly objects. Sensors and actuators provide the interface between the physical world and the cyber world. A sensor is a device that is able to measure physical conditions, e.g. torque size, and transform the physical conditions into information, which includes the attributes, sampling timestamp and spacestamp. On the other hand, an actuator is a device that is capable of changing attributes of an assembly object, e.g. move a part along a certain direction with a certain distance. Generally, one kind of sensor or actuator is usually connected with a unique physical condition or attribute.

Network Layer (NL) makes the ACPSs units interconnected with each other, which is the foundation of data interaction, resource sharing and interoperability. Unlike the traditional network, this network needs to overcome the heterogeneity of ACPSs units so as to provide users with plug and play service. Therefore, ACPS network not only contains the traditional technologies, such as route, access control, and publish/subscribe mechanism, but also some new challenges, e.g. presentation of heterogeneous data. More importantly, current network is mainly about data transmission ignoring time delay while ACPSs require high level of real-time performance. For example, a packet must be transmitted within some time beyond which it will be invalid. So the transmission with time constraint will be one of the main challenges.

Co-processing Layer (CL) is essentially a data management center in some degree, which is mainly responsible for the storage, organization and management of the data received from NL, aimed at supplying the right application with the right data at the right time. On the other hand, it can also do some preprocessing of the collected data, and provide decision support for the users. In detail, when the position and posture of assembly parts is captured, it can automatically compute the feasible paths through co-simulation and select the best one through collaborative optimization. All the results will be submitted to the users and then an order is returned correspondingly, which is interpreted according to some predefined rules , so as to control the actuator units operating cooperatively.

Application Layer (AL) supports resource management, real-time monitor and global clock management. Through global clock, the whole system can be identified with uniform time stamp to assure the correct event sequence and task deadline, which reflects the strict timeliness of ACPS. Furthermore, AL can provide the users with individual services and send user orders to the CL. It encapsulates the bottom information in various application modules so that the users can manipulate processing ignoring the bottom details. Finally, it also offers interactive interface and users can observe the visual assembly process and interact with the cyber world through TV, PC or mobile devices. Here we choose augmented reality (AR) because it's more direct and lifelike.

4   Conclusions

This paper applies CPS to the traditional assembly industry, and proposes a new system named ACPS to improve the safe, real-time and reliable monitor and control of assembly process. Unlike other CPS domain, it presents special characteristics related with the assembly industry, based on that a four-layer architecture is suggested, which consists of PL, NL, CL and AL.

However, there are still many open research challenges for this architecture. For example, how can physical assembly events be represented in the cyber world, and how can actuators operate accordingly? [10]How can event model reflect the character of temporality and spatiality? Data management and dispatching are other challenges, while security issues pose a challenge for any system.

References

1. Lee E.A.: Cyber-Physical Systems-Are Computing Foundations Adequate? (2006)
2. Lee, I., et al.: Challenges and Research Directions in Medical Cyber-Physical Systems. Proceedings of the IEEE 100(1SI), 75–90 (2012)

3. Lee, E.A.: Cyber Physical Systems: Design Challenges. In: 2008 11th IEEE International Symposium on Object Oriented Real-Time Distributed Computing (ISORC), pp. 363–369. IEEE (2008)

4. Shi, J., et al.: A survey of Cyber-Physical Systems. In: 2011 International Conference on Wireless Communications and Signal Processing (WCSP), pp. 1–6 (2011)

5. Wan, J.F., et al.: Advances in Cyber-Physical Systems Research. KSII Transactions on Internet and Information Systems 5(11), 1891–1908 (2011)

6. Wu, F.J., Kao, Y.F., Tseng, Y.C.: From wireless sensor networks towards cyber physical systems. Pervasive and Mobile Computing 7(4), 397–413 (2011)

7. Tan, Y., Goddard, S., Perez, L.C.: A prototype architecture for cyber-physical systems. ACM SIGBED Review 5(1), 26 (2008)

8. Dillon, T.S., et al.: Web-of-things framework for cyber-physical systems. Concurrency and Computation-Practice & Experience 23(9SI), 905–923 (2011)

9. Koubâa, A., Andersson, B.: A vision of Cyber-Physical Internet. CISTER Research Unit, Polytechnic Institute of Porto, ISEP/IPP (2009)

10. Poovendran, R.: Cyber-Physical Systems: Close Encounters Between Two Parallel Worlds. Proceedings of the IEEE 98(8SI), 1363–1366 (2010)

A Robust Physics-Based 3D Soft Tissue Parameters Estimation Method for Warping Dynamics Simulation

Xiangyun Liao, Zhiyong Yuan[*], Zhaoliang Duan, Weixin Si, Si Chen, Sijiao Yu, and Jianhui Zhao

School of Computer, Wuhan University, Wuhan, Hubei 430072, China
{xyunliao,dzlwhu,wxsics,cnlingmocs,sijiaoyucswhu}@gmail.com,
zhiyongyuan@whu.edu.cn, jianhuizhao@whu.edu

Abstract. Soft tissue warping is one of the key technologies in dynamic simulation of many surgical procedures. To achieve high performance simulation of 3D soft tissue warping, the research of physical parameters estimation of the warping model is of great significance. Through the construction of parameters estimation platform which consists of an optical tracking system PPT2 (Precision Position Tracker with 2 Cameras) and pressure acquisition devices, we obtain the nodal displacements of tetrahedron finite element model and external forces on it. Then we calculate the parameters of 3D soft tissue by using reverse engineering method and verify the parameters by comparing the calculated nodal displacements and the measured nodal displacements of the soft tissue. The experimental results show that the Physics-based 3D soft tissue parameters estimation method we proposed have achieve accurate agreement of calculated nodal displacements and the measured nodal displacements and it has the properties of accuracy and robust;

Keywords: soft tissue warping, Physics-based 3D soft tissue parameters estimation method, reverse engineering.

1   Introduction

Surgery simulation based on virtual reality has wide applications and development perspectives in the future development of medicine. In the research of virtual surgery simulation, many researchers provided real-time simulation and rendering technology based on visual feedback physically-based deformations and stunning simulations. 3D soft tissue often has complicated elastic behavior due to its biological nature [1]. To simulate the physical state of 3D soft tissue more realistically, we must pay attention to the parameters estimation and representation of 3D soft tissue.

There are many ways of 3D soft tissue parameters estimation [2-5]. Biomechanical model is of high accuracy, while it is very complex because of the requirements of precise geometrical model and accurate estimation of the external force [6]. Data driven method has a wide application in computer graphics and the researchers can

---

[*] Corresponding author.

get stress and strain value more easily from the parameters estimation data, but the material properties of this method will also be changed as the changes with time [7]. Recently some research teams have minimally invasive parameters estimation test in animals and humans [8][9], estimating the nonlinear properties and changes with time. Some researchers accurately reconstruct the motion of the surface of the body by applying the three-dimensional trajectories for the dense marker set to a subject-specific polygonal model [10]. All these methods need a kind of inverse solution to obtain the unknown materials properties of soft tissue. The difficulties of this method are the accurately representation of the characteristics and the high computational complexity of the finite element model. Therefore, we need to study a precise and robust method for measuring the parameter of soft tissue.

In this paper, we present a precise and robust method of 3D soft tissue parameters estimation method. First, we utilize a motion capture system called PPT2 to obtain the mark point on 3D soft tissue. Then we obtain the pressure exerted on the experimental material through the pressure acquisition devices and calibrate the pressure sensor with BP neural network. After discretization of 3D soft tissue, we construct a linear tetrahedron finite element model and obtain nodal displacements through the PPT2, and we calculate the parameters of 3D soft tissue using reverse engineering method.

2   Parameters Estimation Platform Construction

2.1   Acquisition of Nodal Displacements

We utilize an optical tracking system PPT2 to obtain the warping position set of test materials. PPT2 tracks large areas (more than 10 x 10 meters) while maintaining high precision and accuracy. It presents excellent value for the real-time demands of immersive virtual reality applications. After calibrating the cameras of PPT2, we manually mark on the test material and extract the pixel coordinates of feature points. Optical sensors mounted in the corners of a room track active LED markers. As sensors make their way through the room, the sensors acquire data. Through rapid processing, the sensors' data are converted to the accurate three-dimensional location of the markers.

2.2   Pressure Acquisition

The major task of pressure acquisition is to obtain the pressure exerted on the experimental material through some devices. These devices include SDI-2F miniature pressure sensor and AT91SAM7S64-based application board.

We collected nearly a thousand set of samples. Each sample consists of seven data that is the actual pressure and six voltages measuring by sensor. The actual pressure is obtained by the electronic scale with accuracy of 2g. Different quality objects are placed on the pressure sensor which records the voltage and the corresponding object quality to train BP network by using δ learning rule. We first obtain a pressure

value, and then multiply by the acceleration due to gravity to calculate the forces according to the trained neural network corresponding to the quality.

The 3D soft tissue parameters estimation platform that consists of PPT2 and pressure acquisition devices is shown in Fig. 1.
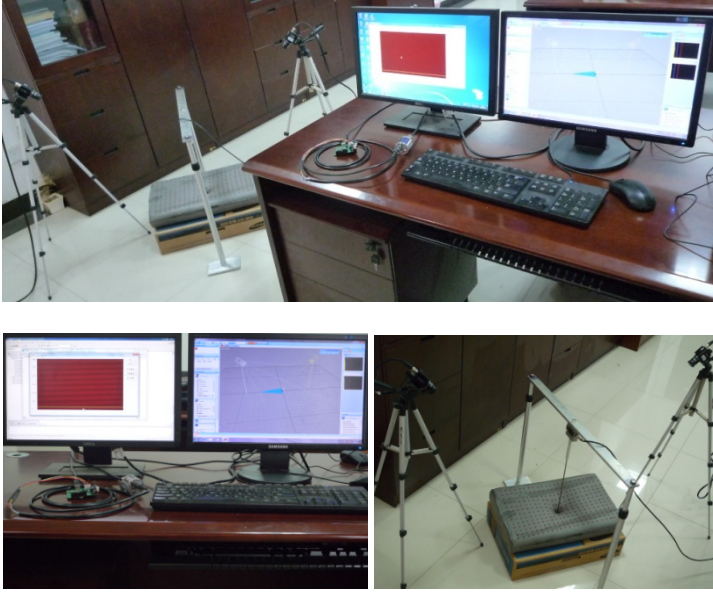
Fig. 1. 3D soft tissue parameters estimation platform

3   Physics-Based Soft Tissue Parameters Estimation Method

We adopt Physics-based linear finite element model to estimate the 3D soft tissue parameters. The 3D soft tissue is represented by an isotropic linear tetrahedron finite element model. In addition to the tetrahedron finite element model's space coordinates, its element stiffness matrix and global stiffness matrix is determined by Young's modulus E and Poisson ratio v which are the key parameters and respectively defines material elasticity and describes material compressibility.

3.1   Global Stiffness Matrix Parameterization

For isotropic linear tetrahedron finite element model, the element stiffness matrix of tetrahedron finite element is

$$\boldsymbol{K}^e = \int_{V_e} \boldsymbol{B}^T \boldsymbol{D} \boldsymbol{B} \, dV = \boldsymbol{B}^T \boldsymbol{D} \boldsymbol{B} \cdot \int_{V_e} dV = V_e \boldsymbol{B}^T \boldsymbol{D} \boldsymbol{B} \tag{1}$$

Constant matrix $\boldsymbol{B}$ and volume V_e are dependent on the initial position of the element's nodes. Elastic matrix $\boldsymbol{D}$ is a 6×6 symmetric matrix with parameters

Young's modulus E and Poisson ratio v. By the requirements that elastic matrix is positive-definite and the elastic material is isotropous, the Poisson ratio v meets $0 < v < 0.5$. Here we divide elastic matrix D in the following form

$$D = E(\varphi(v)D_1 + \phi(v)D_2) \tag{2}$$

Where D_1 and D_2 are constant matrix, then the element stiffness matrix K^e and global stiffness matrix K can be written as

$$K^e = E(\varphi(v)K_1^e + \phi(v)K_2^e) \tag{3}$$

$$K(E,v) = E(\varphi(v)K_1 + \phi(v)K_2) \tag{4}$$

with

$$K_1^e = V_e B^T D_1 B, \quad K_2^e = V_e B^T D_2 B \tag{5}$$

After assembling K_1^e and K_2^e we obtain K_1 and K_2.

3.2   Soft Tissue Parameters Estimation and Verification

During the experiment, we act a force on a finite element node and acquire nodal displacements and the force's value by using PPT2 and pressure acquisition devices. Then we adopt the static finite element equation to estimate the soft tissue parameters(

$$K(E,v)U = F \tag{6}$$

Define

$$H(v) = \varphi(v)K_1 U + \phi(v)K_2 U \tag{7}$$

$$H(v) = (h_1(v), h_2(v), \cdots, h_n(v))^T, \quad F = (f_1, f_2, \cdots, f_n)^T \tag{8}$$

We have

$$EH(v) = F \tag{9}$$

Our goal is to solve E and v which meet equation (10).

$$(\hat{E}, \hat{v}) = \arg\min_{(E,v)} \left\{ \sum_{i=1}^{n} \|Eh_i(v) - f_i\|^2 \right\} \tag{10}$$

We calculate p sets of experimental data, the corresponding parameters are (E_1, v_2), ..., (E_p, v_p) and the forces acted on finite element node m are $f_m^{(1)}$, ..., $f_m^{(p)}$. We construct a parameters reference model which allows us to estimate the soft tissue parameters of other sets of data to be verified. Suppose the force acted on finite element node m is f_t, we define a vector $W_t = (w_{t,1}, w_{t,2}, \cdots, w_{t,p})^T$.

$$w_{t,i} = \sum_{j=1}^{p} \left| \boldsymbol{f}_t - \boldsymbol{f}_m^{(j)} \right| \Big/ \left| \boldsymbol{f}_t - \boldsymbol{f}_m^{(i)} \right|, \qquad i = 1,2,\cdots,p \tag{11}$$

Suppose the soft tissue parameters are E_t and v_t, define $\boldsymbol{E} = \left(E_1, E_2, \cdots, E_p \right)$ and $\boldsymbol{v} = \left(v_1, v_2, \cdots, v_p \right)$, we estimate the soft tissue parameters as

$$E_t = \boldsymbol{W}_t \boldsymbol{E} \big/ sum(\boldsymbol{W}_t) \tag{12}$$

$$v_t = \boldsymbol{W}_t \boldsymbol{v} \big/ sum(\boldsymbol{W}_t) \tag{13}$$

Where $sum(\boldsymbol{W}_t)$ is the sum of elements in $\boldsymbol{W}_t$, we have

$$E_t \left(\varphi(v_t) \boldsymbol{K}_1 + \phi(v_t) \boldsymbol{K}_2 \right) \boldsymbol{U}_c = \boldsymbol{F} \tag{14}$$

Here we adopt conjugate gradient method to solve equation (14), $\boldsymbol{U}_c$ is the calculated nodal displacements. Suppose $\boldsymbol{U}_{ep}$ is the measured nodal displacements, the average deviation avg_dev between $\boldsymbol{U}_c$ and $\boldsymbol{U}_{ep}$ is calculated by equation (15).

$$avg\_dev = \sqrt{\sum_{i=1}^{n} \left\| \boldsymbol{U}_c(i) - \boldsymbol{U}_{ep}(i) \right\|^2 \Big/ n} \tag{6}$$

The process of the Physics-based soft tissue parameters estimation method is shown in Fig. 2.
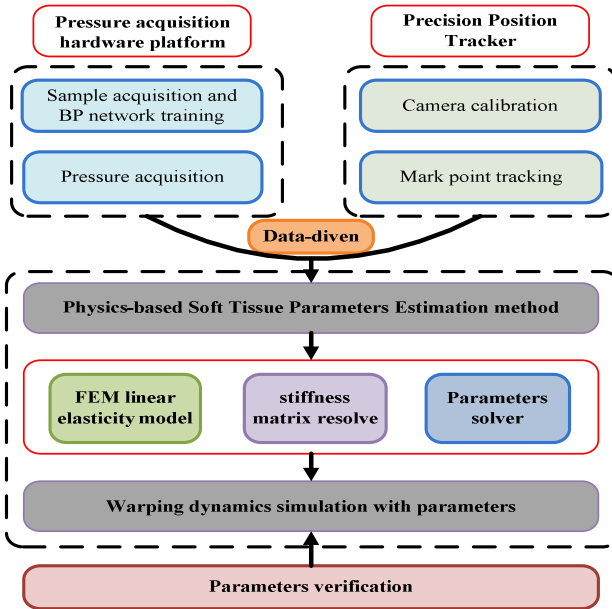
Fig. 2. Structure of Physics-based soft tissue parameters estimation method

4   Experimental Results

Our experimental platform includes: (1) hardware: Intel® Xeon® CPU, 2.40GHz, 4GB memory, SDI-2F miniature pressure sensor, AT91SAM7S64-based application board; (2) software：Copy CAD 6.0.04, MATLAB 2009a, PPT2. When the warping simulation of soft tissue is solved by using our method in this paper, we exported the soft tissue model in the STL format and rendered it in Copy CAD 6.0.04.

4.1   Tetrahedron Finite Element Model Construction

We construct a linear tetrahedron finite element model after obtaining the nodes' space coordinates through the PPT2. For the soft tissue materials we adopt in the experiment, finite element number of the tetrahedron finite element model is 1584, the number of finite elements node is 598, the obtained nodes and constructed tetrahedron finite element model is shown in Fig. 3.
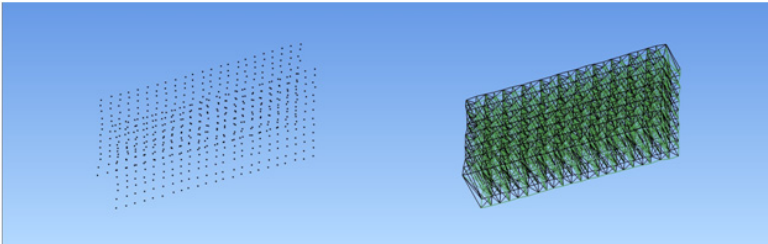
Fig. 3. The measured nodes and tetrahedron finite element model

4.2   Parameters Estimation and Verification

We adopt four sets of data to calculate Young's modulus E and Poisson ratio ν and the results are shown in Table 1.

Table 1. Calculation results of experiment data used for construction of reference models

Data sets	E	ν	F
e1	7.9720×10^3	0.06	-19.9724
e2	5.9830×10^3	0.06	-21.4326
e3	7.3386×10^3	0.07	-27.0382
e4	8.8096×10^3	0.06	-30.6152

Three sets of data are used for verification. We substitute Young's modulus and Poisson ratio in Table 1 in equation (12) and (13), and obtain the new Young's modulus and Poisson ratio of v1,v2 and v3, then substitute the new parameters in equation (14) and (15) to solve the average deviation between measured displacement and calculated displacement for verification. The results are shown in Table 2.

Table 2. Calculation results of experiment data used for verification

Data sets	E	V	avg_dev	F
v1	7.3583×10^3	0.0659	0.0024	-16.8658
v2	7.4199×10^3	0.0666	0.0022	-18.5906
v3	7.2190×10^3	0.0652	0.0034	-24.0786
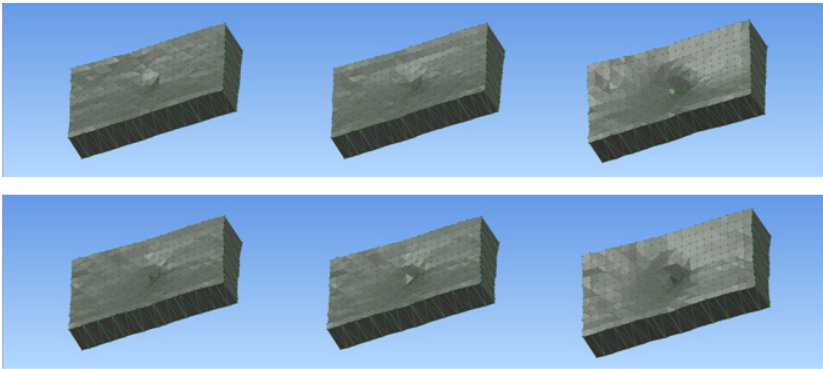
Fig. 4. The measured warping effects and the calculated warping effects

In order to verify the parameters we obtained, we substitute the parameters in the static finite element equation and calculate the nodal displacement of finite element nodes and render the warping effect.

As shown in Fig. 4, the figures above show the warping effects by using measured nodal displacements of data sets v1, v2 and v3, which is obtained by the optical tracking system; the figures below show the warping effects of calculated nodal displacements in verification data sets v1, v2 and v3. The results in Table 2 and Fig. 4 show accurate agreement of calculated warping effects and measured warping effects.

5   Conclusion and Discussion

In this paper, we present and verify a precise and robust 3D soft tissue parameters estimation method. Firstly we obtain the pressure exerted on the experimental material through the pressure acquisition devices and construct a linear tetrahedron finite element model and obtain nodal displacements through the PPT2. By using the acquired pressures and nodal displacements, we then calculate the parameters of 3D soft tissue by using reverse engineering method. Then we verify the obtained parameters and substitute them to calculate the nodal displacement, experimental results show the accurate agreement of the warping effect we calculated and the warping effect we obtain by experiment. In the future work, we will utilize more

precise miniature pressure sensor and position tracking method, and further improve the accuracy of the parameter estimation method by combining with nonlinear finite element model.

References

1. Nealen, A., Muller, M., Keiser, R., Boxerman, E., Carlson, M.: Physically based deformable models in computer graphics. Computer Graphics Forum 25(4), 809–836 (2006)
2. Schoner, J.L., Lang, J., Seidel, H.-P.: Measurement-based interactive simulation of viscoelastic solids. Computer Graphics Forum (Proc. Eurographics) 23(3), 547–556 (2004)
3. Wang, H., O'Brien, J.F., Ramamoorthi, R.: Data-driven elastic models for cloth: modeling and measurement. ACM Transactions on Graphics 30(4), Article 71 (July 2011)
4. Kim, T.-Y., Vendrovsky, E.: DrivenShape – a date-driven approach for shape deformation. In: Eurographics/ACM SIGGRAPH Symposium on Computer Animation, pp. 49–55
5. Bickel, B., Bacher, M., Otaduy, M.A., Matusik, W., Pfister, H., Gross, M.: Caputre and modeling of nonlinear heterogeneous soft tissue. ACM Transactions on Graphics 28(3), Article 89 (August 2009)
6. Sueda, S., Kaufman, A., Pai, D.K.: Musculotendon simulation for hand animation. ACM Trans. Graph. (Proc. SIGGRAPH) 27(3) (2008)
7. Peterlík, L., Sedef, M., Basdogan, C., Matyska, L.: Real-time visio-haptic interaction with static soft tissue models having geometric and material nonlinearity. Computers & Graphics 34, 43–54 (2010)
8. Park, S.I., Hodgins, J.K.: Capturing and animating skin deformation in human motion. ACM Transaction on Graphics (SIGGRAPH 2006) 25(3), 881–889 (2006)
9. Tautges, J., Zinke, A., Kruger, B., Baumann, J., Weber, A., Helten, T., Muller, M., Seidel, H.-P., Eberhardt, B.: Motion reconstruction using sparse accelerometer data. ACM Transactions on Graphics 30(3), Article 18 (May 2011)
10. Samur, E., Sedef, M., Basdogan, C., Avtan, L., Duzgun, O.: A robotic indenter for minimally invasive measurement and characterization of soft tissue response. Medical Image Analysis 11, 361–373 (2007)

An Efficiency-Driven Deterministic Optimization Approach for Sensor Placement in Image-Based Forest Field Measurement

Luis Diago[1], Nobuyoshi Muto[2], Lu Yang[1], Zheng Gong[1], and Ichiro Hagiwara[1,3]

[1] Department of Mechanical Science and Engineering, Tokyo Institute of Technology,
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
[2] Research Institute of Local Industries and Economy, Japan
[3] Institute for Advanced Study of Mathematical Sciences (MIMS), Meiji University,
1-1-1, Higashi-Mita, Tama-ku, Kawasaki, 214-8571, Japan

Abstract. In forest management, one key objective of forest field measurement is to measure the Diameter at Breast Height (DBH) of each tree in a specified area. Nowadays the widely employed way is to measure the trees manually one by one which usually takes several days. However much work can be considerably saved by adopting image based measurement approach which includes several steps. Among these steps, careful planning of the sensor placement is an essential preparation step which has significant impact on the effectiveness of the following steps. In this paper, a concept named Trees Per Location (TPL) is proposed to evaluate the efficiency in sensor placement. Based on TPL, we present a novel automatic sensor placement algorithm suitable for image based forest field measurement. The key feature of the proposed algorithm is that the impact of the camera orientation on optical constraints is attenuated due to the fact that in outdoors, the orientation of camera is not easy to control compared with the location of camera. Our method generates a plan composed of a series of sensor viewpoints and a shortest path that traverses each viewpoint exactly once. The plan guarantees that the total number of images needed to be taken is minimum and the travel distance of the path is the shortest while our plan satisfies the constraint that each tree appears in at least one image without being blocked by any other trees. Experiments are carried out on a sample forest from the PlotNet database and a real forest in order to compare the proposed TPL with other mainstream algorithms and validate the proposed method for DBH measurement.

1 Introduction

Traditionally forest information is gathered by field sampling methods [4]. For more demanding investigations and analyses, measurement is conducted using remotely sensed data. Among several possibilities, LiDAR (Light Detection And Ranging) data are widely available commercially [5]. LIDAR is an optical means with high costs per hectare and LiDAR data are only being utilized by a subset of forest landowners and land managers at present.

As such, recent research is focused on applying image based measurement methods for forest inventories. The process of image based measurement consists of several steps

[8]. Among these steps, sensor placement is the procedure to determine the number of sensors, the position and pose of each sensor so that minimum cost is achieved while high accuracy is maintained. Careful planning of the sensor placement is an essential preparation step which has significant impact on the effectiveness of the following steps.

However, most existing researches on image based forest measurement do not pay much attention to the important sensor placement step. Comparatively, in the research area of model based 3D objects measurement, extensive work is carried out dealing with the sensor placement problem and fruitful results are achieved in recent years. The "best-next-view" (BNV) based methods are adopted by many researchers [2][7]. BNV is defined as the next sensor pose which would enable the greatest amount of previously unseen three-dimensional information to be acquired. The critical problem with these methods is the global optimization of sensor planning.

To achieve high efficiency and quality, the optimal spatial distribution of the viewpoints should be determined too. Chen et al. [3] proposed an automatic approach to determinate the optimal sensor positions and a shortest path through these viewpoints. However, their method is only applied to the measurement of small-scale objects, not suitable for forest field measurement because the constraints considered in their method are not appropriate for large-scale inventories such as forests.

To sum up, all previous works on image based forest measurement do not emphasize in the sensor placement problem while existing sensor placement algorithms applied to other areas do not fit into large-scale forest inventories. So in this paper, we present a novel automatic sensor placement algorithm suitable for image based forest field measurement. A concept named Trees Per Location (TPL) is proposed to evaluate the efficiency in sensor placement. Based on TPL, the proposed algorithm generates a plan composed of a series of sensor viewpoints and a path that traverses each viewpoint exactly once. The plan guarantees that the total number of places needed for data acquisition is minimum and the travel distance of the path is the shortest while our plan satisfies the constraint that each tree appears in at least one image without being blocked by any other trees. Our algorithm is evaluated with extensive experiments conducted on synthesized and real life sample forests and the key-technologies introduced in the paper considerably reduce field measurement time and facilitate the planning in current forest management systems.

2 TPL Based Sensor Placement Algorithm

First the forest area is discretized into a high resolution grid containing M^2 cells and the TPL of each grid cell is calculated. Then the grid cell with the highest TPL is picked out as one optimized sensor position followed by the TPL of each of the remaining grid cells updated. This procedure repeats until all trees in the area are marked captured. Last the optimized path that traverses each optimized location exactly once is found with a genetic algorithm to minimize the total travel distance. The Table 1 lists all the camera model parameters either used or to be determined. The forest model F is defined as:

$$F = \{\boldsymbol{t}_i = (\boldsymbol{l}_i, r_i)| \quad ||\boldsymbol{l}_i - \boldsymbol{l}_j|| > r_i + r_j : j \neq i\} \tag{1}$$

where $(\boldsymbol{l}_i, r_i)$ denotes the location and radius of each tree $\boldsymbol{t}_i$. Four contraints (visibility, focus, field of view and resolution) are included in proposed algorithm.

Table 1. Optical constraints parameters

$\boldsymbol{v}_i$	Azimuth of tree relative to camera
β_i	Half angular diameter
θ^c	Pan angle of camera
$\boldsymbol{v}^c$	Unit vector along the optical axis
$\boldsymbol{l}^c$	Location of camera
ψ	Angle between $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$
ϕ	Angle between $\boldsymbol{v}_i$ and $\boldsymbol{v}^c$
a	Diameter of the aperture of the lens
d	Distance from lens to image plane
f	Effective focal length of the lens
w	Size of the tree in the image plane
c	Limiting blur circle diameter
D_1	Far limit of the depth of field
D_2	Near limit of the depth of field
D_i	Minimum depth of tree
D_h	Hyperfocal distance
d_h	Hyperfocal image distance
α	Field of view angle
I	Size of sensor
R	Resolution of image

The visibility constraint $G^1_{u,v}(\boldsymbol{t}_i)$ can be expressed as:

$$
G^1_{u,v}(\boldsymbol{t}_i) := \begin{cases} true, \ \beta_i + \beta_j \leq \psi \\ true, \ \beta_i + \beta_j > \psi \ \text{AND} \\ \quad ||\boldsymbol{l}_j - \boldsymbol{l}^c|| > ||\boldsymbol{l}_i - \boldsymbol{l}^c|| \\ false, \text{otherwise} \end{cases} \tag{2}
$$

The trees not being blocked by any other trees relative to camera location $\boldsymbol{l}^c_{u,v}$ are marked ready for the subsequent tests.

The philosophy implied in the other three optical constraints is that when the image based field measurement is carried out, we should not need to worry about the viewing strategy at each viewpoint. We only need to stand at the selected optimized sensor positions and make a full $360°$ turn to take several images that can constitute a panorama of surrounding trees. In this way, although some more images may be required to take, much time and labor will be saved from manipulating the orientation $\boldsymbol{v}^c$ of camera since in outdoors, the pan angle θ^c of camera is not easy to control compared with the location $\boldsymbol{l}^c$ of camera which can be retrieved by reading the handheld GPS device.

Specifically, for the focus constraint, in order to make sure the tree is in focus no matter in which direction the camera is oriented, the limit situation illustrated in Fig. 1 is considered. The tree $\boldsymbol{t}_i = (\boldsymbol{l}_i, r_i)$ is tangent to both the near focal plane and the field of view bound. From the illustrated geometric relationships we have
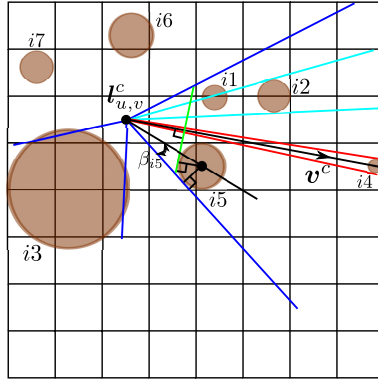
Fig. 1. Computation of TPL

$$||l_i - l_{u,v}^c|| = \frac{D_2}{\cos\left(\frac{\alpha}{2} - \beta_i\right)} + \frac{r_i}{\cos\left(\frac{\alpha}{2} - \beta_i\right)}$$

$$= \frac{D_2 + r_i}{\cos\left(\frac{\alpha}{2} - \arcsin\frac{r_i}{||l_i - l^c||}\right)}. \tag{3}$$

So if the following statement is true, we say that the tree t_i satisfies the focus constraint.

$$G_{u,v}^2(t_i) : \cos\left(\frac{\alpha}{2} - \beta_i\right)||l_i - l_{u,v}^c|| - r_i \geq D_2,$$

$$D_2 = \frac{D_h}{2} = \frac{f}{2}\left(1 + \frac{a}{c}\right) \tag{4}$$

As to the field of view constraint, we do not check whether the tree lies inside a certain field of view defined by v^c and α. Instead, we only confirm whether the tree can be enclosed in a single image. The statement

$$G_{u,v}^3(t_i) : \beta_i = \arcsin\frac{r_i}{||l_i - l^c||} \leq \alpha/2 \tag{5}$$

holds true for the trees that satisfy the softened field of view constraint which is independent of camera orientation. The tree $i3$ in Fig. 1 is an example considered out of the field of view in our algorithm.

For the resolution constraint, when the camera orientation $v_{u,v}^c$ is identical to v_i (i.e., $\phi = 0$), the size w of the tree in the image plane will be the smallest, such as the one $i4$ in Fig. 1. So if w in this limit situation is greater than or equal to the diagonal length of a pixel, the tree will be resolvable by the sensor located at $l_{u,v}^c$ regardless of the camera orientation.

$$G_{u,v}^4(t_i) : w \geq \sqrt{\left(\frac{I_h}{R_h}\right)^2 + \left(\frac{I_v}{R_v}\right)^2} \tag{6}$$

$$w = 2d\tan\beta_i = 2d_h\tan\beta_i = 2f\left(1 + \frac{c}{a}\right)\tan\beta_i$$

Table 2. Proposed sensor placement plan

$P^o = \varnothing$
$C^o = \varnothing$
while $F! = \varnothing$
   for each $l^c_{u,v} \in C$
      $F_{u,v} = \varnothing$
      for each $t_i \in F$
         if $G^4_{u,v}(t_i) == true$ **then**
            if $\bigcap_{k=1}^3 G^k_{u,v}(t_i) == true$ **then**
               $F_{u,v} = F_{u,v} \bigcup \{t_i\}$
            endif
         endif
      endfor
      $e_{u,v} = |F_{u,v}|$
   endfor
   $(u^o, v^o) = \underset{(u,v)}{\mathrm{argmax}}\, e_{u,v}$
   $F = F \backslash F_{u^o,v^o}$
   $C = C \backslash \{l^c_{u^o,v^o}\}$
   $P^o = P^o \bigcup \{F_{u^o,v^o}\}$
   $C^o = C^o \bigcup \{l^c_{u^o,v^o}\}$
endwhile

The TPL $e_{u,v}$ of the specified camera location $l^c_{u,v}$ is assigned the value equal to the number of trees in forest that pass all the above four tests. The larger TPL value of a camera location is, the more amount of DBH information of trees can be acquired at this location.

The next step determines the minimum number of viewpoints needed to gather the DBH information of all trees in the forest and their corresponding positions. The detail of the proposed algorithm is shown in Table 2. P^o is the generated optimized sensor placement plan. C^o is the selected optimal sensor positions set and $C^o \subset C = \{l^c_{u,v} | u, v \in \mathbb{N}; u, v \leq M\}$.

To achieve high efficiency and quality, the optimal sensing sequence should be determined too. The last step of our method handles this issue by finding a shortest possible tour that visits each location $l^c_{u^o,v^o}$ in set C^o exactly once. In our scheme, the shortest tour is found with a genetic algorithm [6].

3   Experiments

First we compare our algorithm with the global optimization method [3] and the next best view (NBV) method [2]. For the experiment, a 200×200 lattice (i.e., $M = 200$) is constructed covering a sample forest from the project PlotNet [1]. The name of the database is TOEF_GasPlot and it includes 559 trees with average DBH of $8.98cm$ located at N $42°40'$ E $141°35'$ in Hokkaido, Japan. The size of the forest area A is also $50m$ long by $50m$ wide. According to the records, all data were gathered with girth

(a) Our TPL method   (b) Global Optimization [3]

(c) NBV method [2]
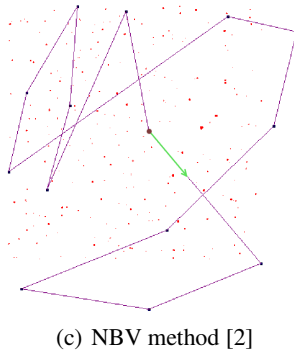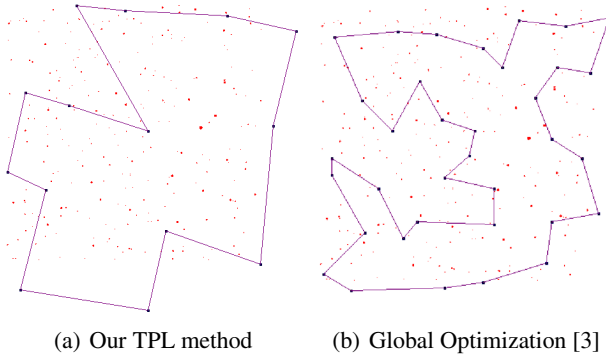
(d) Kinect
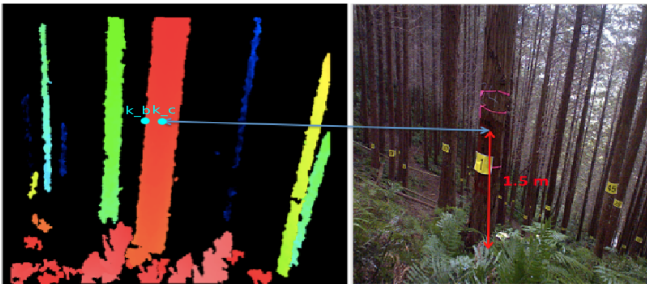
(e) DBH measurement

Fig. 2. Experimental results

Table 3. Comparison of sensor placement algorithms using PlotNet database

	TPL	Global[3]	NBV[2]
Images	14	35	14
Distance	248.0m	299.8m	363.3m

measurement from middle to late April in 1998, 1999, 2000, 2001, 2002 and 2003. The comparison results are reviewed in Table 3 and illustrated in Fig. 2. In this figure, the DBH information measured in the year 2003 is visualized. Clearly our TPL algorithm is much more effective than the other two. From the table we notice that though the global method requires more sensor positions, the travel distance of this method after optimization is always shorter than the unoptimized one of the NBV method. It proves the significance of optimization of the sensing sequence.

From January 18 to 21, 2012, in coordination with the government of Ehime Prefecture and Ishizuchi Forestry Association, an area of $225m^2$ (15m x 15m) was also selected to make measurements of the DBH of 45 trees. We measured the DBH using the traditional method and compared the results with measurements made by the Microsoft Kinect sensor in Fig. 2(d) which can output not only image but also depth. The depth information (k_b and k_c) is adopted to calculate the DBH using (7).

$$DBH = \frac{k_b^2 - k_c^2}{2k_c},\tag{7}$$

The points used to measure k_b and k_c are ilustrated in Fig. 2(e). Table 4 shows the minimum (Min), maximum (Max), average (Ave) and standard deviation (Std) of the absolute differences (Diff.) in the DBH of 45 trees computed by the traditional method and using the Kinect sensor. The last column shows the percent of error (%error) produced by the proposed method as compared with the traditional method. As result we achieve an average error of 26% between the two DBH measurements.

Table 4. Experimental results of DBH measurement for 45 trees by the traditional method and the proposed method using Microsoft Kinect sensor

	Diff.(cm)	Traditional (cm)	Kinect (cm)	% error
Min	1.54	20	18.45	7.7
Max	10.4	26	36.41	40.06
Ave.	5.52	22.6	25.93	26.3
Std.	2.90	4.32	6.74	17.5

4   Conclusion and Future Work

In this paper, a concept named Trees Per Location (TPL) is proposed to evaluate the efficiency in sensor placement. Based on TPL, our proposed algorithm guarantees that the total number of images needed to be taken is minimum and the travel distance of

the path is the shortest while our plan satisfies the constraint that each tree appears in at least one image without being blocked by any other trees. The key feature of the proposed algorithm is that the impact of the camera orientation on optical constraints is attenuated due to the fact that in outdoors, the orientation of camera is not easy to control compared with the location of camera which can be retrieved by reading the handheld GPS device. The key-technologies introduced in the paper considerably reduce field measurement time and facilitate the planning in current forest management systems.

References

1. PlotNet Forest Database, `http://eco1.ees.hokudai.ac.jp/plotnet/home`
2. Banta, J., Abidi, M.: Autonomous placement of a range sensor for acquisition of optimal 3-d models. In: Proc. of Industrial Electronics, Control, and Instrumentation, vol. 3, pp. 1583–1588 (1996)
3. Chen, S., Li, Y.: Automatic sensor placement for model-based robot vision. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 34(1), 393–408 (2004)
4. Holopainen, M., Kalliovirta, J.: Modern data acquisition for forest inventories. In: Forest Inventory. Managing Forest Ecosystems, vol. 10, pp. 343–362. Springer, Netherlands (2006)
5. Hosoi, F., Omasa, K.: Estimating vertical plant area density profile and growth parameters of a wheat canopy at different growth stages using three-dimensional portable lidar imaging. ISPRS Journal of Photogrammetry and Remote Sensing 64(2), 151–158 (2009)
6. Khan, F., Khan, N., Inayatullah, S., Nizami, S.: Solving TSP problem by using genetic algorithm. IJBAS 9(10), 79–88 (2009)
7. Li, Y., Liu, Z.: Uncertainty-driven viewpoint planning for 3d object measurements. In: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA 2003, vol. 1, pp. 127–132 (2003)
8. Reiterer, A., Lehmann, M., Fabiankowitsch, J., Kahmen, H.: Quality control for building industry by means of a new optical 3d measurement and analysis system. In: Observing our Changing Earth. International Association of Geodesy Symposia, vol. 133, pp. 771–779. Springer, Heidelberg (2008)

hMETIS-Based Offline Road Network Partitioning[*]

Yan Xu and Gary Tan

School of Computing National University of Singapore
13 Computing Drive
Singapore 117417
{Xuyan,gtan}@comp.nus.edu.sg

Abstract. Distributed transportation simulation is an important technology for evaluating large-scale traffic applications and control policies, before they are implemented in real-world traffic systems. Offline road network partitioning is the first step towards distributed transportation simulation. Currently, road network portioning algorithms, like METIS, are designed to reduce the execution time. However, a slower execution time is acceptable for offline road network partitioning, if it leads to more efficient road network partitions. This paper introduces *hMETIS-based offline road network partitioning*. One experiment based on Singapore expressways shows that compared with the famous METIS-based offline road network partitioning, hMETIS-based offline road network partitioning reduces the number of vehicles crossing partitions by 9.8% on average, with a similar load imbalance and an acceptable execution time. For distributed traffic simulations, where there are large amount of data exchanged between partitions, *hMETIS-based offline road network partitioning* is one candidate solution to reduce the simulation time and increase the scalability.

Keywords: distributed transportation simulation, offline road network partitioning, hMETIS.

1   Introduction

Transportation simulation is an important area of discipline in traffic engineering and transportation planning, in order to evaluate new traffic infrastructures, intelligent transportation applications, traffic control strategies and lane use plans, before they are implemented in the real-world traffic systems. To boost the simulation speed and the size of the simulated scenario, one option is to execute transportation simulations on distributed computers. *Offline road network partitioning* is the first step to space-parallel distributed transportation simulation. *Offline* means that the road network partitioning is pre-calculated before traffic simulations.

The purpose of *offline road network partitioning* is to divide a large traffic road network into *k* disjoint small partitions, so that each computer simulates only a partial

road network. Fig. 1 shows the structure of offline road network partitioning. There are four levels in the structure. The bottom level is the road network level. This level represents traffic facilities that constitute the road network. This level can be modeled as nodes and links.   The second level is the vehicle movement level. This level represents vehicle behaviors and person behaviors on the road network. This level can be modeled as speed, flow and density on links. The third level is the traffic control level. This level represents traffic applications and policies to control traffic facilities and traffic flow. The top level is the partitioning level. This level represents the solution to divide the road network. This level can be modeled as the list of zones in each partition. Different types of zones can be defined for different requirements. In this paper, each zone is centered by one node and cuts neighbor links in the middle.
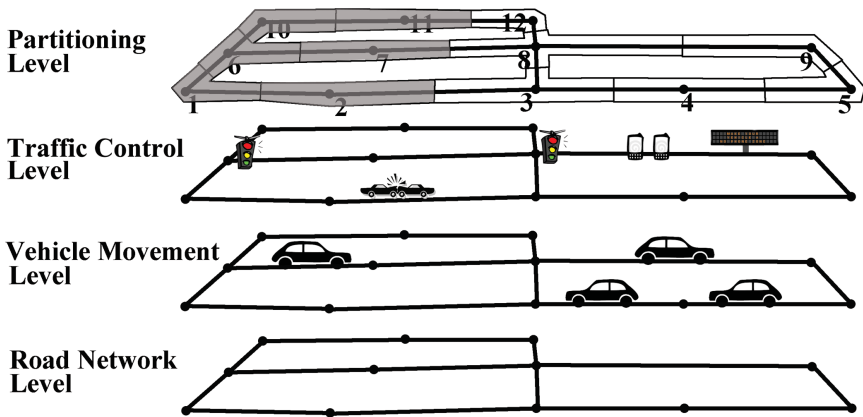
Fig. 1. The structure of offline road network partitioning

The benefit of *offline road network partitioning* is that each computer simulates only the vehicle movement and traffic control in its own partial road network, which might reduce the total simulation time. The cost of *offline road network partitioning* is that computers need to send messages to each other, if for example, one vehicle moves from one partition to another, or the traffic signal control in one partition needs the traffic flow data in another partition. A good *offline road network partitioning solution* should maximize the benefit and minimize the cost at the same time. The efficiency of road network partitions is measured by three major parameters [1] [9]:

- **load imbalance:** The load of one partition is the time to simulate the vehicle movement and traffic control in its partial traffic road network. Load imbalance is the ratio of the heaviest partition load over the average partition load. However, the time to simulate one partition is not easy to be directly modeled. A practical way is to use the number of simulated vehicles or the geographical size of a partition to estimate the load of that partition.
- **data exchange:** It is the time cost caused by sending messages between partitions. It includes vehicles crossing messages in the vehicle movement level

and traffic control messages between partitions in the traffic control level. However, the time cost caused by sending messages is not easy to be directly modeled. A practical way is to use the total data size or the total number of messages to estimate the cost.

- **execution time:** It is the time taken by the partitioning algorithm to generate road network partitions, which does not include I/O time.
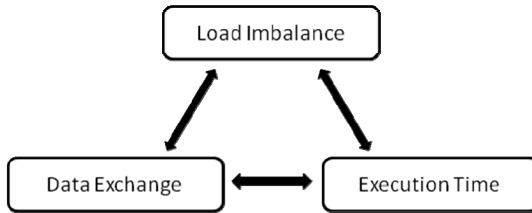
Fig. 2. Three major measurements of road network partitions

METIS [4] is the most popular offline traffic road network partitioning solution in distributed transportation simulators [1, 5, 8, 9], because it can always give good road network partitions in a small execution time. METIS is designed to reduce the execution time. However, to the best of our knowledge, *offline road network partitions* are always pre-calculated before traffic simulations, and then kept in databases or file systems. Thus, a slower execution time is acceptable, if the road network partition solution leads to more efficient road network partitions.

This paper introduces a *hMETIS-based offline road network partitioning solution*. Preliminary work on this has been reported in [10]. hMETIS[2] is a set of *hypergraph partitioning algorithms*, which are designed for *Very Large Scale Integration* (VLSI) applications. A *hypergraph* is a generalization of a normal graph, where a *hyperedge* can connect any n ($n \geq 2$) vertices [7]. hMETIS can also be used in *offline road network partitioning*. According to our experiments based on Singapore expressways, compared with *METIS-based offline road network partitioning*, *hMETIS-based offline road network partitioning* reduces the number of vehicles crossing partitions by 9.8% on average, with a similar load imbalance and an acceptable execution time. This research is part of project SimMobility [6], which is investigating a distributed transportation simulator in Singapore. The remainder of the paper is organized as follows. Section 2 describes the *hMETIS-based offline road network partitioning*. Section 3 presents experiments to investigate the efficiency of the *hMETIS-based offline road network partitioning solution*. Section 4 draws the conclusions.

2　hMETIS-Based Offline Road Network Partitioning

In this section, we present the problem formulation of *hMETIS-based offline traffic road network partitioning*, and the state of the art algorithm to solve the problem.

2.1   Problem Formulation

The road network is formulated as a hypergraph $G = (V, E)$, where V denotes the set of vertices and E is the set of edges. Each vertex in V represents a zone (as shown in the partitioning level in Fig. 1) and each edge in E represents one type of data exchange between m zones ($m >= 2$).   Each vertex has one weight, representing the workload to simulate that zone. Each edge also has a weight, representing the size of the data exchanged between the zones. A k-way road network partitioning is to divide G into k disjoint partitions, in order that each partition has the same workload, and the total weight of cut edges is minimized.

---

hMETIS-based offline road network partitioning

Inputs: $G=(V, E)$
Nodes and Weights:
$V = \{ (1, 1), (2, 2), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1) \}$
Edges and Weights:
$E = \{ (1, 2, 1) \}$

Targets:
Divide G into k disjoint partitions, so that:
1: the total nodes' weight in each partition is about the same
2: the total weight of cut edges is minimized
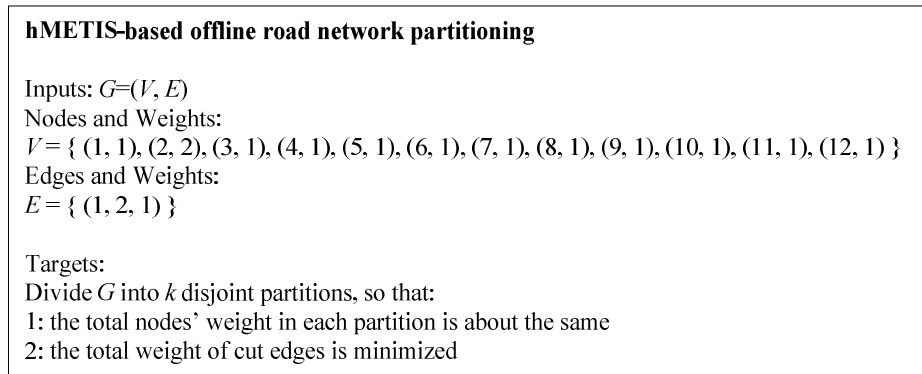
Fig. 3. Example of offline road network partitioning

   In order to explain the problem formulation, we present one example. As shown in Fig. 1, there are 12 zones in the road network, and thus there are 12 vertices in the graph. Assume the total number of vehicles in these 12 zones within the simulation period is $\{1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$. Since the number of vehicles is a reasonable estimation of the simulation load, the weights of the 12 zones are estimated as $\{1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$. Assume that there is only one vehicle crossing from zone 1 to zone 2, then there is one data exchange between zone 1 and zone 2 in vehicle movement level, and the data exchange can be modeled as one edge in zones $\{1, 2\}$. The weight of the edge is 1, representing the number of vehicles. The data exchange between partitions is modeled as the total weight of cut edges. Finally, the problem formulation of the *hMETIS-based offline traffic road network partitioning* is shown in Fig. 3.

2.2   Algorithm

When the paper is written, the state of the art algorithm to solve *hypergraph-based offline road network partitioning* is hMETIS. As shown in Fig. 4, hMETIS is implemented based on the *multilevel graph partitioning paradigm* [2].
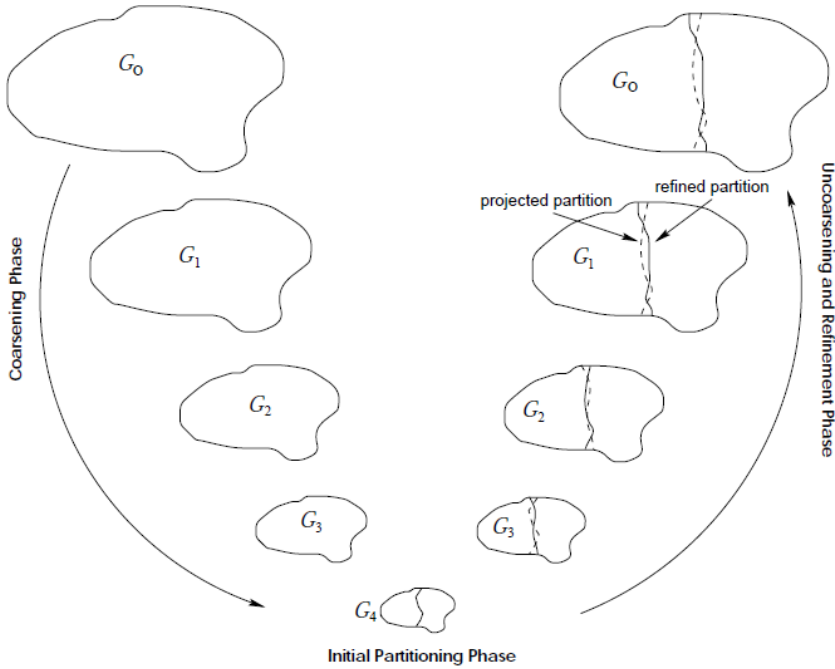
Fig. 4. Multilevel graph partitioning paradigm

The multilevel paradigm consists of three phases: *coarsening phase*, *initial partitioning phase*, and *uncoarsening and refinement phase*. In the *coarsening phase*, a series of successively smaller hypergraphs are derived from the input hypergraph. Each successive hypergraph is constructed from the previous hypergraph by collapsing together a set of adjacent vertices. This process continues until the size of the hypergraph has been reduced to just a few hundred vertices. In the *initial partitioning phase*, a partitioning of the coarsest and hence, smallest, hypergraph is computed using relatively simple approaches. Since the coarsest hypergraph is usually very small, this step is very fast. Finally, in the *uncoarsening and refinement phase*, the partitioning of the smallest hypergraph is projected to the successively larger hypergraphs by assigning the vertices that were collapsed together to the same partition as that of their corresponding collapsed vertex. After each projection step, the partitioning is refined using various heuristic methods to iteratively move vertices between partitions as long as such moves improve the efficiency of the partitioning solution. This phase ends when the partitioning solution has been projected all the way to the original hypergraph. The details of the algorithms in hMETIS are in [2].

3   Experiments

In this section, we present one experiment, in order to compare the efficiency of METIS-based and hMETIS-based *offline road network partitioning solutions*. The experiment is to divide the Singapore expressway road network into k partitions, for distributed transportation simulation ($2 \leq k \leq 100$).

3.1   Configuration

The Singapore expressway road network consists of 1954 nodes and 3182 links. Each node has two-dimensional coordinates and each link consists of a start node and an end node. Firstly, the road network is changed into a zone-based graph. Each zone is centered by a node and cut in the middle of neighbor links, as shown in the partitioning level in Fig. 1. Traffic flow and average speed (every 15 minutes) on each link in Singapore on October 5, 2011 are used to estimate the load of each zone and the data exchange between two contiguous zones. The load of one zone is the total number of vehicles on that zone in the day. The data exchange between two contiguous zones is the total number of vehicles, crossing the boundary in the day. Then, the road network partitioning is formulized as the format shown in Fig. 3.

In this experiment, we are using METIS 5.0 and hMETIS 2.0. In METIS, the default configuration of gpmetis algorithm is used. In hMETIS, as suggested by the manual, when the partition size is smaller than 17, *recursive bisection algorithm* is used; otherwise, *direct k-way partitioning algorithm* is used. The details of these two algorithms are in [2, 3]. To make the partitions comparable, we specify the maximum load imbalance of METIS and hMETIS to be 1.05, so that we can focus on the size of data exchange and the execution time. The experiment is run on a 2.83GHz Intel Core 2 Quad processor, with 4GB main memory, Ubuntu Linux 10.10.

3.2   Experiment Results

The partition comparisons between METIS-based and hMETIS-based *offline road network partitioning solutions* are shown in figures 5 to 7. Fig. 5 shows the comparison on the number of cross vehicles between partitions. We can see that the *hMETIS-based offline road network partitioning solution* has fewer vehicles crossing partitions, compared to METIS-based solution. The average reduction is 9.8%, and the best reduction is 26.5%. Fig. 6 shows the comparison on load imbalance. We can see that, when the number of partitions is smaller than 60, the load imbalance of the hMETIS-based solution is similar to the METIS-based solution. When the number of partitions increases higher, the load imbalance of the METIS-based solution turns worse. The reasons are explained later. Fig. 7 shows the comparison on execution time. We can see that the *hMETIS*-based solution requires more execution time. The average and maximum execution times are 3.02/5.12 sec, which is much higher than the METIS-based solution. However, for distributed transportation simulation on city-scale road networks, like Singapore Expressways, the execution time of the *hMETIS*-based solution is still acceptable. This experiment infers that if the execution

time is acceptable, the *hMETIS-based offline road network partitioning solution* reduces data exchanges between partitions by 9.8% on average, with similar load imbalance.
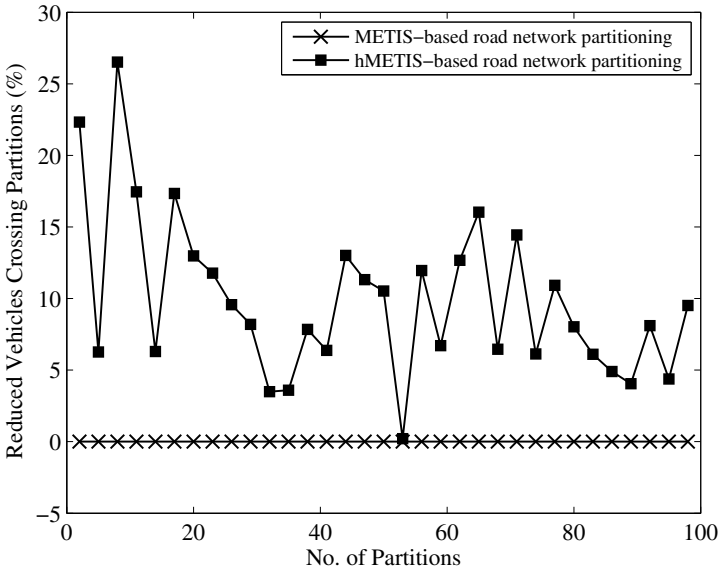
Fig. 5. Comparison of Cross Vehicles when partitioning the Singapore Expressway
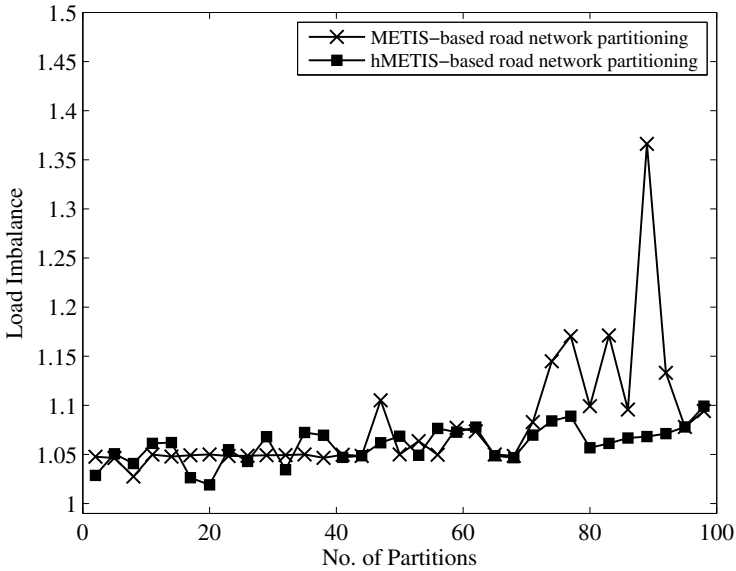
Fig. 6. Comparison of Load Imbalance when partitioning the Singapore Expressway
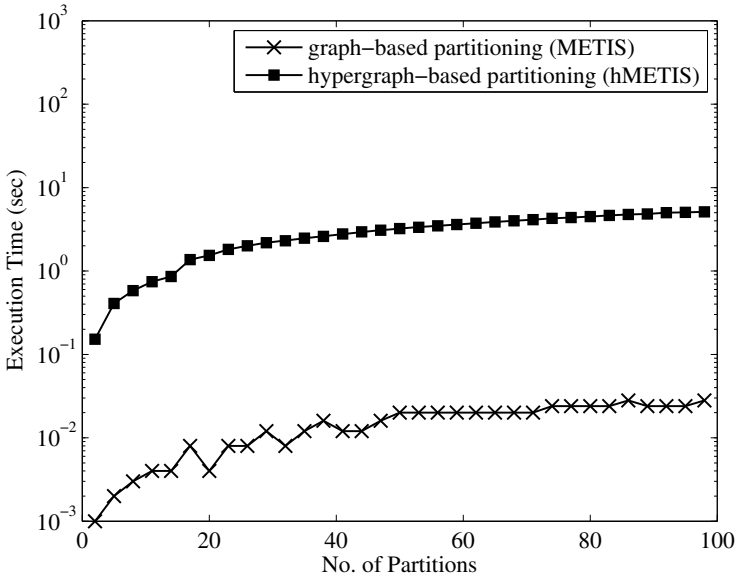
Fig. 7. Comparison of Cross Vehicles when partitioning Singapore Expressway

3.3   Discussion

The experiment results show that *hMETIS-based offline road network partitioning solution* reduces data exchanges between partitions by 9.8% on average, compared with the famous *METIS-based offline road network partitioning solution*. Based on our observations, there are two primary reasons to explain the reduction of data exchange. First, it is the *first choice coarsening scheme* [3] in the implementation of hMETIS. The idea of *first choice coarsening scheme* is to allow more than two vertices to be coarsened in one time. The *first choice coarsening scheme* has been proven to be able to protect the clusters of vertices that naturally exist in the source hypergraphs. However, based on our experiments, first choice coarsening schema also works in normal graphs. Second, instead of selecting a single initial partitioning in graph-based solution, many different initial partitioning solutions are kept in hypergraph-based solution. The reason is that the partitioning of the coarsest hypergraph that has the smallest data exchange may not necessarily be the one that will lead to the smallest data exchange in the original hypergraph. However, to keep many different initial partitioning solutions causes more execution time. It is also the reason that the hMETIS-based solution has better load imbalance than the METIS-based solution, when the number of partitions is larger than 60. Besides, we cannot guarantee the reduction of data exchanges between partitions theoretically. Generally, the reduction depends on the topology of the simulated road network and the distribution of traffic flow on the road network.

4   Conclusions

Distributed transportation simulation is important to evaluate large-scale traffic applications and control policies, before they are implemented in the real-world traffic systems. *Offline road network partitioning* is the first step towards the space-parallel distributed transportation simulation. *Offline road network partitioning* can be formulized as graph partitioning problems and also hypergraph partitioning problems. The experiments based on Singapore expressways shows that *hMETIS-based offline road network partitioning* can reduce the number of vehicles crossing partitions by 9.8% on average. In cases of distributed transportation simulations, where there are large amount of data exchanged between partitions, our proposal is one candidate solution to reduce the simulation time and increase the scalability. Besides, while we focus on offline distributed transportation simulations, our proposal can also be used to generate the initial road network partitions for real-time transportation simulations.

References

1. Charypar, D.: Efficient algorithms for the microsimulation of travel behavior in very large scenarios. Ph.D thesis, ETH Zurich (2008)
2. Karypis, G., Aggarwal, R., Kumar, V., Shekhar, S.: Multilevel hypergraph partitioning: applications in VLSI domain. IEEE Transactions on VLSI Systems, 69–79 (1999)
3. Karypis, G., Kumar, V.: Multilevel k-way hypergraph partitioning. VLSI Design, 1–16 (2000)
4. Karypis, G., Kumar, V.: A fast and highly quality multilevel scheme for partitioning irregular graphs. SIAM Journal on Scientific Computing, 359–392 (1999)
5. Nagel, K., Rickert, M.: Parallel implementation of the TRANSIMS micro-simulation. Parallel Computing, 1611–1639 (2001)
6. SMART, Future Urban Mobility (2012), `http://smart.mit.edu/research/future-urban-mobility/future-urban-mobility.html`
7. Voloshin, V.I.: Introduction to graph and hypergraph theory, 2nd edn., pp. 135–160. Nova Science Publishers (2009)
8. Yang, W.: Scalability of dynamic traffic assignment. Ph.D thesis, Massachusetts Institute of Technology (2009)
9. Cetin, N.: Large scale parallel graph-based simulations. PhD thesis, ETH Zurich (2005)
10. Xu, Y., Tan, G.: Offline Road Network Partitioning in Distributed Transportation Simulation. Short paper. In: Proceedings of the 26th Workshop on Principles of Advanced and Distributed Simulation (July 2012)

Exploring a P2P Based Collaborative Feature Modeling through a Procedural 3D CAD Language

Jiacai Wang and Ichiro Hagiwara

Department of Mechanical Science and Engineering, Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan
Institute for Advanced Study of Mathematical Sciences (MIMS), Meiji University
1-1-1 Higashi-Mita, Tama-ku, Kawasaki-shi, Kanagawa, 214-8571, Japan
{tz12011,ihagi}@meiji.ac.jp

Abstract. Real time collaborative feature-based CAD modeling within geographically dispersed participants is one of the current research hot spots in collaborative design community. Most existing collaborative design prototypes are based on C/S network architecture which has some weaknesses, such as a single point of failure and bottleneck, low expansibility and higher maintenance costs. This paper proposed a more effective and efficient scheme of real time collaborative parametric feature-based CAD modeling through concurrently programming 3D CAD model macro file with a procedural 3D CAD scripting language within a Peer-to-Peer (P2P) based collaborative editing system of our PRC platform, developed on the top of an improved JXTA. This paper describes the hybrid P2P network infrastructure, the procedural 3D CAD language, and a hierarchical, dynamic locking approach for consistent maintenance in P2P based collaborative editing. The rudimentarily developed prototype has verified that the proposed scheme is feasible.

Keywords: Collaborative design, Collaborative feature-based CAD modeling, Peer-to-peer, Real time collaborative editing, locking, Consistency maintenance.

1   Introduction

Under the context of global economy and concurrent engineering, complex product development has already shifted to systematic, integrated, and collaborative processes through taking advantage of increasingly globalized market and widely distributed resources to enable new products to launch the right market in the right place at right time. Collaboration is the successful coordination and communication among multiple functional departments (e.g., demand planning, sourcing, procurement, manufacturing, assembly and distribution) as well as external upstream and downstream trading partners, including designers, manufacturers, material suppliers, distributors and other vendors. Collaboration has already been a key factor for the success of complex product development.

   Collaborative design is the process in which participants from different disciplines share their knowledge about both the design process and the design content, in order to create shared understanding on both aspects, to integrate and explore their

knowledge and to achieve their common objective: the new designed product. Obviously, it is essential and necessary for participants working in geographically dispersed product development teams to be supported sufficiently by an effective and efficient distributed collaborative environment.

P2P network are getting popular in the recent Internet computing environment. It was crowned by Fortune as one of the four technologies that will shape the Internet's future. Although numerous studies have been carried out in the collaborative design research area, only a limited amount of researches were concentrated on P2P network, and scarcely any in the field of mechanical engineering. More, most developed collaborative design tools and systems were aimed mostly at the level of visualization, review, data sharing, and multi-media conference, did not support concurrently direct interoperation of CAD model, rather, real time co-design which is one of the most critical, meaningful and practical tools for collaborative design activities.

Motivated by the ultimate goal of establishing a product realization collaboration (PRC) environment targeted in making good use of computing power and other resources from various participants, widely deployed, large scalability, supporting highly dynamic and heterogeneous, fault-tolerance and load-balancing, etc., for large companies or virtual enterprises (VE) which are comprised of many SMEs (Small and Medium sized Enterprises), to support effectively engineering product collaborative design and manufacturing within globally dispersed participants with different backgrounds and contexts across professional, organizational and geographical boundaries, during the whole product realization process under concurrent engineering policy, the main objective of this research is to develop one kind of the most representative collaborative design applications, a P2P based real time collaborative parametric feature-based 3D CAD modeling system, located on the top of our hybrid P2P based PRC platform. The real time collaborative parametric feature-based 3D CAD modeling is mainly achieved by concurrently programming a tree-structured 3D CAD macro file with a novel procedural 3D CAD scripting language within a P2P based collaborative editing system of our PRC platform.

This paper is organized as the follows. Section 2 is devoted to introduce some related works. Section 3 introduces our hybrid P2P based PRC network infrastructure based on improved JXTA. Section 4 describes a hierarchical dynamic locking algorithm for real time collaborative editing. A novel procedural 3D CAD scripting language for real time collaborative parametric feature-based CAD modeling is presented in Section 5. Finally, concluding remarks are given in Section 6.

2   Related Works

Over the last decades, collaborative product design has received considerable attention from academia and growing interest from industrial community. Research efforts in the areas of collaborative product design substantially contributed to the improvement of collaborative design environment. Various academic prototypes have been developed to support geographically distributed collaborative product design activities. Some famous high-end commercial CAD/CAM, PDM/PLM systems also have provided a broad range of significant proportion of collaborative design

functionalities, amplified by continuous enhancements in interactive and visualization capabilities, and profiting from the availability of ever faster and more powerful hardware. Collaborative product design has experienced some major technological innovations and paradigm shifts.

Nearly all of these prototypes are C/S (client/server) or B/S (browser/server) architectures. In this kind of collaborative design environment, the system grouping, operation, and communication all have to rely on the central server, single point failure and bottleneck always rise because of limited bandwidth.

Bidarra, R., et al. [1] developed a Web based collaborative feature modeling system with C/S architecture, named WebSPIFF. The fat server offers all feature modeling functionalities. The thin clients implemented by Java3D could directly visualize and manipulate a feature model in real-time interactive manner by the use of visualization model, selection model and feature skeleton. A "traffic light" was used to ensure only one client to do exclusive modifications.

Li, W.D., et al. [2] developed a collaborative design environment with manipulation client/modeling server scenario to accomplish a feature-based design task. Based on feature-to-feature relationships, a distributed feature manipulation method was used to filter the varied information of a working part during a co-design activity to facilitate efficient information exchange for large-size 3D models. They chose the way of: Open CASCADE models (from) ― VRML models ― Java3D-based VRML browser for display and manipulation, in which they also designed a new visualization format based on features and VRML to preserve the information for the high-level features.

CollabCAD^{TM} [3] is a commercial collaborative framework for CAD & PLM, and a J2EE based distributed 2D&3D CAD/CAM/CAE software system with modeling client/communication server model for the collaborative industrial product design. It is a 3-tier distributed computing architecture, developed using Java3D and RMI. CollabCAD employs event-transmission to allow other participating sites to observe what is happening at one site by executing received events individually.

OneSpace.net^{TM} [4] is another commercial system with manipulation client/modeling server model to provide some collaborative and connective facilities to support Web-based multi-party 3D CAD design reviews smoothly, and to enable key documents and engineering data to be shared together safely and securely with suppliers, remote team members and customers by integrating seamlessly into PLM, ERP, and other applications.

In contrast with C/S or B/S architectures, the field of research on P2P-based collaborative design system is relatively young, and very few publications can be found, in particular hardly any in the field of mechanical product design.

Fan, L.Q., et al. [5] developed a distributed collaborative design framework, involving multiple parties across multiple domains and multiple enterprises' boundaries, with a hybrid of grid and peer-to-peer technology. In order to access computational resources for design, analysis and process simulation, a meta-scheduler was designed and implemented. It helped in resource discovery and optimal utilization of resources. It integrated model compression techniques to enable efficient data transfer and visualization in a distributed mode. A test bed was established to demonstrate a distributed collaborative design and manufacturing environment.

3   P2P Based Collaborative Platform

P2P technology is a loosely defined term that encompasses several definitions. For this research, P2P refers to the end-user application-level computing environment. Intel's David Barkai formally defines P2P computing as "a network-based computing model for applications where computers share resources via direct exchanges between the participating computers." There are two classes of P2P networks: structured and unstructured. In unstructured P2P networks, there are three categories: pure P2P, hybrid P2P, and centralized P2P systems.

P2P is a distributed application architecture which comprises a class of systems and applications that employ distributed resources to perform a critical function in a decentralized manner, by partitioning tasks or workloads among peers. Within P2P model, peers are equally privileged participants in the application. Each peer in the P2P network is referred to as a node, and would set aside a portion of its resources. The resources encompass computing power, storage capacity, data/content, network bandwidth, and presence (computers, human, and other resources), etc., directly available to other network participants, without the need for central coordination by servers or stable hosts. The critical function can be distributed computing, data/content sharing, communication and collaboration, or platform services.

3.1   Why Select a Hybrid P2P Paradigm as Our PRC Solution?

Our PRC applications should be accessible efficiently at anytime and anywhere, to a massive number of concurrent participants geographically dispersed. In addition, PRC network infrastructure need support and accomplish some particular engineering tasks, e.g. massive engineering analyzing and computing. As a result, PRC is targeted in making good use of computing power and other resources from various participants, widely deployed, large scalability, supporting highly dynamic and heterogeneous, fault-tolerance and load-balancing, etc.

Decentralized P2P approaches are more efficient at providing such guarantees for PRC than C/S model. Firstly, some advantages of P2P paradigm make P2P networking more effective than C/S model in PRC applications, such as sharing various resources, better scalability, enhanced load balancing, dynamic information repositories, redundancy and fault tolerance, low setup and maintenance cost, etc. Secondly, P2P network can greatly increase the utilization of information, bandwidth, and computing resources, which are currently the most valuable three components of the Internet. Thirdly, the current acknowledged applications based on P2P network are focused on three different application domains: computing, collaboration and file sharing, which are just the needed main applications for collaborative design systems. Fourthly, to a great extent, peers communications in P2P network are just similar with participants interactions within collaborative design activities. Fifthly, as a matter of fact, the appearance of real P2P applications has fostered the popularity of P2P networks, which have therefore proven their stability and correctness as a substrate for global concurrent access to distributed resources.

Combined with considering some disadvantages of P2P paradigm such as the security, peer and resource management, apparently, it is beneficial to follow a hybrid solution in addressing such large number of the identified requirements. In fact, the hybrid solution suggested in this research combines a number of mechanisms. Hybrid P2P systems are interesting alternatives to pure system designs since they can overcome the limitations of the original approaches. Exploitation has shown that hybrid systems are usually the ones that are widely deployed and extensively used. Though initially hybrid P2P systems were targeted at merging together the P2P and C/S paradigms in different services, they have been extended to explore a much wider range of combinations. For example, super-nodes are expected to be endpoints with high bandwidth and fast access to the collaborative network and act as landmarks for each network domain. Doing so, they have enough abilities to successfully address the complex requirements of our PRC environment.

3.2   Implementation

Based on literature review and informal discussions, we select JXTA [6] as the main development platform for PRC infrastructure. JXTA (Juxtapose) is a hybrid P2P platform for developing interoperable P2P applications. It is an open source project initiated by Sun Microsystems, and now carried on by many contributors from the open-source community. Unlike most P2P systems which are mainly developed for specific purpose, JXTA provides a basic framework on top of which developers can concentrate on functionality of their own applications, without worrying about low-level details of the P2P system.
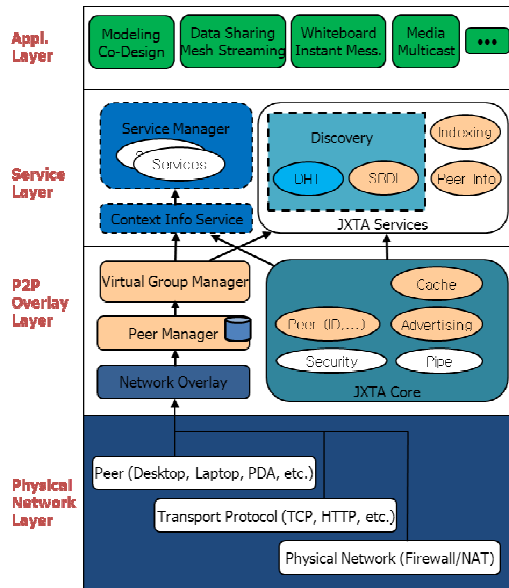
Fig. 1. PRC platform architecture

Hierarchical architecture almost always accompanies the large-scale, complex distributed systems. Based on JXTA, we explore a well-balanced hybrid P2P hierarchical architecture towards accomplishing the fulfillment of the targeted requirements. PRC platform architecture enables the effective usage of JXTA by successfully addressing their flaws, e.g. improving a loosely-consistent distributed hash table (DHT) in JXTA with pure DHT mechanism in Chord [7].

PRC platform architecture includes four layers of physical network, overlay network (expanded from JXTA core layer), services (expanded from JXTA service layer), and applications, as shown in Figure 1. Further, we develop a hybrid P2P based PRC prototype, on the top of which collaborative parametric feature modeling is achieved in a P2P based real time collaborative editing system with a procedural 3D CAD scripting language.

4   Consistency Maintenance in P2P Based Real-Time Collaborative Editing System

This is one of the representational scenarios in PRC, in which a geographically distributed product design team is working together, collaboratively programming the shared 3D CAD model scripting files within a real time collaborative editing system (RTCES) to achieve real time co-design and solve the interdisciplinary and complex problems, increasingly sharing knowledge, ideas and experiences, and systematically interacting as closely and productively as all the team members within the same room.

Distributed architectures and replicated approaches are typically employed in current RTCES in order to enable concurrent changes among the active users, and maintain high local responsiveness, etc. CCI (convergence, causality preservation, and intention preservation) model, proposed by Sun, S. et al. [8], is the current benchmark standard for consistency maintenance in RTCES. Locking is one of the techniques used to ensure consistency and data integrity in distributed systems, but locking has the disadvantage of reducing concurrent access. Operational transformation (OT) is the most popular way to maintain real time, high responsiveness while striving to achieve the CCI model, and ensure consistency among copies of a shared document in RTCES that employ replication of document state.

However, traditional OT algorithms act the whole document as a linear sequence of data, costly with regard to computation, communication, and the total memory required in storing the history buffers among all users. Additionally, OT approaches do not scale for a large number of operations and a large number of users. Further, the performance of OT algorithms degrades as the document size increases, so it is advantageous to minimize the employed document size. On the contrary, P2P paradigms may allow each peer to manage a section (sub-tree) of the document tree, are effective in load balancing work among peers and avoiding a single point of failure and bottleneck in processing peer actions.

In PRC, we adopt an adjustable locking policy that is established on a per-section (per-node) basis. At various depths (levels) within the tree-structure document, we implement a hierarchical, dynamic locking approach and data structure which are incorporated with existing OT algorithm for our P2P based RTCES scenarios, so that all active peers are able to concurrently make changes (operations) locally within

locked (or partially-locked) sections of a tree structure document and adopt OT as needed (when many users make changes to a single section of the document), in order to reduce the communication and computation costs, solve the problem of CCI, and improve to be more scalable in supporting larger collaborations.

We begin by assuming that initially the RTCES contains no users; all active peers have a copy of the shared document on their local site within the distributed editing systems; a peer will always own the largest sub-tree that does not conflict with any other user's sub-tree. Whenever a peer enters or leaves the RTCES (or moves from one section to another), locks are promoted or demoted.
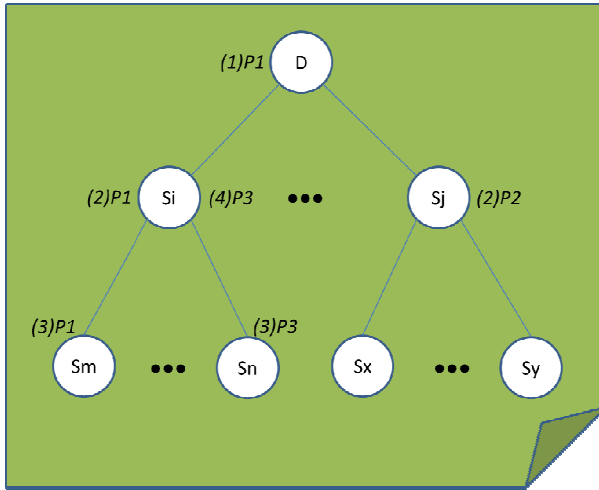
Fig. 2. Distributing dynamic looking at various levels within the document tree

Figure 2 demonstrates the distributing dynamic locking at various levels within the document tree when a user enters or leaves the editing session within RTCES. Here, assuming the dynamic locking policy has been adopted and sharing is not permitted.

(1) When a user, *P1*, first enters or initiates a collaborative session to edit a CAD model file *D* with tree structure in the P2P based RTCES, this user *P1* is the only one in this session and consequently has the entire document *D* updated and cached in his computer. Since there is no contention, user *P1* edits the entire document *D* without conflict from any other users.

(2) When another user, *P2*, enters the same session and desires to edit a disparate section rooted at node *Sj* within the document *D, P1*'s portion of the document is reduced to accommodate the new user such that the contention between *P1* and *P2* is removed. User *P1*'s lock is demoted to node *Si* that is the largest sub-tree which does not conflict with any other user's sub-tree, and currently his editing active section is belong to. Any changes made to the sections rooted at *Sj* are communicated to user *P2*. If the section requested to edit by *P2* is as same as the section which *P1* is currently editing, *P2*'s request will be rejected (or an OT-based multi-writer policy may be adopted).

(3) At this time, if the third user, *P3*, enters the same session and desires to edit a sub-section of *Si*, rooted at *Sn*. According to the aforementioned same policy, we suppose that user *P1*'s lock is demoted to node *Sm*, and any changes made to the sections rooted at *Sn* are communicated to user *P3*. At this point, user *P1* contains the current version of the section rooted at *Sm*, User *P2* contains the current version of the section rooted at *Sj*, and user *P3* contains the current version of the section rooted at *Sn*. If changes are made by user *P1*, *P2*, or *P3* at this point, those changes are not sent to any other user since they are in disparate sections.

(4) If user *P1* leaves the session or moves to another section, user *P3*'s lock on *Sn* should be promoted to *Si*, and changes made to the section rooted by *Sm* should be sent to user *P3*. We know this replay of changes to user *P3*'s copy of section rooted at *Sm* is non-conflicting with any changes made to section rooted at *Sn* because *Sm* and *Sn* are disparate. As a result, the history buffer rooted at *Si* may be defined as the merging of the history buffers of *Sm* and *Sn* incurring little computational cost.

5   Collaborative Parametric Feature-Based CAD Modeling with a Procedural 3D CAD Scripting Language

Our previous real time collaborative design modification [11, 12] was implemented by interoperating 3D CAD file in VRML format within a shared workspace on the P2P based RTCES, based on JXTA. However, VRML format had some limitations to represent current parametric feature-based CAD modeling in practical engineering applications.

Product models can be generated with parameters, features and constraints in parametric feature-based CAD systems. Such models can easily be edited for a variety of downstream applications. Parameters are dimensional variables within a geometric model. They provide indications of what dimensions are permissible to change. Features are parametric shapes with some attributes and have a semantically higher level than primitive closed regular sets. By feature based modeling, a designer need not work from the low level or individual curve and surface elements, also it forms a basis for linking CAD with downstream manufacturing applications. Constraints and specified relationships between geometric or topological elements together provide invariant characteristics in the model, often in the interest of maintaining product functionality during modification. The designer's choice of parameters, features, and constraints constitutes an important part of what is known as design intent. Following Bill Anderson's definition [13], design intent is defined as the functional requirements provided by customers; that is, a set of geometric and functional rules which the final product have to satisfy. The design intent is represented by parameters, constraints, features and design history. By analyzing the modeling sequence, the designer's design intent can be implicitly translated.

Existing CAD data format standards such as STEP and IGES are neutral, geometric-based formats that are commonly used to exchange basic geometric and topology data of shape models of the boundary representation (B-rep) between different CAD systems. These format standards cannot be used to transfer and

preserve the design information concerning the construction history, features, constraints, and other parametric-based CAD data. Hence, design-intent, semantic-level information, and the ability to perform subsequent high-level modifications will be lost after exchanging models between different CAD systems. Therefore, achieving such higher-level functionalities often requires a time-consuming, error-prone, tedious process of manually recreating the model in the target CAD system. As a result, these format standards are inadequate for modifying, extending, and performing other important higher-level functionalities on CAD models within other collaborative engineering applications.
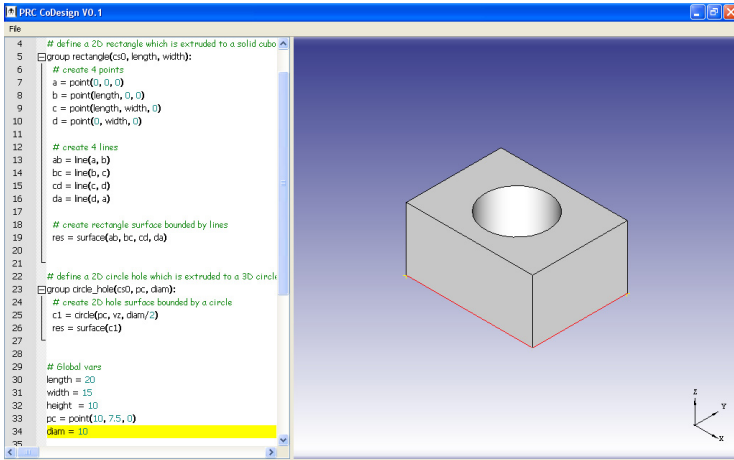
Fig. 3. Collaborative feature modeling by collaboratively programming 3D CAD model file

Based on techniques adapted from programming language research, we analogy and treat CAD modeling in CAD system as source code programing in software development environment, apparently their outputs, CAD models correspond to program files written in a kind of program language. Within a feature-based modeling system, CAD model can be built entirely by a sequence of insertion, modification, and deletion operations on some predefined design features. Correspondingly, feature-based CAD model can be programmed with a kind of structured, object-oriented language, wherein the design features are defined as subroutines or classes.

To image CAD co-modeling in collaborative product design as the distributed source code programing in the collaborative software development, there are many similar scenarios. A procedural 3D CAD scripting language has been developed rudimentarily to support real time collaborative parametric feature modeling in PRC.

With such procedural programming approach, a parametric feature based 3D CAD modeling is achieved by accompanying editable shape representation with a program written in a compact high-level scripting language, and saved as a CAD model macro file in the form of hierarchical tree structure, as shown in Figure 3. The geometry interpreter (or compiler) translates the elements of the editable shape representation through a suitable sequence of geometric operations carried out by the underlying geometric core modeler, based on Open CASCADE.

6   Conclusions and Future Works

This paper explores a hybrid P2P based real time collaborative parametric feature-based 3D CAD modeling through concurrently programming a 3D CAD model macro file with a procedural 3D CAD scripting language within a P2P based collaborative editing system of our PRC platform, developed on the top of improved JXTA.

However, there are still some technical problems to be addressed later, such as undo/redo problem in collaborative editing system, interface between our procedural 3D CAD scripting language and some common CAD format standards, famous commercial CAD software, etc. More, the functions of our procedural 3D CAD scripting language should be further enhanced in future.

References

1. Bidarra, R., van den Berg, E., Bronsvoort, W.F.: A Collaborative Feature Modeling System. Transactions of the ASME 2, 192–198 (2002)
2. Li, W.D., Ong, S.K., Fuh, J.Y.H., Wong, Y.S., Lu, Y.Q., Nee, A.Y.C.: Feature-based design in a distributed and collaborative environment. Computer-Aided Design 36, 775–797 (2004)
3. CollabCAD^{TM}, National Informatics Centre, India, `http://www.collabcad.com`
4. OneSpace^{TM}, CoCreate Inc., `http://www.onespace.com`
5. Fan, L.Q., Senthil Kumar, A., Jagdish, B.N., Bok, S.H.: Development of a distributed collaborative design framework within peer-to-peer environment. Computer-Aided Design 40, 891–904 (2008)
6. The JXTA Project, `http://www.jxta.org`, `http://jxta.dev.java.net`
7. The Chord official web site, `http://pdos.csail.mit.edu/chord/`
8. Sun, S., Ellis, C.: Operational Transformation in Real-Time Group Editor: Issues, Algorithms, and Achievements. In: Proceedings of ACM Conference on Computer Supported Cooperative Work, Seattle, USA, pp. 59–68 (1998)
9. Sun, C., Jia, X., Zhang, Y., Yang, Y., Chen, D.: Achieving convergence, causality-preservation, and intention-preservation in real-time cooperative editing systems. ACM Transactions on Computer-Human Interaction 5(1), 63–108 (1998)
10. Sun, C., Sosič, R.: Optional Locking Integrated with Operational Transformation in Distributed Real-Time Group Editors. In: Proceedings of the 18th ACM Symposium on Principles of Distributed Computing, Atlanta, GA, USA, pp. 43–52 (1999)
11. Wang, L., Wang, J., Hagiwara, I.: Investigation to Peer-to-Peer-based collaborative working platform for product development. International Journal of Internet Manufacturing and Services 1(2), 194–212 (2008)
12. Wang, L., Wang, J., Sun, L., Hagiwara, I.: A Peer-to-Peer Based Communication Environment for Synchronous Collaborative Product Design. In: Luo, Y. (ed.) CDVE 2007. LNCS, vol. 4674, pp. 9–20. Springer, Heidelberg (2007)
13. Anderson, B.: ENGEN data model: a neutral model to capture designintent. In: PROLAMAT 1998 (1998)
14. Altidor, J.: A programming language approach to parametric CAD data exchange. In: DETC2011-48530, Proceedings of the ASME 2011 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, IDETC/CIE 2011, Washington, DC, USA, August 28-31 (2011)
15. Mun, D., et al.: A set of standard modeling commands for the history-based parametric approach. Computer-Aided Design 35, 1171–1179 (2003)

Development of Software Module for Model Reference Adaptive Control Simulation Based on the LabVIEW

Yankai Wang, Zhaoyu Zhang, and Wenjie Qiu

Key Laboratory of Dynamics and Control of Flight Vehicle, Ministry of Education, Beijing Institute of Technology, Beijing, China
wangyankai@bit.edu.Cn

Abstract. Missile attitude control system usually adopts fixed gain with feedback control and the design method of combining with using the classical autopilot approach. But the missile's dynamic characteristics and parameters varies with the environment changes dramatically, the traditional method is sometimes hard to meet the guidance performance requirements. According to the missile control system as the research object, based on the theory of adaptive control of missile attitude stabilization loop control method, and LabVIEW control simulation development environment to achieve the model reference adaptive control simulation software module, and the design of missile attitude control system simulation experiment. This design method can be used for missile attitude control system analysis, design and effectively assisting the designer to design the control system, simulation, optimization and programming, the linear system and nonlinear system by computer simulation, through adjustment of controller parameters to obtain satisfactory control effect online. It has a certain practical reference value for the actual design work.

Keywords: model reference adaptive control(MRAC), LabVIEW, attitude control system.

1   Introduction

The trend of the development of current missile attitude control system is digital, adaptive realizable, intellectualized and multi-functional. With the development of computer technology and modern control theory, advanced control theory adaptive control has been introduced into digital missile control system and it certainly will become the major development trend of stable control of missile. Therefore, research on design method of missile attitude control system is a very important subject. Missile attitude control system usually adopts design method which combines fixed gain with feedback control, namely method of classical autopilot. However, dynamic characteristics and parameters of missile change greatly are according to environmental conditions. Dynamic characteristics with violent changes of random disturbance and dynamic parameter of system are sometimes hardly to meet hardly meet the requirements of guidance performance. Various changes in features of modern war put higher requirements for maneuvering performance of missile. Therefore, finding a kind

of design method of missile attitude control system based on advanced control theory and control technology has become became a common hotspot for researchers in the field of missile attitude controlling [5].

Graphic program developing environment LabVIEW has been developed from the simplest tool—data collecting and instrument controlling to the graphic frame of design and publishing virtual instrument software for scientific and technical personnel since its appearenced in 1986. It has also become the standard software platform of test and measurement industry and control industry. A few latest versions of LabVIEW contain control design and simulation tool kit which contain a series of module, such as modeling and transformation of transfer function, time-domain and frequency-domain performance analysis of transfer function and system identification. The tool kit has various graphic tools, such as bode diagram and root locus diagram which can be used in verification of controller's performance, data processing, modeling and system analysis. Application of LabVIEW also includes the usage of RT control design module function in MathScript to design and analyze controller and transform Simulink modules into LabVIEW program block diagram model. LabVIEW has a broad application prospects in field of control design and simulation.

This paper regards control system of missile as research object. A control method of missile attitude stable loop based on adaptive control theory will be designed in this paper. The calling function of MATLAB Script node for MATLAB will be used to achieve model reference adaptive controller design of attitude control of missile based on LabVIEW and the detailed simulation result analysis. This paper will also combine theoretical analysis with design simulation to explore the adaptive controller design of missile attitude channel. It has thoretical and practical value.

2   Missile Longitudinal Motion Adaptive Control System Modeling

Establishing mathematical model of description of missile motion is the foundation for analyzing, designing and simulating missile control system. Regarding missile as a rigid body, then its motion will be considered as the combination of translational motion of center of mass and rotation motion around center of mass. That is six degrees of freedom which decide instantaneous position and attitude of the center of mass of rigid body. For the simplicity of modeling of attitude control system, firstly establish the motion model of attitude channel in the quasi-body coordinate system. Then obtain the dynamics transfer function by small disturbance linearization so that the missile attitude control system will eventually be established.

The change of aerodynamic parameters is due to the flight velocity, the flight height and the atmospheric density. So the mass and the center of mass will change along with the fuel consumption. Therefore, the mathematical model parameters of missile can change in a wide range. When the mathematical model parameters changes in a small range, the harmful effect of parameter variation on control quality can be eliminated or reduced by using conventional feedback control, optimal control and compensation control. But for the case of wide range parameter variation, in order to ensure that the system can still work or be close to the optimal working state, the adaptive control problem is proposed [1].

2.1   Classical Attitude Control Loop of Missile

The space motion of missile can be decomposed into longitudinal motion in the vertical plane and lateral motion in the horizontal plane to simplify the motion problem of missile. Because angular motion causes change in missile attitude, angular motion stabilization equals to missile attitude stabilization. The device to stabilize the parameters of elevation angular motion elevation is usually called longitudinal autopilot. And devices to stabilize the parameters of yaw and tilt angular motion are respectively called yaw autopilot and tilt autopilot. Stable loop of missile's center of mass is designed to control missile's flight along the predetermined orbit. For axisymmetric missile and aerodynamic missile, longitudinal autopilot and yaw autopilot are the same. In this paper, only the stable loop of center of mass is studied.

There is a series of disadvantages existing in the model of missile body, such as high oscillation property in transient process, low dynamic quality and poor anti-interference ability. In order to overcome these disadvantages, a perfect guidance and control system is required. Close loop which consists of autopilots and missile body is used in the classical control theory design scheme of missile stable loop. That is attitude control system stable loop. An autopilot may contain several sensitive elements, and every sensitive element can be formed into a loop. So a stable loop itself could be a multi loop. We usually choose the design scheme of stable control loop which contains an angular rate gyro and a linear accelerometer. The damping loop composed by angular rate gyro can be used to adjust the equivalent damping of missile. The overload loop composed by linear accelerometer can be used to ensure the steady state accuracy of missile and to improve the anti-interference ability. To design a pitching channel autopilot, we can use an autopilot consisted of one linear accelerometer and on angular rate gyro to realize. The block diagram of the pitching channel autopilot as shown in Figure 1.
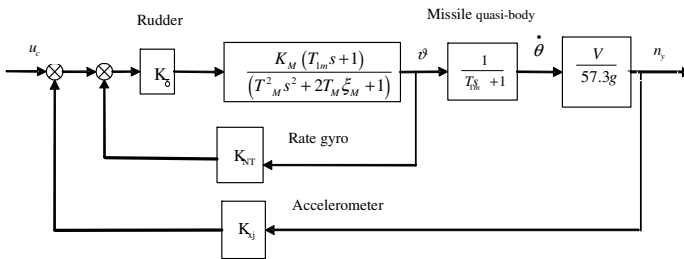
Fig. 1. The block diagram of the pitching channel autopilot

Simplification of the system model includes: ①the actuator loop of system can be simplified into constant K_δ when deriving the transfer function of controlled object; ② the rate gyro feedback loop of system can be used to improve the damping ratio, stability and robustness. The time constant of velocity measurement gyro is much less than the time constant of missile body. So the original second-order system can be considered as an inertialess amplification, section of which the transfer coefficient is

K_{NT} ; ③the control loop of missile is an instruction control loop which combines the damping loop and a lateral linear acceleration negative feedback. It is used to measure the lateral linear acceleration $v_d\dot{\theta}$ (actually it is used to measure overload n_y).

2.2   Design Method of Adaptive Controller

2.2.1   Model Reference Adaptive Control

The model reference adaptive system is a common adaptive system. It has already been many design approaches, some of them have matured, and some of them are still developing. Based on the viewpoint of engineering, we hope that the design system can maintain balance between performance and complexity. In order to simplify the adaptive system, we hope that linear or nonlinear equations don't need to be solved to determine the adaptive law. So the design problem of model reference adaptive system can be regarded as auto adjustment problem of the parameters and states of system when they deviate from their equilibrant positions [2].
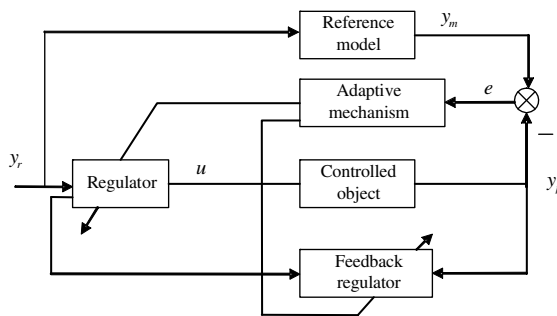
Fig. 2. The basic structure of model reference adaptive control system

The basic structure of model reference adaptive control system is shown as in Figure 2. It is consists of reference model, adjustable system including controlled object, preregulator and feedback regulator and adaptive device. The requirements of performance of an adjustable system such as overshoot, transition time and pass band can be directly set up by reference model. The output of the reference model y_m represents anticipant dynamic response of system. It makes model reference adaptive control different from other control approach, because there is no need for system to transform the performance index. When there is difference between the model and the actual output of system, the difference can be detected by comparator and the parameter of regulator including previa controller and feedback controller will be changed or the auxiliary signal of regulator will be generated after adaptive device making a decision. Then the output of actual system adjustable system will be consistent with the output of reference model. Eventually, the generalized error ε will tend to minimum value or reduce to zero.

2.2.2   Narendra Stable Adaptive Controller

The approach to compose adaptive law by using state variables of the controlled object involves all state variables. However, for lots of actual control systems, it is impossible to obtain all state variables. Considering the realizability of the designed model reference adaptive control approach, we use Lyapunov direct method to design an adaptive law according to the input and output data of controlled object. The adaptive law will adjust the parameter of adjusting controller and will directly change the data added to the input end of controlled object so that the transfer function of adjustable system which is made of controller and controlled object is consistent with the transfer function of reference model. It is a way to simplify parameter identification and system structure. It has obvious superiority. Among approaches using the input and output data, some of them (such as single input single output adaptive system) involve all-order derivatives of the output of controlled object and generalized error. Therefore, differentiators are needed in the adaptive device. As a result, the anti-disturbance ability of adaptive system will be decrease. To avoid this kind of result, K.S.Narendra proposed an approach of stable adaptive controller which needed derivatives of the output of controlled object and generalized error. A brief explanation is given below.

Firstly there are some requirements when the approach is applying. The reference model $G_m(s)$ show strictly $G_m(s) = k_m N_m(s) / D_m(s)$. Both $N_m(s)$ and $D_m(s)$ are Hurw it polynomial of which the leading coefficient is 1. And the order numbers are respectively m and n. k_m is the gain of reference model. To realize complete matching of adjustable system and reference model, the adaptive controller must have enough adjustable parameters. When the denominator of transfer function of controlled object is n-order and numerator is m-order, plus magnification factor, controller has n+m+1 adjustable parameter. Therefore, the adaptive device must have n+m+1 adjustable parameter to correspond to. According to the requirement of modeling, consider the case of n-m=2 and the schematic diagram of the designed Narendra adaptive controller is shown as in figure 3 below.
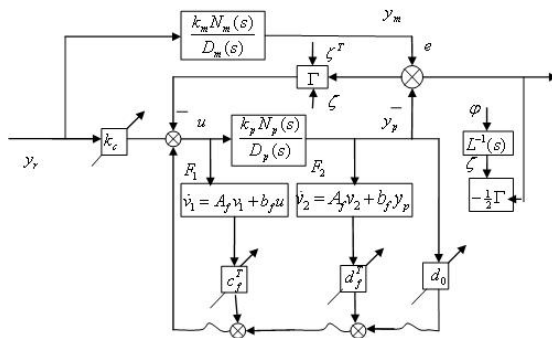
Fig. 3. The schematic diagram of the designed Narendra adaptive controller

Conclusively, applying Lyapunov stable theory to the design of model reference adaptive controller can both improve performance of control system and ensure stability of system. This is a feasible method and achieved wide application in engineering practice. However, for a practical system, there isn't a general way to find the best Lyapunov function. It differs according to the level of designer. The adjusting law and performance of control system are directly decided by Lyapunov function. It is also the reason why this kind of algorithm is restricted by some factors [2].

3 Model Reference Adaptive Controller Design Based on LabVIEW

Graphic program developing environment LabVIEW has developed from the most simple tool of data collecting and instrument controlling to the graphic frame of design and publishing virtual instrument software for scientific and technical personnel since its appearance in 1986. It has also become the standard software platform of test and measurement industry and control industry. A few latest versions of LabVIEW contain control design and simulation tool kit which contains a series of module, such as modeling and transformation of transfer function, time-domain and frequency-domain performance analysis of transfer function and system identification. The tool kit has various graphic tools, such as bode diagram and root locus diagram which can be used in verification of controller's performance, data processing, modeling and system analysis. Applications of LabVIEW also include using RT control design module function in MathScript to design and analyze controller and transforming Simulink modules into LabVIEW program block diagram model. LabVIEW has a wide application prospect in field of control design and simulation. First, select the reference model and controlled object model of adaptive controller. Then, analyze the performance stability, anti-disturbance ability and parameter variation ability of adaptive controller according to Narendra adaptive algorithm. At last, design the interface using LabVIEW. The output of reference model can be traced by the output of controlled object when the program is run continuously and achieve a more perfect control effect[6].

3.1 Simulation Software Design Method of Controller

There are several ways to realize adaptive control in LabVIEW. We can call external compiling C or C++ program by using CIN node. We can also compile or call MATLAB program by using MATLAB Script node. Or we can use graphical programming language of LabVIEW to compile. The approach of using MATLAB Script node has features such as easy to realize, multi input multi output, large information content of one time processing and able to debugging MATLAB program in advanced. The design procedure of simulation software of model reference adaptive controller is given below [3] .

① establish the function script of Mathscript node, introduce the function compiled in MATLAB into the script file, eliminate the function definition sentence and graph drawing sentence in the original file. ② select and add input and output from the quick-menu of the script node, the added input is in the form of array or number, output in the form of oscillogram. ③ express module by using the transfer function in the control design and simulation tool kit so that the numerator, denominator and the coefficient of transfer function of input can be transformed into the form of transfer function. ④ adjust the parameter input by using adaptive gain, and set up a knob input on front panel so that during continuous running on-line modification of adaptive parameter can be simulated to achieve the optimal control effect. ⑤ add missile attitude control model option card and adaptive controller instruction option card, complete the whole design procedure of Narendra adaptive controller of missile attitude channel.

3.2   Implementation of Model Reference Adaptive Control Based on LabVIEW

The designed interface is shown in Figure 4~5. There are four option cards: the model and parameter of missile attitude control, design block diagram of adaptive controller, parameter input of controlled object and reference model (Figure 4), and adjusting parameter input of controller (Figure 5). There are three simulation oscillograms display (including output of reference model ym(t), output of controlled object yp(t), generalized error e(t)=ym(t)-yp(t), control variable u(t)).
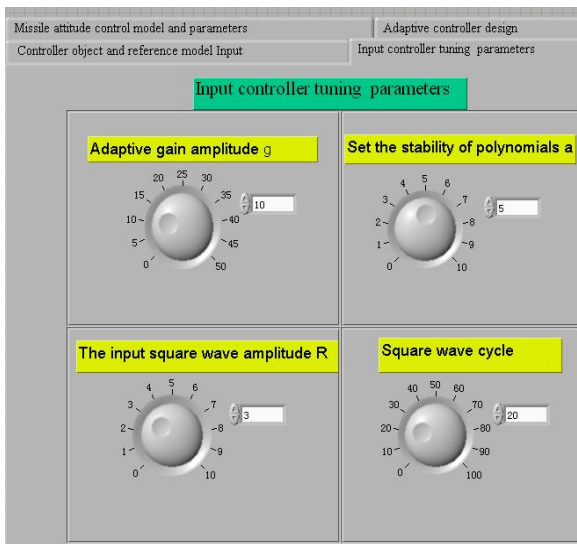
Fig. 4. Parameter input of controlled object and reference model
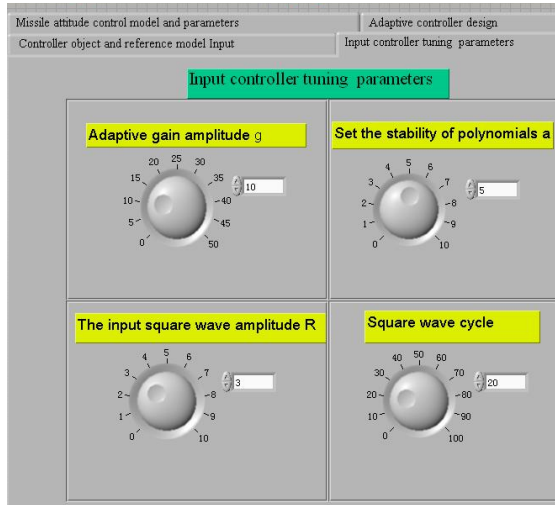
Fig. 5. Adjusting parameter input of controller

Figure 4 is the input module of controlled object and reference model. Figure is input module of controller adjusting parameter. The nonzero initial values need to be set up before program running. During continuous running of program, control effect can be observed from three oscillograms on the right. By on-line modification of controller adjusting parameter, the best control effect can be observed. The design is direct and simple. Add the numerator, denominator and coefficient of transfer function from high order to low order, during the program running, the transfer function will be automatically displayed on the right side where controlled object and reference model were displayed. It is convenient to if the input model is right. Figure shows block diagram of missile attitude control model. Model of controlled object is in the black block. The table below shows some related parameters of missile pitching channel. Figure contains two parts: operating instruction and design drawing of Narendra adaptive controller. The operating instructions include: ① Gp(s) is missile transfer function model that is not added to autopilot acceleration loop. ② Gm(s) is transfer function of reference model which has desired dynamic performance. ③ the numerator of Gp(s) and Gm(s) should keep the same. ④ press the continuous running button after setting up the controller adjusting parameter, and parameters can be adjusted on line according to output graphics (control result graphics) so that the best control result is achieved. ⑤ the marked yellow parts are must required items.

3.3   Simulation Software Operation

Front panel is the user operation interface of LabVIEW. On the program block diagram plate, objects on the front panel are controlled by using icons and lines. And the inputs

on front panel can be transferred to the corresponding program block diagrams on the background. The outputs of program block diagrams can be displayed in form of oscillogram and text on the front panel. In this way, after users input data on the front panel, they can also receive the outputs they need. It is convenient to debug program according to the output result. The interface is given in the Figure 6 below, during the program debugging.
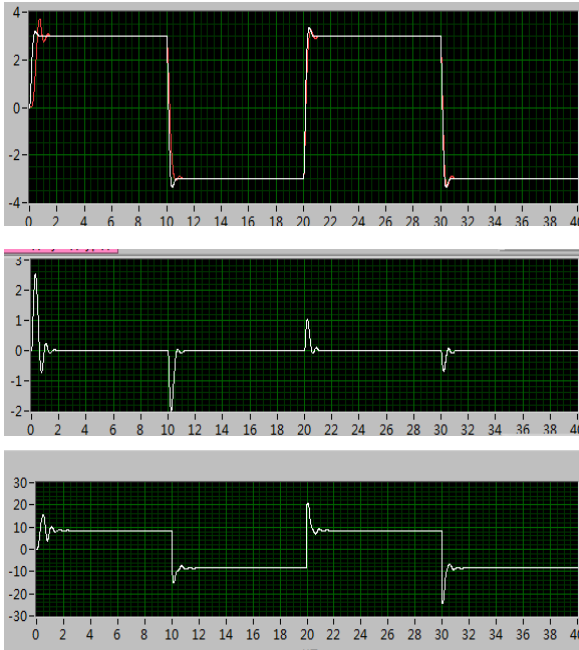
Fig. 6. Simulation Function interface

4   Conclusions

In this paper, firstly we get the transfer function of missile attitude channel control by processing small disturbance simplification for missile motion model. Then we get classical attitude control loop, and design the attitude controller based on model reference adaptive theory for missile longitudinal channel by combining classical control system with adaptive controller. The control effect of different kinds of model reference adaptive approach is compared and analyzed. Narendra adaptive control algorithm is chosen to be analyzed deeply. And LabVIEW is used to conduct the simulation software design for model reference adaptive controller. The on-line tracking of model reference output by controlled object output can be realized and achieve a perfect control effect. It shows the design model reference adaptive simulation software can effectively conduct design, simulation and analysis for missile

attitude with the presence of small disturbance. Virtual instrument technology and adaptive control is organically combined by using LabVIEW. Both application ranges are broadened. At the same time, the development platform of LabVIEW has a friendly man-machine interface and convenient operation, and it has improved the performance and development efficiency of control system.

References

1. Shin, Y., Calise, A.J.: Adaptive autopilot designs for an unmanned aerial vehicle. AIAA 2005-6166 (2005)
2. Pang, Z., Cui, H.: System identification and adaptive control of MATLAB simulation. Beihang University press, Beijing (2009)
3. Xu, M., Yu, Y.-M.: Using MATLAB in LabVIEW. Journal of Shandong University of Technology (Sci. & Tech.) 19(4), 92–95 (2005)
4. Sang, B., Jiang, C.S.: Integrated Guidance and Control for a Missile in the Pitch Plane Based upon Subspace Stabilization. In: CCDC 2009: 21st Chinese Control and Decision Conference, Guilin (2009)
5. Shi, K., Zhang, X.-Y.: Design and simulation of model reference adaptive control for air-to-surface missile. Flight Dynamics 6(28), 51–54 (2010)
6. Yang, Q., Dang, X.: Realization of Decoupling Control System Based on LabVIEW. Computer Engineering 31(2), 228–231 (2005)
7. Wu, Y.-H., Wang, Q.-Z.: Realization of Adaptive Control Reference Model Based on LabVIEW. Computer Technology and Development 18(11), 179–182 (2008)

MDA-Based Meta-modeling Technique
for Collaborative Simulation

Cheng Ma[*] and Tianyuan Xiao

The National CIMS ERC, Department of Automation,
Tsinghua University, Beijing 100084, China
mac05@mails.tsinghua.edu.cn

Abstract. Heterogeneity among multi-disciplinary models of complex products models hinders the sharing of information between subsystems, decreases the efficiency of collaborative development and becomes the bottle-neck of system integration and collaborative simulation. This paper analyses the current progress on model consistency description in collaborative simulation system for complex product development, and presents an UML Profile-based meta-model, CSMM (Collaborative Simulation Meta Model), referring MDA standard, to fulfill the consistency description of collaborative simulation model and collaborative behavior information. Also, the process for building collaborative simulation models of complex products based on CSMM meta-model is given. A case study has verified that this meta-model is able to support multi-granularity modeling and multi-resolution simulation in complex product development process.

Keywords: collaborative simulation, MDA, meta-model, CSMM.

1   Introduction

Complex products have the features of multi-discipline and multi-subsystem [1]. The product model is usually divided into different subsystems in accordance with the constitutional structures and subject fields, such as mechanical, electric, hydraulic, fluid, control, etc. And there are a large number of constraints and coupling among various subsystems [2]. The process for developing complex products is to reuse and integrate heterogeneous models described with different specifications and granularities and developed by modelers in different organizations in order to establish a system which is able to reflect different technique principles. To achieve the target for information sharing among different disciplines in the process of developing complex products and remove the obstacles in system integration and collaborative simulation brought by the heterogeneity of models, an integrative, reusable and hierarchical collaborative simulation unified meta-model is required for supporting the fast development of distributed collaborative simulation system. Meanwhile, in order to meet the constantly changing collaborative requirement, this meta-model must be able

to support the combination and reuse of the existing simulation models and collaborative behavior information.

MDA (Model Driven Architecture) is a software development methodology proposed by OMG (Object Management Group) on July, 2001 [3]. MDA provides supports for aspects such as modeling language, modeling method, mapping and transformation of models, verification for consistency and completeness of models by a series of standards. Applying MDA method to collaborative simulation system can improve the simulation application efficiency and bring good maintainability and expandability to the simulation system.

2   Related Study

In recent years, a lot of researches have been carried out to support the combinability, portability and reusability of heterogeneous simulation models. And the study of building unified meta-model for describing multi-disciplinary simulation models based on MDA has made some progress. Mature meta-model specifications or standards for modeling meta-models are BOM (Base Object Model), SRML (Simulation Reference Markup Language) and SMP2 (Simulation Model Portability 2.0).

The idea of MDA is reflected in BOM as it separates the design model and implementation model by defining conceptual model and object model separately [4]. However, during the mapping process between the conceptual model and the object model, the information of behavior semantic which cannot be expressed in object models is lost. In addition, BOM describes model behaviors in the form of state machine information table, which can neither express hierarchical and concurrent models nor support expressions of aggregation relationships between models.

SRML is the standard for model description in XMSF (eXtensible Modeling and Simulation Framework) [5]. Based on the theory of meta-model in MDA and Web Service technology, XMSF is the integration in a series of general modeling and simulation standards. It manages various simulation models and resources by building meta-data and establishing resource library of meta-models to guarantee the combinability and the scalability of the simulation system. However, SRML describes simulation behavior by inserting Java scripts into XML and is not technology-independent in the strict sense.

SMP2 is the portable specification for simulation models proposed by European Space Agency, borrowing ideas from MDA standard [6]. SMP2 uses a XML-based self-defined meta-model, SMDL (Simulation Model Definition Language), to describe the design and assembly information of the simulation model. SMDL is a self-describing meta-model which is not based on MOF (Meta Object Facility) standard. Therefore it is hard to support meta-modeling in various formal systems. Also, QVT (Query/View/Transformation) and other standards cannot be adopted to support the model transformation, which makes other heterogeneous models are difficult to transform into SMP2 models. As lack of support from CASE tools, SMDL increases the difficulty in modeling and hinders the promotion for development field of complex products.

These meta-modeling techniques or modeling languages all have the ability to describe multi-disciplinary heterogeneous simulation model at the conceptual level. But as the focuses on solving problems are different, they all have some limitations. Some of them just apply MDA thought into the modeling frame but don't strictly follow MDA standards such as MOF and QVT. Therefore, they can support heterogeneous simulation model integration only in theory, but have many shortcomings in practice.

Based on MDA method, this paper proposes a meta-model oriented to collaborative simulation for complex products following MOF standard in order to guide the modeling of simulation model and interaction semantics on the conceptual level and promote information sharing and system integration among various disciplines in the practice of complex products development.

3   Collaborative Simulation Meta-Model (CSMM)

3.1   Design Ideas

As new disciplines and concepts increase ceaselessly in simulated systems, it is hard to find a meta-model that contains all modeling structure sets of all fields. Therefore, this paper is not intended to propose a meta-model which is big and complete enough to cover all model semantics in simulation field. In this paper, the models are described as features of interaction capabilities and interfaces when they serve as participants of the collaborative simulation. What's more, the static relationships among models caused by the physical relevance as well as interactive behavior shown in plot of collaborative simulation tasks are also described. The process for establishing meta-model in collaborative simulation includes abstracting the common concept of model elements during the collaborative simulation, standardizing and formalizing these concepts, and analyzing the semantics and syntax dependence among these concepts so as to acquiring the ability to describe various meta-data in collaborative simulation.

This paper uses "entity" and "relationship" as the basic modeling elements to describe conceptual model in the design of the Collaborative Simulation Meta-Model (CSMM), and combines relevant concepts and mechanisms in object-oriented modeling methodology to extend CSMM.

3.2   Collaborative Simulation Meta-model

This paper establishes collaborative simulation meta-model by tailor the UML core meta-model and expands UML class diagram and UML sequence diagram based on UML profile mechanism, which enables it to describe the general concepts and collaborative scenarios in collaborative simulation field. The overall structure of CSMM is shown in Fig.1, including packets of CSMMKernel, CSMMDataTypes and CSMMInteractions.
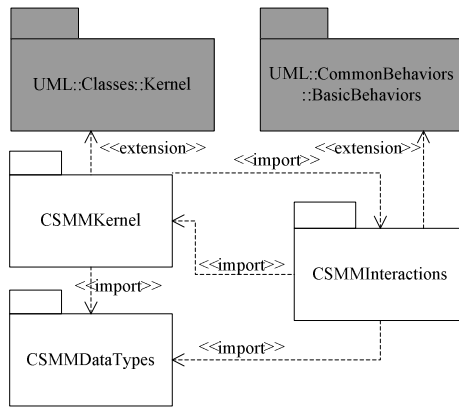
Fig. 1. Overall structure of CSMM

UML Profile extension mechanism extends the semantics of meta-class in UML standard by defining stereotypes and tagged values in order to express the semantic concept in a specific field. The stereotypes and tagged values defined in the three packets are separately introduced in the following sections.

(1) CSMMKernel packet

As a set of various static modeling elements required for collaborative simulation concept modeling, CSMMKernel packet includes simulation model description, simulation model interface information, model relationship, event description, static information of simulation task and etc. The stereotypes and tagged values in CSMMKernel packet and the relationship among them are shown in Fig. 2 and Fig.3.
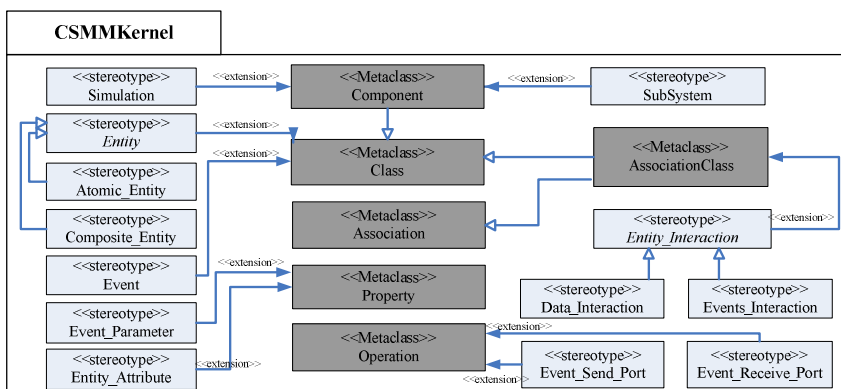
Fig. 2. Stereotypes in CSMMKernel packet
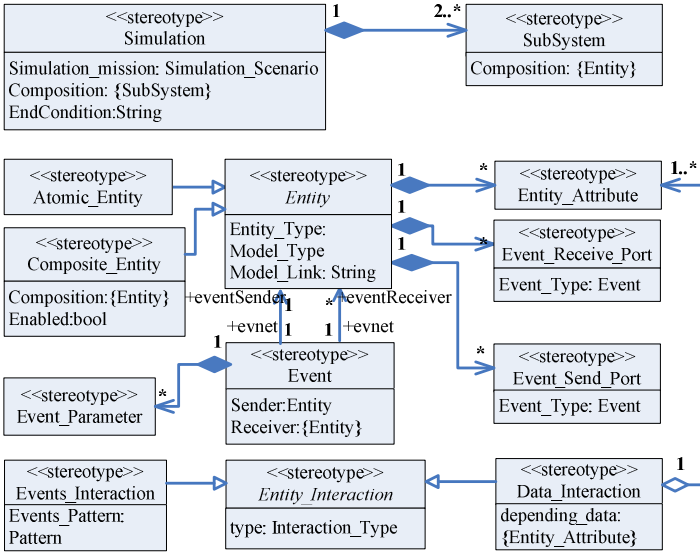
Fig. 3. Relationship among meta-models in CSMMKernel packet

As the most basic module in conceptual model, "*Entity*" is used to describe concepts, objects and systems in the real world. In CSMM, "Entity" is an abstract class, with two subclasses, "*Atomic_Entity*" and "*Composite_Entity*". Atomic_Entity is the minimum unit participating in collaborative simulation and it is indivisible. Composite_Entity consists of two or more entities, either Atomic_Entities or Composite_Entities. To meet the requirements for simulation tasks, Composite_Entity can be depolymerized into some sub-entities and sub-entities can also be polymerized into a Composite_Entity. The system totally depolymerized into Atomic_Entities is the collaborative simulation system with the highest resolution. As Composite_Entity aggregates the collaborative capacities of all Atomic_Entities which form the Composite_Entity, the attributes and event input/output ports shown is the union for the attributes and event input/output ports of all Atomic_Entities which consist of the Composite_Entity when it participating in a low resolution collaborative simulation.

"*Relationship*" is used to describe the possible relation among entities, including inheritance relationship (generation relationship), hierarchical relationship (inclusion relation between whole-part) and interaction relationship. Interaction relationship can be divided into data dependent relationship and event trigger interaction relationship. In CSMM, "*Entity_Interaction*" is used to describe the information interactive relationships among entities except inheritance relationship and hierarchical relationship during the process of collaborative simulation. it is an abstract class with two sub-classes, "*Data_Interaction*" and "*Event_Interaction*". Data_Interaction indicates there exists the data dependency relationship among simulation entities. CSMM defines Data_Interaction as a binary association class. What's more, it is of one-way direction, the beginning entity of the association is the receiver of dependent data and the terminal of association is the owner of dependent data. Event_Interaction

indicates that there exists event trigger interactive relationship among simulation entities. A series of event trigger interactions among some entity classes follow a predefined behavior pattern which is described in the form of extended sequence diagram in CSMM.

(2) CSMMDataTypes packet

By extending the basic datatypes of UML, CSMMDataTypes packet includes a series of predefined datatypes supporting collaborative simulation and various user-defined datatypes. The stereotypes and tagged values in CSMMDataTypes packet are shown in Fig.4.
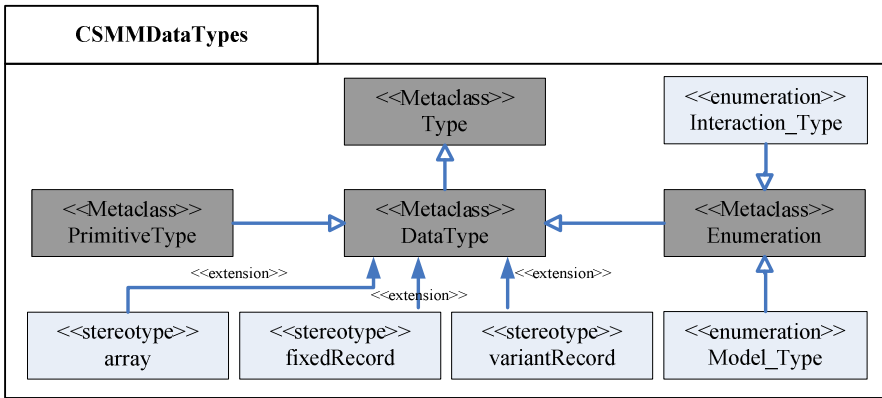
Fig. 4. Stereotypes in CSMMDataTypes packet

(3) CSMMInteractions packet

CSMMInteractions packet includes dynamic behavior information such as pattern-based interaction relationship among models and collaborative simulation scenarios. The stereotypes of CSMMInteractions packet are shown in Fig.5.
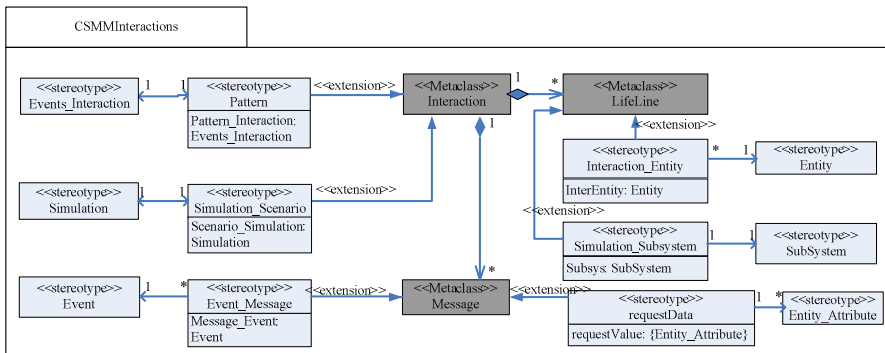
Fig. 5. Stereotypes in CSMMInteractions packet

3.3   Main Features of Collaborative Simulation Oriented Meta-model

Collaborative simulation meta-model, CSMM, provides a consistent solution for describing, modeling and managing collaborative information during the process of integrating multi-discipline subsystem models. CSMM meta-model has the following features

(1) Compliance with MOF specifications
(2) Accurate formalized description
(3) Independent of platform or middleware technologies
(4) Graphic modeling environment and model representation
(5) Supporting to inherent, reload, and combine model, or verify parameters to reuse the existing models
(6) Supporting static and dynamic collaborative information description simultaneously
(7) Supporting hierarchical modeling and multi-resolution simulation

4   The Modeling Process Based on CSMM

The CSMM-based modeling process can be divided into the following four stages.

4.1   Model Design Stage

The main work in this stage is to prepare class information of various kinds of Atomic_Entities, Composite_Entities and Entity_Interaction and it can be divided into the following six steps:

(1) Designing the stereotype of Atomic_ Entity class, including Entity_Attributes and its datatype, as well as operation function declarations for Event_Send_Port and Event_Receive_Port which describes the capability of Atomic_ Entity in sending and responding to events. Each event port corresponds with an event class.

(2) Designing the stereotype of Composite_Entity classes to describe the aggregation and combination relationships among atomic entities. Hierarchical composition mechanism is adopted to build Composite_Entity class, in order to describe multi-resolution and multi-hierarchical models. The tagged value named "composition" record the entity classes covered in the adjacent lower layer of the Composite_Entity, and which are the basic meta-data for model polymerization and depolymerization to support multi-resolution simulation.

(3) Separately calculating the union sets of Entity_Attribute, Event_Send_Port and Event_Receive_Port of all Atomic_Entities covered by Composite_Entities, Entity_Attribute, Event_Send_Port and Event_Receive_Port of the Composite_Entity are worked out.

(4) Designing the inheritance Relationship among entity classes. Sub-entity classes inherit all Entity_Attribute, Event_Send_Port and Event_Receive_Port from the parent entity class.

(5) Designing the data dependency relationship among entities. Data dependency relationship shall be determined in accordance with the data flow among entity classes and a Data_Interaction class is built among entity model classes where dependency exists. Data_Interaction class starts from dependency entity class and is directed to the depended entity class. The tagged value named "depending_data" of Data_Interaction records the dependent Entity_Attributes.

(6) Designing the Event interaction relationship among entities. An Event_Interaction class is built among entities associated by sending and receiving events, and the Pattern sequence diagram for each Event_Interaction is designed. Important to note that this Pattern sequence diagram is the sequence diagram of classes. Each Event_Message in the sequence diagram corresponds with a specific event class.

As an example, the output modeling product of model design stage is shown in Fig. 6. In this model, Atomic_Entity classes A, A', B and C are built. Entity class A' inherits from entity class A. Class A has data dependency on class B. Class B has event interaction dependency with class C, and the event interaction is described by extended sequence diagram Pattern1.
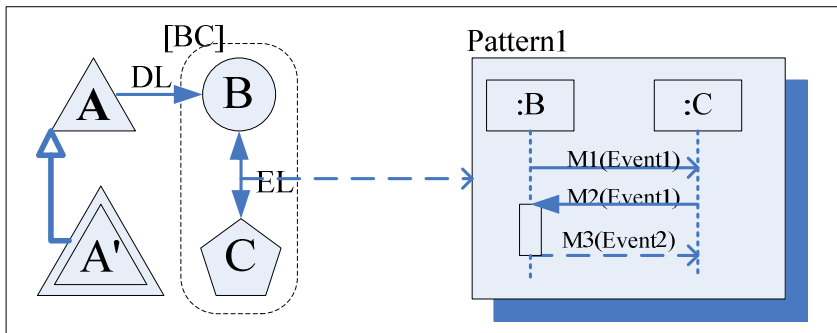
Fig. 6. Modeling product output during model design stage

4.2   Instantiation Stage

In accordance with simulation requirements, Atomic_Entity class, Composite_Entity class and Entity_Interaction are gained during instantiation period. If there are instances of Composite_Entity class in the system, hierarchical model instances tree shall also be built.

The output modeling product of instantiation stage is shown in Fig.7. By instantiation, instances of entity class A, a1 and a2, instance of entity class A', a1', instances of entity class B, b1 and b2, instances of entity class C, c1,c2 and c3, instances of composite entity class BC, bc1 and bc2, instances of Data_Interaction class between a1 and b2, dl1, instances of Data_Interaction between a1' and b1, dl2, instances of Event_Interaction between b2 and c1, el1, instance of Event_Interaction between b1 and c3, el2 are obtained. Model instances tree including bc1 and bc2 is built to record the hierarchical relationship.
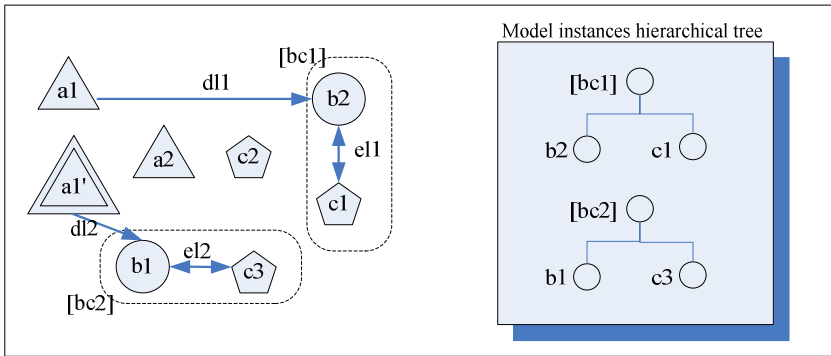
Fig. 7. Modeling products output during instantiation stage

4.3   Subsystem Determination Stage

The main work of Subsystem determination stage are analyzing collaborative simulation target, comprehensively considering the layout of simulation model and Simulation runtime environment, selecting collaborative entity models to participate in the current simulation tasks, determine subsystem function and delimit the components of subsystems. It can be divided into the following two steps:

(1) The function of each collaborative simulation subsystem is determined in accordance with simulation tasks. Select entity instances from all entity simulation instances for participating in the simulation and delimit the components of subsystems.

(2) Entity_Interaction types can be determined in accordance with the boundaries of subsystems. The following three situations are included.

✓ Association outside the subsystem —— If the entity classes connecting the association are outside of the entities set consisting of all subsystems, this connection is unconcerned with the current subsystem, and the current system doesn't have to deal with it. Set the "Interaction_Type" tagged value of the association instance as "outside", which means that it is an interaction outside the subsystem.

✓ Association among subsystems —— Association beyond the boundaries of subsystems. In other words, in entity instances connected to the association connection have both entities inside the entity set consisting of a specific subsystem and entities outside the entity set consisting of the subsystem. It is deemed as association among subsystems in collaborative simulation which is also a collaborative behavior where attention shall be paid during the process of collaborative simulation. Set the "Interaction_Type" tagged value of the association instance as "between" to mark it as interaction among subsystems.

✓ Association inside subsystems——If the entity instances connected to interactive association are all in the entity set consisting of a specific subsystem, this association is an association inside subsystem. These interactions need to be deal within the subsystem, but not interactions occurred between the simulation

subsystem. Set the "Interaction_Type" tagged value as "inside" and mark it as interaction inside subsystem.

In all entity instances connected to Entity_Interaction, there are entity instances which belong to the entity set consisting of a specific subsystem and entity instances which don't belong to the entity set consisting of all subsystems. This situation indicates that there are entities which participate in collaborative simulation but not be grouped into subsystems. CSMM meta-model stipulates that this situation is illegal. If it happens in the modeling process, the user will be noticed that there are some mistakes in the model in model validation stage. The user shall check and reset the components for subsystems.

It is Allowed to divide the composition of Composite_Entity instance into different subsystems. In this case, the enabled flag "Enabled" should be set to "false", means that the Composite_Entity can participate in the simulation only in the form of which composition sub-entity individual.

The grade of simulation resolution shall be determined in accordance with the target of collaborative simulation tasks and then set the "Enabled" flag of the Composite_Entity in accordance with the resolution grade. If low resolution is adopted in the simulation, set "Enabled" to "true" and then the composite entity is effective. If high resolution is adopted in the simulation, set "Enabled" to "false". It means that the composite entity is ineffective.

The output modeling product during subsystem determination stage is shown in Fig.8. The overall model is divided into two subsystems in accordance with collaborative simulation target. The subsystem boundary crosses the Composite_Entity bc2 and makes it ineffective. dl1 and el2 in Fig.8 are the associations among systems which requires to be mapped in collaborative simulation scenario interaction diagram. dl2 and el1 are both associations inside subsystems. And there is no association outside subsystems in Fig.8.
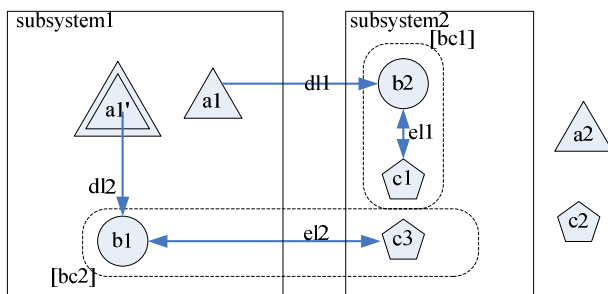
Fig. 8. Modeling products output during subsystem determination stage

4.4   Simulation Configuration Stage

The main work in simulation configuration stage includes designing the collaborative scenario among subsystems in accordance with collaborative simulation tasks and configuring simulation parameters.

(1) Designing Simulation_Scenario sequence diagrams correspond with the interactions among subsystems in accordance with simulation tasks. Simulation_Scenario here is the sequence diagram of the instances. And then mapping the associations among subsystems output during subsystem determination stage to the sequence diagram

- ✓ Mapping the Data_Interaction instances among subsystems to synchronous message "requestData" in the Simulation_Scenario sequence diagram. The data are acquired by the parameters of the return message.
- ✓ Mapping the Event_Interaction sequence diagram of instances to a sub-segment in the Simulation_Scenario sequence diagram.

(2) Set simulation parameters such as simulation step and simulation end condition etc.

Simulation_scenario sequence diagram acquired in simulation configuration stage is shown in Fig.9. There are two associations among subsystems. Data_Interaction dl1 is mapped to requestData message and its return message in simulation sequence diagram. The instance of pattern interaction diagram corresponding with Event_Interaction el2 is mapped to sub-segment of simulation scenario sequence diagram. According to the simulation task, the simulation is doing data request and event interaction between two subsystems circularly until the simulation end time is reached.
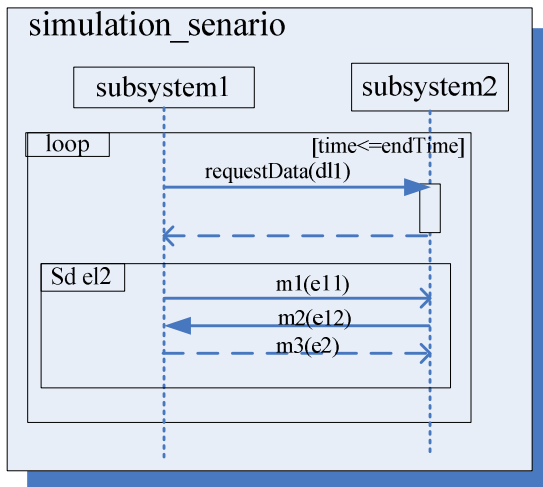
Fig. 9. Simulation scenario sequence diagram output in simulation configuration stage

5   Conclusions

In order to remove the obstacles in system integration and collaborative simulation, and promote information sharing among different disciplines models in the complex product development, this paper presents a collaborative simulation meta-model,

CSMM, based on UML profile technology by referring to object-oriented modeling methodology on the basis of extending the entity-relationship model. Also, it explains in detail the stereotypes, tagged values, limitations and the semantics of meta-models. Also, the process for modeling complex products collaborative simulation model based on CSMM meta-model is also introduced.

The CSMM describes static and dynamic collaborative semantics information such as data interaction and event collaboration among simulation models uniformly. Meanwhile it supports multi-granularity modeling and multi-resolution simulation in complex product development process.

References

1. Chai, X., Li, B., Xiong, G., et al.: Research and Implementation on Collabrative Simulation Platform for Complex Product. Computer Integrated Manufacturing Systems 8(7), 580–584 (2002)
2. Wang, K., Xiong, G.: Cooperative Design and Simulation for Complex Product. Computer Integrated Manufacturing Systems 9(S), 15–19 (2003)
3. OMG Staff Strategy Group. Model Driven Architecture, `http://www.omg.org/2000`
4. SISO-STD-003.0-DRAFT-V0.12,26 2005(8), Simulation Interoperability Standards Organization. Guide for Base Object Model (BOM) Use and Implementation
5. Brutzman, D., Tolk, A.: JSB Composability and Web Services Interoperability via Extensible Modeling & Simulation Framework (XMSF), Model Driven Architecture (MDA), Component Repositories, and Web-based Visualization (November 2003/December 28, 2011), `http://www.movesinstitute.org/xmsf.html`
6. European Space Agency. SMP 2.0 Metamodel. European Space Agency, France (October 2005)

Research and Application on Simulation Technology of Natural Environment

Jing Li, Shaoning Zhang, and Jingyan Han

Beijing Electromechanical Engineering Institute, 100074 Beijing, China
siriuslee2012@hotmail.com, zsnhorse376@yahoo.com.cn,
2668750134@qq.com

Abstract. The natural environment affects and restricts the performance of high tech vehicle and platform. Research shows that modeling and simulation technology of natural environment influence on the design, development and application of modern vehicle all round. The environmental effect on vehicle is analyzed, and the research scheme and the key technology breakthrough in the process of study are introduced. The simulation result indicates that this kind of research method has advanced in technology.

Keywords: natural environment simulation, vehicle, mesoscale weather numerical model, data assimilation.

1 Introduction

Complicated meteorological hydrological environment (including ocean, atmosphere and space environment) affects and restricts the operational effectiveness of high tech vehicle. After the Kosovo war, The U.S. research institution analyzes the results of precision strikes, and then it is found that twenty-five percent of task is forced to cancel because of weather. At the end of the 1990s, American troops by using numerical simulation method to study the influence of weather on a local conflict which involves five thousand targets, the result shows that they can save $310 million just in accurate ammunition spending by mastering the environmental effect. So the modeling and simulation technology of meteorological hydrological environment can influence on the propulsion of modern vehicle all round in the future.

2 The Natural Environment Effect Analysis on Vehicle

Natural environment elements in the report include atmospheric environment conventional weather and unconventional weather, marine hydrological environment and electromagnetic environment (lightning), etc. These elements effect on the flight performance, flight path, control and stability, homing performance and onboard equipments working performance of the vehicle.

Wind may endanger the control characteristics and security of vehicle. Wind field closely relates to aerodynamic lift and air Resistance, and the drastic change of wind

velocity field brings greater difficulty for flight control and makes vehicle up or down out of control. Otherwise, wind can restrict the design of the vehicle maximum-range.

Atmospheric temperature influences aeroengine thrust output and sound speed calculation. The change of temperature brings calculation error of actual flight Mach number. When simulation calculation can't reflect actual Mach number, the command of vehicle aerodynamic characteristics appears the deviation accordingly.

Atmospheric pressure influences the height control of vehicle. Along with regional difference, phenomena of atmosphere pressure surface drift appeared. The drift causes flying altitude measurement error, and influences the flight control further.

Unconventional weather such as cloud, icecap and thunderstorm is harmful for the flight-safe and calculation of various guidance way.

Lightning is a vehicle killer. The strong electromagnetic pulse which radiates form it can damage electronic instrument and cause abnormal data link communication.

Study on ocean environment such as flow field, wave is important safeguard to movement stability and gesture correctness for vehicle.

At this stage, the standard atmosphere mode is used in design and simulation of vehicle generally. For the long time flying vehicle which flies in atmospheric boundary layer, standard atmospheric hypothesis is insufficient to provide a powerful support for vehicle design and efficiency analysis because it can't factually reflect how natural environment influences on its performance and control precision.

3   The Research of Natural Environment Simulation

Firstly, the numerical value modes will be chosen to replay atmospheric environment in a certain region. Based on this, the model of atmospheric environment conventional weather and unconventional weather will be reached with the appropriate modeling method and observation data, the model of electromagnetic environment will be reached with the academic method and actual testing together, and the model of the ocean flow field will be reached with the actual the numerical value modes and actual testing together.

Through the research on the arithmetic of the transmission of the high altitude atmosphere infrared radiation, the model of infrared transmission and background radiation of the high/low altitude atmosphere can be reached using basic principles calculating the radiation of atmospheric background under the condition of the non-local thermodynamic equilibrium.

Then, the dynamic environment database will be built to save the environmental data model and to support data inquiry and access.

At last, based on SEDRIS, the research on the technology of environmental data representation and interchange, standard and specification will be developed and environmental simulation platform will be attained. This platform can be used in the vehicle simulation to test and upgrade its performance. In figure 1, the design frame of environmental simulation platform and the in-out relation of modules are described.
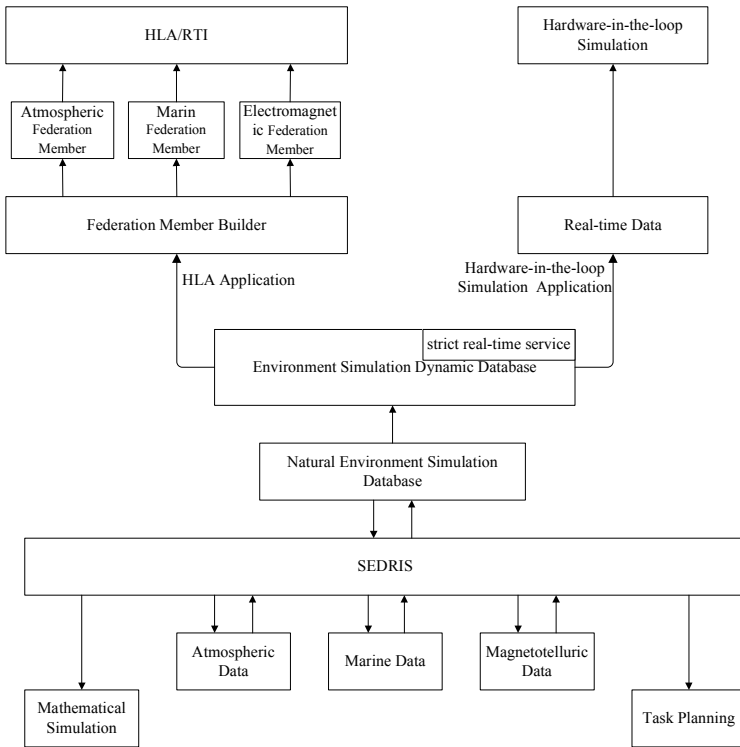
Fig. 1. The frame of environmental simulation platform

4   The Key Technology Research of Environmental Simulation Platform

The environmental simulation platform can realize functions such as M&S of atmospheric environment conventional weather and unconventional weather, M&S of marine hydrological environment and M&S of electromagnetic environment (lightning), standards and specifications of natural environment, the natural environment effect on vehicle, etc. The key technology of the platform is broken through, including the typical environmental simulation technology based on mesoscale weather numerical model, the data assimilation modeling technology of ocean tidal current and circulation and the performance effect simulation technology of natural environment on long time flying vehicle.

4.1   The Typical Environmental Simulation Based on Mesoscale Weather Numerical Model

Firstly, quantitative evaluation method of mesoscale weather numerical model is presented to choose a certain numerical model which is highly reliable and applies in

local region. Figure 2 is the tech flow chart of the quantitative evaluation method of mesoscale weather numerical model. By chosen statistic indicators such as average error, root mean square, correlation coefficient, results that various numerical models simulate weather elements, including temperature, wind field and geopotential height, are evaluated from three aspects of systematical error, average deviation and anomaly correlation.

As is explored above, calculation platform is developed on SGI Altix350 high performance server to comparing mesoscale weather numerical models. Applied together with observation data preprocessing, objective analysis and assimilation, mode initialization and diagnostic analysis of mode output, the typical weather process in local region can be replayed. The mesoscale numerical models such as MM5, RAMS WRF and AREM are usual models at home and abroad. Using above models, the numerical model evaluation software on the high performance server replays the atmospheric model of the temperature field, wind field and geopotential height field and scores statistically the data in specified region. At last the software analyzes the statistic results and evaluates numerical simulation ability of various models.

Through simulation validation, it is proved that the quantitative evaluation and reappearance technology is viable. It can replay conventional weather realistically and achieve data model of wind field, temperature field and pressure field per month.
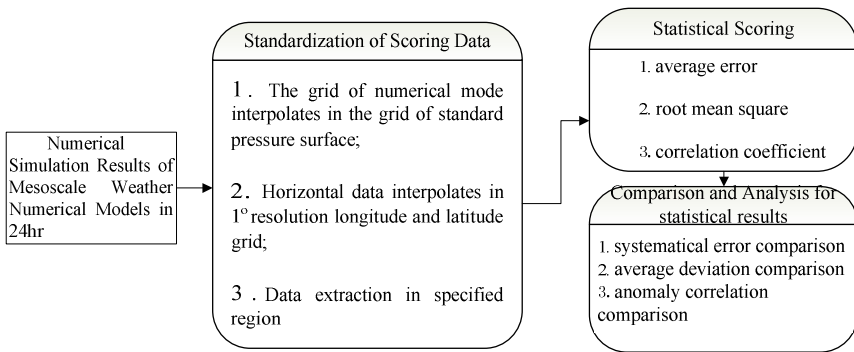
Fig. 2. The tech flow chart of the quantitative evaluation method of mesoscale weather numerical model

4.2 The Data Assimilation Modeling Technology of Ocean Tidal Current and Circulation

Data assimilation technology is important to reduce simulation error of ocean tidal current and circulation.

Nudging is a kind of utility method which assimilates satellite data into the numerical simulation data, but at the same time, it brings the dynamic process distortion. Variational method is more appropriate than Nudging, but some multi-grid ocean model can't use this method because that its number of iterations is overabundant and computational complexity is large.

So based on the ocean model (POM), a new method is applied to data assimilation technology of ocean tidal current and circulation. It is a little small calculation burden and high precision through adding a forced term (F_ς) in right end of continuous control equation of POM in σ coordinate Systems and setting up an adjustable parameter.

In order to validate that the simplified method is effective, through comparing the POM simulated results of the M2, S2, K1 and O1 tidal components based on assimilated data with the simulated results based on unassimilated data, it indicates that assimilated data deviation is smaller than unassimilated data deviation and simulation result based on assimilated data is closer to actual observations by compared simulation data with the charts from satellite data.

4.3   The Performance Effect Simulation of Natural Environment on Long Time Flying Vehicle

The performance effect simulation research of natural environment on long time flying vehicle is achieved by the development of environmental simulation platform based on HLA. It contains the natural environment simulation database, KD-RTI, the dynamic effect model of environment on equipment, 2-dimensional display module and 3-dimensional display module, and the platform is used to analyze vehicle performance considering natural environment influence.

The platform development process is as follows. The environment simulation database is base of platform running, and it manages air-sea environment data in local region. According to simulation scenario which includes area, time, precision and environment information sets, the database creates run-time database and uses main memory database to ensure real-time performance. Then, federation member of a certain long time flying vehicle based on its guidance math model is developed. Finally, on condition that natural environment affects whole range of the vehicle, simulation experiment makes use of air-sea environment data booking based on HLA-RTI and carries out.

The simulation experiment makes environment platform as technical support, and it not only validates some key functions, such as simulation modeling of air-sea environment, vehicle dynamic response model to environment of vehicle federation member and real-time dynamic data service of environment simulation database and so on, but also assesses the guidance and control performance.

5   Summary

Generally, the standard atmosphere model is used to the research of vehicle. The effect of constant wind is considered at most. But, with the improvement of requirements for flight time and flight distance, the environment model, such as standard atmosphere and constant wind, is not enough to scientific research. So the study method of natural environment simulation in this article in combination with theory analysis and observation statistics analysis and environment numerical model

simulation is provided. The research result can comprehensively evaluate the guidance and control system performance of vehicle flying in atmospheric layer under the condition close to actual natural environment. The research method has advanced and innovative in technology.

References

1. Perez, J.A., Hughes, J.G.: The DOD Modeling and Simulation Integrated Natural Environment Program. In: Proceedings of the 2000 Spring SIW Conference, 00S-SIW-073 (March 2000)
2. Richbourg, R., et al.: The Environmental Data Coding Specification: The Standard for Identification of Environmental Objects and Properties. In: Proceedings of the 2002 Spring SIW Conference, 02S-SIW-049 (March 2002)
3. Carswell, J.: The SEDRIS Transmittal Format (STF). In: Proceedings of the 1999 Spring SIW Conference, 99S-SIW-058 (March 1999)
4. Pokhariyal, D., et al.: Aircraft flight dynamics with simulated ice accretion. In: AIAA-20001-0541 (2001)
5. Xu, L., et al.: Research on Modeling Methods of Atmospheric Environment Simulation. Journal of System Simulation 18(suppl. 2) (August 2006)

Research on Multi-spectral Infrared Translation Technique Based on Fiber Array Plane

Cheng Tang[1], Zhuo Li[2], Wenhua Kong[1], and Lili Zhou[1]

[1] Beijing Electro-mechanical Engineering Institute, Beijing 100074, China,
[2] Beijing Institute of Technology, Beijing 100081, China
tchvsbxm@126.com, lizhou@Bit.edu.cn, kongwenhua@263.net,
ZLL197949@hotmail.com

Abstract. To satisfy the infrared detector simulation need for different wave bands, the research on multi-spectral infrared translation technique based on "fiber array plane", was developed. It achieved that multi-spectral infrared scene simulator's manufacturing, and the system contains video to visible transducer, visible to infrared transducer and multi-spectral optical system.

Keywords: Fiber array plane, multi-spectral infrared scene simulator, multi-spectral infrared optical system.

1   Introduction

To satisfy the simulation need of medium wave infrared and long wave infrared for infrared imaging detector, when developing infrared scene simulation equipments, it often developed two different infrared scene simulators. This way is unscientific in both utilize rate and economy benefit. Aim at the problem above, this article puts forward a multi-spectral infrared image transition technique based on "fiber array plane" which can achieve two wave bands (medium wave, long wave) infrared scene simulator's developing to satisfy the simulation need of medium wave and long wave detectors.

2   Multi-spectral Infrared Image Conversion Technique Based on "Fiber Array Plane"

Fiber array plane is composed of two parts: pedestal and fiber array, its structure is shown in Fig. 1. Fiber array is sealed in a vacuum device. When the visible light shines on one ending- face, the other absorbs the incident visible light which plated with absorb membrane. The membrane converses light energy to heat energy, and radiates infrared light, further generate infrared temperature distributing image. The radiation intention of infrared light and the intention of visible light have linear relation. The radiating wavelength of peak value satisfies Wien displacement law ($\lambda T=2897.8\pm0.04\mu Mk$)). It evolutes from membrane infrared image transition technique.
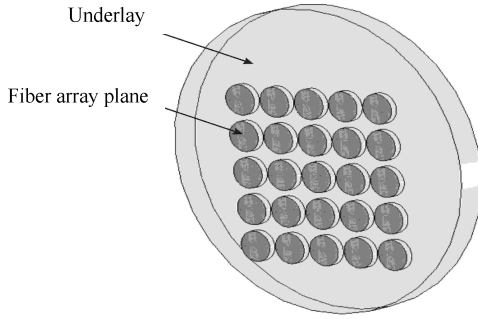
Fig. 1. Fiber array plane composing structure picture

3　Multi-spectral Infrared Scene Simulator Based on Fiber Array Plane

3.1　System Composing and Working Course

Multi-spectral infrared scene simulator based on fiber array plane was composed of three parts, which are video to visible image transducer, visible to infrared image transducer and multi-spectral optical projection system, just as shown in Fig.2. It works as follows: Computer Image Generation System (CIG) generates video signal which has infrared aim and its background character. The video signal is translated to visible image by video to visible image transducer. The visible image is casted to incidence window of visible to infrared image transducer, and is conversed to generate a corresponding wave band infrared radiating image. The image is projected to imaging detector by multi-spectral optical projection system.
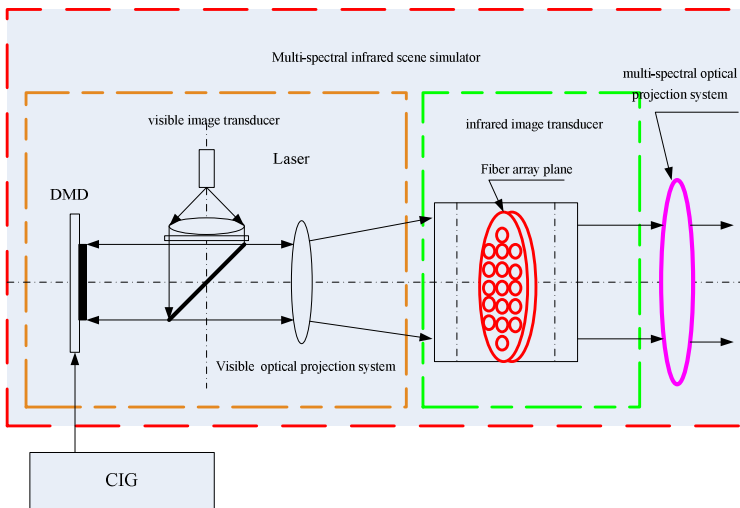
Fig. 2. Multi-spectral infrared scene simulators composing picture based on Fiber array plane

3.2   The Function and Developed for Key Assembly

3.2.1   Video to Visible Image Transducer

Video to visible image transducer is used to change video signal with infrared character that generated by computer image generation system, to visible image signal. This transducer is composed of microdisplay device、 backlighting、 uniformity of luminous intensity system and visible optical projection system. And hereinto, DMD
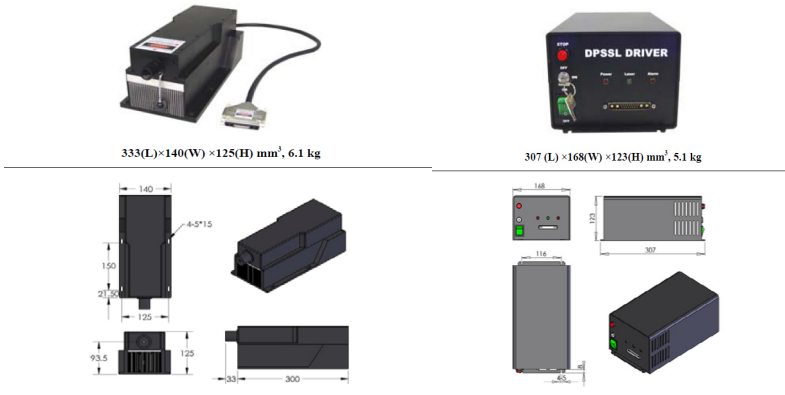
Fig. 3. The actual picture of green laser and laser power supply
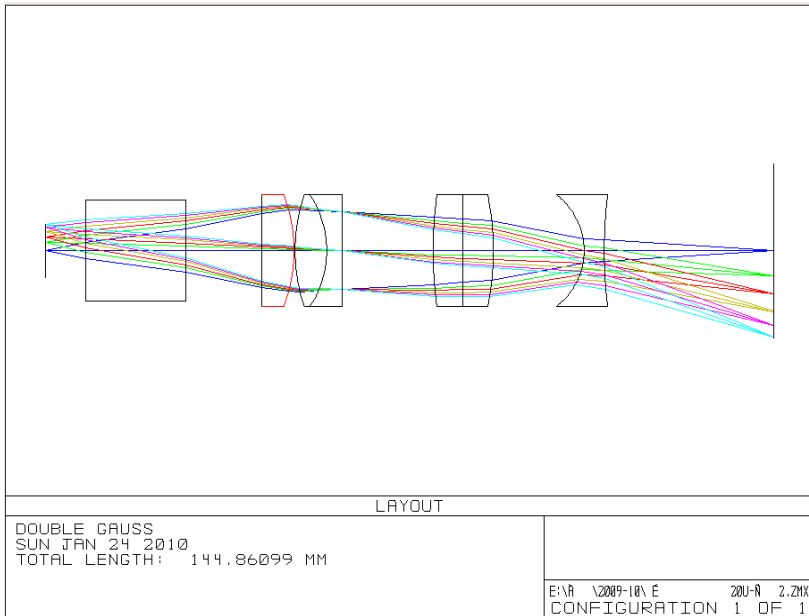
Fig. 4. System design of visible optical projection

and LCOS which was familiar to you at present market can be used for microdisplay device; the green laser can be used for backlighting, and its structure is shown in Fig.3; the laser collimation lens, uniformity board and   beam-splitter are adopted to actualize the uniformity of luminous intensity; The system of visible optical projection has two functions, one is projecting image in the microdisplay device onto the incidence surface (fiber array plane) of infrared image transducer; another is transforming visible optical pixel dimension into the pixel dimension which is requested by infrared image. It's finite distance imaging and high resolution with projection lens, the system design of visible optical projection was illustrated with Fig.4.

3.2.2   Visible to Infrared Image Transducer

Visible to infrared image transducer is the hard core of this system, which used to change visible image to infrared image that can be identified by detector. This transducer is composed of visible window, infrared window, fiber array plane, cooling system, and vacuumize system. And hereinto, the process of fiber array plane design consists of three steps. Firstly, establish the physics model of fiber array plane based on theory analyze calculate; secondly, analyze and design the model by finite element thermal analysis; finally, confirm the target parameter through of repeated experiment validation. The principium templet of fiber array plane is shown in Fig.5. The cooling system and vacuumize system made sure that fiber array plane worked at the temperature below room temperature, and generated a invariableness underlay temperature which can reduce the heat change between fiber array and outer, and avoid the heat interfere among pixels, confirm the quality of created image. Cooling system adopted large power of semiconductor refrigeration technology, and the materials of Ge that can be transmissive medium/long wave, was used for the infrared window materials of the visible to infrared image transducer.
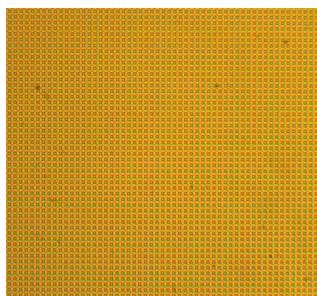
Fig. 5. Principium templet of fiber array plane

3.2.3   Multi-spectral Infrared Optical Projection System

Most of the projection systems for infrared scene simulator are transmissive lens, and one or two transmissive system can be adopted, considered of the work need of medium wave and long wave. On condition that the same transmissive system, its difficulty to reduce the color aberration. While, the infrared scene simulator adopted two system for multi-spectral, it will enhance the manufacture cost, also, the two

systems are discommodiously to use. Therefore, the reflecttion optical system is introduced, medium wave and long wave share the seam optical system. So the design of this reflection optical system possesses merit of large aperture、 zero color aberration and   simpleness system.

4   Image Maked by Multi-spectral Infrared Scene Simulator

The multi-spectral infrared scene simulator was tested with a special infrared image tesing device, the result was shown in Fig.6.
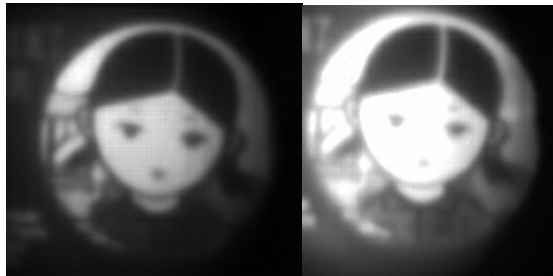
Fig.6-a   Fig.6-b

Fig. 6. Image maked by the principle sample of multi-spectral infrared scene simulator(Fig.6-a medium wave image Fig.6-b long wave image)

5   Conclusions

The principle of infrared imaging with fiber array plane was introduced, and the configuration of multi-spectral infrared scene simulator system was illustrated in this paper. On the basis of selecting fiber array plane as key assembly of the system, the multi-spectral infrared scene simulator was successfully developed, and the principle sample of multi-spectral infrared scene simulator was validated with testing. It can make medium/long wave infrared target image.

References

1. Zeng, Z.-Y.: Analysis of infrared optical system influence in infrared image simulation. Laser & Infrared 10 (2009)
2. Hui, B.: A survey for infrared space optical systems. Optical Instruments 04 (2007)
3. Chang, H.: Optical system design of DMD-based dual-band infrared scene projection. Journal of Harbin Institute of Technology 05 (2007)
4. Sun, S.-J.:Key Technology Study of Simulating Infrared Images 03 (2004)
5. Forber, R.A., Au, A., Efron, U., et al.: Dynamic IR scene projection using the Hughes liquid crystal light valve. In: SPIE, vol. 1665, pp. 259–273 (1992)
6. Philippe, B., Frederic, M.: Visible-to-IR-transducer-based infrared image projection systems. In: SPIE, vol. 4027, pp. 368–376 (2000)

A Compact Difference Scheme for Time Fractional Diffusion Equation with Neumann Boundary Conditions

Jianfei Huang[1,*], Yifa Tang[1], Wenjia Wang[2], and Jiye Yang[1]

[1] LSEC, ICMSEC, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, China
`jfhuang@lsec.cc.ac.cn`
[2] Yuanpei College, Peking University, Beijing 100871, China

Abstract. This paper is devoted to the numerical treatment of time fractional diffusion equation with Neumann boundary conditions. A compact difference scheme is derived for solving this problem, by combining the classic finite difference method for Caputo derivative in time, the second order central difference method in space and the compact difference treatment for Neumann boundary conditions. The solvability, stability and convergence of this scheme are rigorously discussed. We prove that the convergence order of this proposed scheme is $O(\tau^{2-\alpha} + h^2)$, where τ, α and h are the time step size, the index of fractional derivative and space step size respectively. Numerical experiments are carried out to demonstrate the theoretical analysis.

Keywords: Fractional diffusion equation, Neumann boundary conditions, Compact difference scheme, Stability, Convergence.

1 Introduction

The subject of fractional calculus has gained considerable popularity and importance due to its attractive applications as a new modeling tool in a variety of scientific and engineering fields, such as viscoelasticity [5], hydrology [1,2,11], finance [3,12,14], and system control [13]. Those fractional models, described in the form of fractional differential equations, tend to be more appropriate for the description of memorial and hereditary properties of various materials and processes than the traditional integer-order models [15]. In particular, modeling of anomalous diffusion in a specific type of porous mediums is one of the most significant applications of fractional derivatives [1,2].

Roughly speaking, a time fractional diffusion equation is obtained from the classical diffusion equation by replacing the first-order time derivative by a fractional derivative of order $0 < \alpha < 1$ [10]. It is derived by considering continuous time random walk problems, which are in general non-Markovian processes [4]. In the last decade, a number of numerical methods have been developed to solve

---

[*] Corresponding author.

the time fractional diffusion equation with Dirichlet boundary conditions. Liu et al. [9] use a first-order finite difference scheme in both time and space directions for this equation, where some stability conditions are derived. Yuste [18] presents a difference scheme based on the weighted average methods for ordinary (non-fractional) diffusion equations. Similar problems are considered by Lin, Li and Xu [7,8] using spectral method. In [8] they design a finite difference/spectrral method based on a finite difference scheme in time and Legendre spectral methods in space. And stability and convergence of this method are rigorously established. In [7] they propose a spectral method in both temporal and spatial discretizations.

Compared with the considerable numerical methods for the time fractional diffusion equation with Dirichlet boundary conditions, only a little work [6,20] has been done on the numerical method for the equation with Neumann boundary conditions. In [6], Langlands and Henry design an implicit numerical scheme for solving this problem. The accuracy and stability of this scheme are investigated, even though there is no proof of global accuracy of the numerical discretization. Zhao and Sun [20] present a box-type scheme by combining order reduction approach and L1 discretization. The stability and convergence is rigorously proved with the accuracy $O(\tau^{2-\alpha}+h^2)$. However, from their theoretical analysis, we know that this box-type scheme can only be applied to solve the equation with homogeneous Neumann boundary conditions. Herein, we will construct a compact difference scheme for solving the time fractional diffusion equation with Neumann boundary conditions and make the corresponding theoretical analysis.

The rest of this paper is organized as follows. In Section 2 the compact difference scheme is constructed by combining the classic finite difference for Caputo derivative in time, the second order central difference in space and the compact difference treatment for Neumann boundary conditions. Section 3 is devoted to study the solvability, stability and convergence of this proposed scheme. Numerical experiments are carried out in Section 4, which verify the effectiveness of our scheme and support the theoretical analysis. Final section is the concluding remarks.

2   Construction of the Compact Difference Scheme

Consider the following time fractional diffusion equation

$$\mathbf{{}_0^C D_t^\alpha} u(x,t) = \frac{\partial^2 u(x,t)}{\partial x^2} + f(x,t), \ x \in (0,L), \ t \in (0,T], \ 0 < \alpha < 1 \qquad (1)$$

with initial condition

$$u(x,0) = \phi(x), \ x \in (0,L) \qquad (2)$$

and Neumann boundary conditions

$$u_x(0,t) = \beta_1(t), \ u_x(L,t) = \beta_2(t), \ t \in (0,T], \qquad (3)$$

where $^{C}_{0}\mathbf{D}^{\alpha}_{\mathbf{t}}u(x,t)$ is the Caputo derivative of order α with respect to time variable t, defined as

$$^{C}_{0}\mathbf{D}^{\alpha}_{\mathbf{t}}u(x,t) = \frac{1}{\Gamma(1-\alpha)} \int_{0}^{t} (t-\tau)^{-\alpha} u_{\tau}(x,\tau)d\tau.$$

We often refer α as the anomalous diffusion exponent, and notice that $\alpha = 1$ is the classical diffusion equation. The case of $0 < \alpha < 1$ models a slow-diffusive flow in which a cloud of diffusing particles spreads at a slower rate than the classical diffusion model predicts.

For the finite difference approximation, denote $h = L/M$, $\tau = T/N$, grid nodes $x_j = jh$ $(0 \le j \le M)$ and $t_n = n\tau$ $(0 \le n \le N)$, here M and N are two positive integers. U^n_j and $(U_{xx})^n_j$ are used to represent the approximations of $u^n_j = u(x_j,t_n)$ and $(u_{xx})^n_j = u_{xx}(x_j,t_n)$, respectively, and $f^n_j = f(x_j,t_n)$.

A classic $2-\alpha$-order finite difference approximation [8,16] of Caputo derivative is described as

$$^{C}_{0}\mathbf{D}^{\alpha}_{\mathbf{t}}u(x,t_{n+1}) = \frac{1}{\Gamma(1-\alpha)} \sum_{k=0}^{n} \int_{t_k}^{t_{k+1}} (t_{n+1}-\tau)^{-\alpha} u_{\tau}(x,\tau)d\tau$$

$$= \frac{1}{\Gamma(1-\alpha)} \sum_{k=0}^{n} \frac{u(x,t_{k+1})-u(x,t_k)}{\tau} \int_{t_k}^{t_{k+1}} (t_{n+1}-\tau)^{-\alpha}d\tau + r^{n+1}(x)$$

$$= \frac{1}{\Gamma(2-\alpha)} \sum_{k=0}^{n} \frac{u(x,t_{n+1-k})-u(x,t_{n-k})}{\tau^{\alpha}} b_k + r^{n+1}(x),$$

where $b_k = (k+1)^{1-\alpha} - k^{1-\alpha}$, and

$$|r^{n+1}(x)| \le c_u \tau^{2-\alpha},$$

where c_u is a constant depending only on u. In addition, we find that

$$b_k = (1-\alpha) \int_{k}^{k+1} s^{-\alpha}ds \qquad (4)$$

thus

$$b_k - b_{k+1} > 0.$$

Define the discrete Caputo fractional differential operator D^{α}_t by

$$D^{\alpha}_t u(x,t_{n+1}) = \frac{1}{\Gamma(2-\alpha)} \sum_{k=0}^{n} \frac{u(x,t_{n+1-k})-u(x,t_{n-k})}{\tau^{\alpha}} b_k.$$

Thus we have

$$^{C}_{0}\mathbf{D}^{\alpha}_{\mathbf{t}}u(x,t_{n+1}) = D^{\alpha}_t u(x,t_{n+1}) + r^{n+1}(x). \qquad (5)$$

As for the difference approximation of $u_{xx}(x_j,t)$, there is a well known second order central difference formula

$$\frac{\partial^2 u(x_j,t)}{\partial x^2} = \frac{u(x_{j+1},t) - 2u(x_j,t) + u(x_{j-1},t)}{h^2} - \frac{h^2}{12} \frac{\partial^4 u(x_j,t)}{\partial x^4} + o(h^2). \qquad (6)$$

For analyzing the truncation error of our compact difference scheme, we should present following lemma.

Lemma 1. *Assume $p(x) \in C^4[0, L]$, then following estimates hold*

$$11p''(x_1) - 2p''(x_2) = -6\frac{p'(0)}{h} + 6\frac{p(x_2) - p(x_1)}{h^2} - \frac{9h^2}{4}p^{(4)}(x_1) + o(h^2) \quad (7)$$

and

$$11p''(x_{M-1}) - 2p''(x_{M-2}) = 6\frac{p'(L)}{h} - 6\frac{p(x_{M-1}) - p(x_{M-2})}{h^2} - \frac{9h^2}{4}p^{(4)}(x_{M-1}) + o(h^2). \quad (8)$$

Proof. It can be directly obtained by some transformations from Lemma 2 in [17,19]. Specifically, we only need to carry out Taylor expansions at the points x_1 for (7) and x_{M-1} for (8) respectively.   □

Now we can use formulas (5)-(8) to discretize (1)-(3), i.e.,

$$D_t^\alpha u_j^{n+1} = (u_{xx})_j^{n+1} + f_j^{n+1} - r^{n+1}(x_j), \quad 2 \le j \le M - 2, \quad (9)$$

$$(u_{xx})_j^{n+1} = \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} - \frac{h^2}{12}\frac{\partial^4 u(x_j, t)}{\partial x^4} + o(h^2), \quad 2 \le j \le M-2, \quad (10)$$

$$11(u_{xx})_1^{n+1} - 2(u_{xx})_2^{n+1} = -6\frac{\beta_1(t_{n+1})}{h} + 6\frac{u_2^{n+1} - u_1^{n+1}}{h^2} - \frac{9h^2}{4}p^{(4)}(x_1) + o(h^2), \quad (11)$$

$$11(u_{xx})_{M-1}^{n+1} - 2(u_{xx})_{M-2}^{n+1} = 6\frac{\beta_2(t_{n+1})}{h} - 6\frac{u_{M-1}^{n+1} - u_{M-2}^{n+1}}{h^2} - \frac{9h^2}{4}p^{(4)}(x_{M-1}) + o(h^2). \quad (12)$$

According to (9)-(12) and omitting the truncation errors, it can be deduced our compact difference scheme for (1)-(3) as follows

$$D_t^\alpha U_j^{n+1} = (U_{xx})_j^{n+1} + f_j^{n+1}, \quad 2 \le j \le M - 2, \quad (13)$$

$$(U_{xx})_j^{n+1} = \frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{h^2}, \quad 2 \le j \le M - 2, \quad (14)$$

$$11(U_{xx})_1^{n+1} - 2(U_{xx})_2^{n+1} = -6\frac{\beta_1(t_{n+1})}{h} + 6\frac{U_2^{n+1} - U_1^{n+1}}{h^2}, \quad (15)$$

$$11(U_{xx})_{M-1}^{n+1} - 2(U_{xx})_{M-2}^{n+1} = 6\frac{\beta_2(t_{n+1})}{h} - 6\frac{U_{M-1}^{n+1} - U_{M-2}^{n+1}}{h^2}, \quad (16)$$

or its equivalent form

$$D_t^\alpha U_j^{n+1} = \frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{h^2} + f_j^{n+1}, \quad 2 \le j \le M - 2,$$

$$11D_t^\alpha U_1^{n+1} - 2D_t^\alpha U_2^{n+1} - 6\frac{U_2^{n+1} - U_1^{n+1}}{h^2} = -6\frac{\beta_1(t_{n+1})}{h} + 11f_1^{n+1} - 2f_2^{n+1},$$

$$11D_t^\alpha U_{M-1}^{n+1} - 2D_t^\alpha U_{M-2}^{n+1} + 6\frac{U_{M-1}^{n+1} - U_{M-2}^{n+1}}{h^2} = 6\frac{\beta_2(t_{n+1})}{h} + 11f_{M-1}^{n+1} - 2f_{M-2}^{n+1}.$$

It's clear that our scheme is consistent with the original fractional diffusion equation. For convenience of computation and analysis, we will focus on scheme (13)-(16).

3   Stability and Convergence of the Compact Difference Scheme

In this section, we will prove the solvability, stability, and convergence of the proposed method by using matrix analysis and the quadratic forms of analysis. First introduce the following vectors at any time level $t^n (n \geq 0)$

$$\overrightarrow{U}^n = (U_1^n, U_2^n, \cdots, U_{M-1}^n)^T, \quad (\overrightarrow{U}_{xx})^n = \left((U_{xx})_1^n, (U_{xx})_2^n, \cdots, (U_{xx})_{M-1}^n\right)^T,$$

$$\overrightarrow{\beta}(t_n) = \left(-6\frac{\beta_1(t_n)}{h}, 0, \cdots, 0, 6\frac{\beta_2(t_n)}{h}\right), \quad \overrightarrow{F}^n = \left(f_1^n, f_2^n, \cdots, f_{M-1}^n\right)^T.$$

Observing the coefficients of terms $(U_{xx})_j^{n+1}$ and U_j^{n+1} in (14)-(16), we can write them into following $(M-1) \times (M-1)$ square matrix respectively

$$A = \begin{pmatrix} 11 & -2 & & & \\ 1 & & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & -2 & 11 \end{pmatrix}, \quad B = \begin{pmatrix} 6 & -6 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -6 & 6 \end{pmatrix}. \tag{17}$$

With these notations, the schemes (14)-(16) can be combined and expressed in the following matrix form

$$A(\overrightarrow{U}_{xx})^{n+1} = -\frac{1}{h^2}B\overrightarrow{U}^{n+1} + \overrightarrow{\beta}^{n+1}. \tag{18}$$

Since A is diagonally dominant matrix, thus A is invertible. (18) can be reduced to

$$(\overrightarrow{U}_{xx})^{n+1} = -\frac{1}{h^2}A^{-1}B\overrightarrow{U}^{n+1} + A^{-1}\overrightarrow{\beta}^{n+1}. \tag{19}$$

Assume that values $\{U_j^k\}$ $(k = 0, 1, \cdots, n, \ j = 1, 2, \cdots, M-1)$ have been obtained, then let's show that values $\{U_j^{n+1}\}$ can be obtained from $\{U_j^k\}$. Substituting (19) into (13), we have the following vector equation

$$D_t^\alpha \overrightarrow{U}^{n+1} = -\frac{1}{h^2}A^{-1}B\overrightarrow{U}^{n+1} + A^{-1}\overrightarrow{\beta}^{n+1} + \overrightarrow{F}^{n+1}, \tag{20}$$

or more concrete form

$$(A + \gamma B)\overrightarrow{U}^{n+1} = \sum_{k=1}^{n}(b_{n-k} - b_{n-k+1})A\overrightarrow{U}^k + b_n A\overrightarrow{U}^0$$

$$+ \Gamma(2 - \alpha)\tau^\alpha \overrightarrow{\beta}^{n+1} + \Gamma(2 - \alpha)\tau^\alpha A\overrightarrow{F}^{n+1}, \tag{21}$$

where $\gamma = \Gamma(2-\alpha)\tau^\alpha/h^2$. It's clear that the matrix $A+\gamma B$ is invertible, thus the system of equations (21) is solvable, i.e., our compact scheme (13)-(16) admits a unique solution.

Before carrying out the analysis of stability and convergence, two lemmas should be provided.

Lemma 2. *Assume that λ is an eigenvalue of the matrix $A^{-1}B$, and $\overrightarrow{X}$ is its corresponding eigenvector. Then the real part of λ is nonnegative.*

Proof. It can be checked that the Greschgorin disks of A are within the right half of the complex plane, thus eigenvalues of the matrix A and A^{-1} have nonnegative real-parts. Let's turn to consider B. Since B can be decomposed into the product of following two matrices C and D

$$B = CD = \begin{pmatrix} 6 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 6 \end{pmatrix} \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix},$$

thus the eigenvalues of C are clear, and notice that D is a real symmetric matrix and the Greschgorin disks of D are also within the right half of the complex plane, we can know the eigenvalues of B is real and nonnegative. This completes the proof. $\qquad\square$

Lemma 3. *Let A be an arbitrary square matrix. Then for any matrix norm, we have $\rho(A) \leq \|A\|$, where $\rho(A)$ represents the spectral radius of A. Moreover, for any $\epsilon > 0$, then there exists a matrix norm, denoted by $\|\cdot\|_\epsilon$, such that $\|A\|_\epsilon \leq \rho(A) + \epsilon$.*

Now we consider the stability and the convergence. The results and proofs are based on (21).

Theorem 1. *For any $0 < \epsilon \leq \tau$, the compact scheme (13)-(16) or its equivalent vector form (21) is unconditionally stable with respect to the initial conditions.*

Proof. Assume that $\overrightarrow{U}_1^m$ and $\overrightarrow{U}_2^m$ are two different solutions of (21) having two different initial values but same boundary conditions. Let $\overrightarrow{\theta}^m = \overrightarrow{U}_1^m - \overrightarrow{U}_2^m$, then it satisfies the following equation

$$\overrightarrow{\theta}^{n+1} = \sum_{k=1}^{n} (b_{n-k} - b_{n-k+1})\overrightarrow{\theta}^k + b_n \overrightarrow{\theta}^0 - \gamma A^{-1}B\overrightarrow{\theta}^{n+1},$$

or

$$\overrightarrow{\theta}^{n+1} = (I + \gamma A^{-1}B)^{-1} \sum_{k=1}^{n} (b_{n-k} - b_{n-k+1})\overrightarrow{\theta}^k + (I + \gamma A^{-1}B)^{-1} b_n \overrightarrow{\theta}^0. \quad (22)$$

According to Lemmas 2 and 3, for any $0 < \epsilon \leq \tau$, it holds that

$$\|(I + \gamma A^{-1}B)^{-1}\|_\epsilon \leq \rho((I + \gamma A^{-1}B)^{-1}) + \varepsilon \leq 1 + \epsilon. \tag{23}$$

For the first time step, that is, $n = 0$, the scheme (22) simply reads

$$\|\overrightarrow{\theta}^1\|_\epsilon \leq (1 + \epsilon)\|\overrightarrow{\theta}^0\|_\epsilon.$$

Assume we have proven

$$\|\overrightarrow{\theta}^k\|_\epsilon \leq (1 + \epsilon)^k\|\overrightarrow{\theta}^0\|_\epsilon, \ k = 1, 2, \cdots, n, \tag{24}$$

we will prove $\|\overrightarrow{\theta}^{n+1}\|_\epsilon \leq (1 + \epsilon)^{n+1}\|\overrightarrow{\theta}^0\|_\epsilon$. From (22), (23) and (24), it can be deduced that

$$\|\overrightarrow{\theta}^{n+1}\|_\epsilon \leq (1 + \epsilon) \sum_{k=1}^{n}(b_{n-k} - b_{n-k+1})\|\overrightarrow{\theta}^k\|_\epsilon + (1 + \epsilon)b_n\|\overrightarrow{\theta}^0\|_\epsilon$$

$$\leq (1 + \epsilon)^{n+1} \left(\sum_{k=1}^{n}(b_{n-k} - b_{n-k+1}) + b_n \right) \|\overrightarrow{\theta}^0\|_\epsilon$$

$$\leq (1 + \epsilon)^{n+1}\|\overrightarrow{\theta}^0\|_\epsilon.$$

Thus

$$\|\overrightarrow{\theta}^{n+1}\|_\epsilon \leq e^{(n+1)\epsilon}\|\overrightarrow{\theta}^0\|_\epsilon \leq e^T\|\overrightarrow{\theta}^0\|_\epsilon.$$

This means our scheme is unconditionally stable with respect to the initial conditions. □

Denote $e_j^{n+1} = u_j^{n+1} - U_j^{n+1}$, $\overrightarrow{e}^{n+1} = \left(e_1^{n+1}, e_2^{n+1}, \cdots, e_{M-1}^{n+1}\right)^T$. Now we aim at deriving the estimates for $\|\overrightarrow{e}^{n+1}\|_\epsilon$ being given in the following theorem.

Theorem 2. *Suppose that u_j^n is the exact solution of (9)-(12) at grid point (x_j, t_n), U_j^n is the compact difference solution of (13)-(16) or its equivalent vector form (21). Then for any $0 < \epsilon \leq \tau$, there exists positive constant M, such that*

$$\|\overrightarrow{e}^{n+1}\|_\epsilon \leq M(\tau^{2-\alpha} + h^2), \ n = 0, 1, \cdots, N - 1. \tag{25}$$

Proof. Subtract (13)-(16) from (9)-(12) and transform their difference into matrix form, since $\overrightarrow{e}^0 = 0$, then there exists a constant c such that

$$\|\overrightarrow{e}^{n+1}\|_\epsilon \leq (1 + \epsilon) \sum_{k=1}^{n}(b_{n-k} - b_{n-k+1})\|\overrightarrow{e}^k\|_\epsilon + c(1 + \epsilon)\tau^\alpha(\tau^{2-\alpha} + h^2). \tag{26}$$

First we will prove $\|\overrightarrow{e}^{n+1}\|_\epsilon \leq cb_n^{-1}(1 + \epsilon)^n(\tau^{2-\alpha} + h^2), \ n = 0, 1, \cdots, N - 1$ by induction.

When $n = 0$, we have

$$\|\overrightarrow{e}^1\|_\epsilon \le c(1+\epsilon)\tau^\alpha(\tau^{2-\alpha} + h^2) = cb_0^{-1}(1+\epsilon)\tau^\alpha(\tau^{2-\alpha} + h^2). \qquad (27)$$

Suppose that we have obtained

$$\|\overrightarrow{e}^k\|_\epsilon \le cb_{k-1}^{-1}(1+\epsilon)^k(\tau^{2-\alpha} + h^2), \ k = 1, \cdots, n, \qquad (28)$$

we want to prove $\|\overrightarrow{e}^{n+1}\|_\epsilon \le cb_n^{-1}(1+\epsilon)^n\tau^\alpha(\tau^{2-\alpha} + h^2)$.
Substituting (28) into (26) gives

$$\|\overrightarrow{e}^{n+1}\|_\epsilon \le (1+\epsilon)\sum_{k=1}^n (b_{n-k} - b_{n-k+1})cb_{k-1}^{-1}(1+\epsilon)^k\tau^\alpha(\tau^{2-\alpha} + h^2) + c(1+\epsilon)\tau^\alpha(\tau^{2-\alpha} + h^2)$$

$$\le cb_n^{-1}(1+\epsilon)^n\tau^\alpha(\tau^{2-\alpha} + h^2)\left(\sum_{k=1}^n(b_{n-k} - b_{n-k+1}) + b_n\right)$$

$$\le cb_n^{-1}(1+\epsilon)^n\tau^\alpha(\tau^{2-\alpha} + h^2).$$

Since for $0 < \epsilon \le \tau$

$$(1+\epsilon)^{n+1} \le e^T,$$

and from (4)

$$\lim_{N\to\infty} b_N^{-1}\tau^\alpha = T^\alpha \lim_{N\to\infty}\frac{b_N^{-1}}{N^\alpha} = \lim_{N\to\infty}\frac{T^\alpha}{(1-\alpha)\int_N^{N+1}(s/N)^{-\alpha}ds} = \frac{T^\alpha}{1-\alpha},$$

thus we can conclude the result

$$\|\overrightarrow{e}^{n+1}\|_\epsilon \le M(\tau^{2-\alpha} + h^2), \ n = 0, 1, \cdots, N-1,$$

where M is a positive constant depending on T and α. $\qquad\square$

4   Numerical Experiments

In this section, we will verify the theoretical results by doing numerical experiments. Two numerical examples are provided below, the first one is a problem with nonhomogeneous Neumann boundary conditions, the second is a problem with homogeneous case.

Example 1. Consider following time fractional diffusion equation with nonhomogeneous Neumann boundary conditions

$$\begin{cases} {}_0^C\mathbf{D}_t^\alpha u(x,t) = \dfrac{\partial^2 u(x,t)}{\partial x^2} + \dfrac{\Gamma(3+\alpha)}{2}t^2 sin(\pi x) + \pi^2 t^{2+\alpha}sin(\pi x), \\ u(0,x) = 0, \ x \in (0,1), \\ u_x(0,t) = \pi t^{2+\alpha}, \ u_x(1,t) = -\pi t^{2+\alpha}, \ t \in (0,1]. \end{cases}$$

It can be checked that the exact solution of Example 1 is

$$u(x,t) = t^{2+\alpha}sin(\pi x).$$
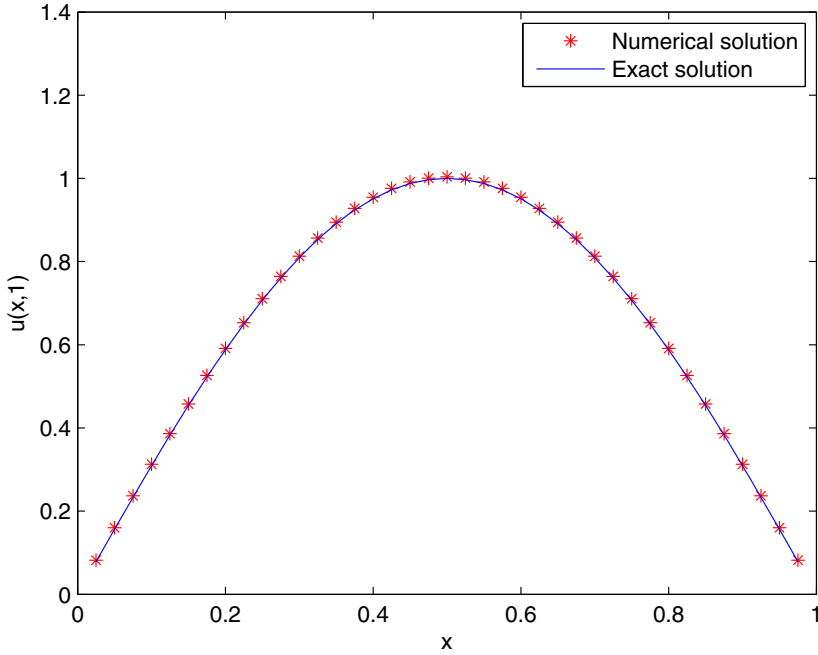
Fig. 1. The comparisons of numerical solution and exact solution for $\alpha = 0.5$, $\tau = 0.025$ and $h = 0.025$

We compute the figure and the errors $\max \left| u(x_i, 1) - U_i^N \right|$, with the index of fractional derivative $\alpha = 0.5$. The comparisons of numerical solution and exact solution is shown in Figure 1. For this case, we take time step size $\tau = 0.025$ and space step size $h = 0.025$. It can be seen that our numerical results are in excellent agreement with the exact solution. In Table 1, we take $h =$5.0e-4, a value small enough such that the space discretization errors are negligible as compared with the time errors, and choose different time step size to obtain the numerical convergence order in time. We can check that these numerical convergence order, almost approach 1.5 $(2-\alpha)$, are consistent with the theoretical analysis in Section 3. In Table 2, with sufficiently small time step size $\tau=$5.0e-5, the convergence order in space of our compact scheme is also checked.

The following numerical example is provided for the comparisons between our compact scheme and the box-type scheme [20].

Example 2. (time fractional diffusion equation with homogeneous Neumann boundary conditions)

$$\begin{cases} {}^{\mathbf{C}}_{\mathbf{0}}\mathbf{D}_{\mathbf{t}}^{\alpha} u(x,t) = \dfrac{\partial^2 u(x,t)}{\partial x^2} + \dfrac{\Gamma(3+\alpha)}{2} t^2 cos(\pi x) + \pi^2 t^{2+\alpha} cos(\pi x), \\ u(0,x) = 0, \ x \in (0,1), \\ u_x(0,t) = 0, \ u_x(1,t) = 0, \ t \in (0,1]. \end{cases}$$

The exact solution of Example 2 is

$$u(x,t) = t^{2+\alpha}cos(\pi x).$$

From Tables 3 and 4, we can find that when $\alpha = 0.3$, the accuracy of our compact scheme is higher than that of box-type scheme; when $\alpha = 0.9$, the situation is entirely different. For other values of α, similar phenomena can be observed. Namely, compared with the box-type scheme, our compact scheme has a higher accuracy for relatively small α.

Table 1. The errors for different step size τ and $\alpha = 0.5$, $h =$5.0e-4

step size τ	max $|u(x_i,1) - U_i^N|$	convergence order
1/10	1.3829e-2	
1/20	5.1179e-3	1.434
1/40	1.8663e-3	1.455
1/80	6.7422e-4	1.469
1/160	2.4227e-4	1.477
1/320	8.6978e-5	1.478
1/640	3.1036e-5	1.487

Table 2. The errors for different step size h and $\alpha = 0.5$, $\tau=$5.0e-5

step size h	max $|u(x_i,1) - U_i^N|$	convergence order
1/10	3.8790e-2	
1/20	7.9521e-3	2.286
1/40	1.8766e-3	2.083
1/80	4.6227e-4	2.021
1/160	1.1526e-4	2.004

Table 3. Our compact scheme

τ	h	$\alpha = 0.3$	$\alpha = 0.9$
		max $|u(x_i,1) - U_i^N|$	max $|u(x_i,1) - U_i^N|$
1/20	1/20	3.2072979e-3	1.0626256e-2
1/40	1/40	6.8332624e-4	4.3191515e-3
1/80	1/80	1.5648106e-4	1.8919361e-3
1/160	1/160	3.8076888e-5	8.5728264e-4
1/320	1/320	9.6547543e-6	3.9432526e-4

Table 4. Box-type scheme

τ	h	$\alpha = 0.3$	$\alpha = 0.9$
		$\max \left| u(x_i, 1) - U_i^N \right|$	$\max \left| u(x_i, 1) - U_i^N \right|$
1/20	1/20	3.3578996e-3	4.7047736e-3
1/40	1/40	8.3107151e-4	2.9664333e-3
1/80	1/80	2.0333094e-4	1.5722128e-3
1/160	1/160	4.9356687e-5	7.7979931e-4
1/320	1/320	1.1878472e-5	3.7526818e-4

5   Concluding Remarks

In this paper we have constructed a compact scheme for the time fractional diffusion equation subjected to Neumann boundary conditions with convergence rate $O(\tau^{2-\alpha} + h^2)$ by combining the classic $2 - \alpha$ finite difference method for Caputo derivative in time, the second order central difference method in space and the compact difference treatment for Neumann boundary conditions. The solvability, stability and convergence of this scheme have been rigorously proved. The numerical examples have verified the theoretical results. Maybe we can conclude that our compact scheme has a higher accuracy than that of the box-type scheme when the value of index α is relatively small.

References

1. Benson, D.A., Wheatcraft, S.W., Meerschaert, M.M.: Application of a fractional advection-dispersion equation. Water Resour. Res. 36, 1403–1412 (2006)
2. Benson, D.A., Wheatcraft, S.W., Meerschaert, M.M.: The fractional-order governing equation of Lévy motion. Water Resour. Res. 36, 1413–1423 (2006)
3. Gorenflo, R., Mainardi, F., Scalas, E., Raberto, M.: Fractional calculus and continuous-time finace. III, The diffusion limit. In: Mathematical Finance. Trends in Math., pp. 171–180. Birkhäuser, Basel (2001)
4. Gorenflo, R., Mainardi, F., Moretti, D., Paradisi, P.: Time fractional diffusion: a discrete random walk approach. Nonlinear Dynam. 29, 129–143 (2002)
5. Koeller, R.C.: Application of fractional calculus to the theory of viscoelasticity. J. Appl. Mech. 51, 229–307 (1984)
6. Langlands, T.A.M., Henry, B.I.: The accuracy and stability of an implicit solution method for the fractional diffusion equation. J. Comput. Phys. 205, 719–736 (2005)
7. Li, X.J., Xu, C.J.: A space-time spectral method for the time fractional differential equation. SIAM J. Numer. Anal. 47(3), 2108–2131 (2009)

8. Lin, Y.M., Xu, C.J.: Finite difference/spectral approximations for the time-fractional diffusion equation. J. Comput. Phys. 225, 1533–1552 (2007)
9. Liu, F., Shen, S., Anh, V., Turner, I.: Analysis of a discrete non-Markovian random walk approximation for the time fractional diffusion equation. ANZIAM J. 46(E), 488–504 (2005)
10. Mainardi, F.: Fractional calculus: Some basicproblems in continuum and statisti-calmechanics. In: Carpinteri, A., Mainardi, F. (eds.) Fractals and Fractional Calculus in Continuum Mechanics. Springer, Wien (1997)
11. Meerschaert, M.M., Tadjeran, C.: Finite difference approximations for fractional advection-dispersion flow equations. J. Comput. Appl. Math. 172, 65–77 (2004)
12. Meerschaert, M.M., Scalas, E.: Coupled continuous time random walks in finance. Phys. A 370, 114–118 (2006)
13. Podlubny, I.: Fractional differential equations. Academic Press, New York (1999)
14. Raberto, M., Scalas, E., Mainardi, F.: Waiting-times and returns in high-frequency financial data: An empirical study. Phys. A 314, 749–755 (2002)
15. Samko, S.G., Kilbas, A.A., Marichev, O.I.: Fractional integrals and derivative: theory and applications. Gordon and Breach, New York (1993)
16. Sun, Z.Z., Wu, X.N.: A fully discrete difference scheme for a diffusion-wave system. Appl. Numer. Math. 56, 193–209 (2006)
17. Sun, Z.Z.: Compact difference schemes for heat equation with Neumann boundary conditions. Numer. Meth. Part. D. E. 25, 1320–1341 (2009)
18. Yuste, S.B.: Weighted average finite difference methods for fractional diffusion equations. J. Comput. Phys. 216, 264–274 (2006)
19. Zhao, J., Dai, W.Z., Niu, T.C.: Fourth-order compact schemes of a heat conduction problem with Neumann boundary conditions. Numer. Meth. Part. D. E. 23, 949–959 (2007)
20. Zhao, X., Sun, Z.Z.: A box-type scheme for fractional sub-diffusion equation with Neumann boundary conditions. J. Comput. Phys. 230, 6061–6074 (2011)

Particle-Based Volume Rendering of Remote Volume Datasets Using FlowVR

Allan Lorant[1], Alexandre Ancel[2], and Kun Zhao[1], Naohisa Sakamoto[3], Koji Koyamada[3], and Bruno Raffin[2]

[1] Graduate School of Engineering, Kyoto University, Japan
[2] Laboratoire d'Informatique de Grenoble (LIG), INRIA, France
[3] Institute for the Promotion of Excellence in Higher Education, Kyoto University, Japan

Abstract. In this paper, we present two different FlowVR systems aiming to render remote data using a Particle-based Volume Renderer (PBVR). The huge size of irregular volume datasets has always been one of the major obstacles in the field of scientific visualization. We developed an application with the software FlowVR, using its functionalities of "modules" which can be mapped and executed on remote computers, and communicating via ssh connections. Using a "Pixels Read-back and Send" and a "Particles Send" algorithms, we succeeded in creating an application designed for developing visualization programs with remote data particle-based rendering.

Keywords: remote large data visualization, volume rendering, PC cluster.

1   Introduction

Recently, some techniques have been developed to render irregular volume datasets. Among them, PBVR [1] generates a set of emissive opaque particles from a given volume dataset, which are then rendered. We wanted to use PBVR to visualize this kind of data, overpassing the problems related to its huge size. Two solutions have been designed: the first one consists in doing the generation of particles and the rendering on a remote machine on which the data is stored, and then in reading back the pixel buffer displayed. This pixels buffer is sent and displayed on the user's computer. The second solution consists in generating particles on the remote machine, and sending those particles to the user's machine which will render them.

Some researches have already been led on the field of distributed visualization of big datasets [2]. What is new in this study is that we chose the software FlowVR [3][4] to create this application. The major advantage of this software is its environment adapted to develop complex visualization applications. The coding with FlowVR starts by creating modules. Each module encapsulates a piece of code, and can communicate with other modules via output and input ports. The software allows multi-processing, and we can easily map each module on a particular machine, the communication being done via ssh connexions.

This paper describes the created global application which encapsulates the two previously cited systems, and compares their performances depending especially on the size of the dataset and the size of the visualization window.

2   FlowVR

FlowVR is a professional software designed to allow the user to develop and run high performance interactive applications on PC clusters and Grids. FlowVR, with its concept of "modules", is adapted to virtual reality and scientific visualization. The development of a new application consists in two steps: First, we develop modules, which encapsulates portions of code. This code can be multi-threaded or parallel. Then, the modules are linked together by their output and input ports, and they are mapped onto the target architecture, and assembled into a network.
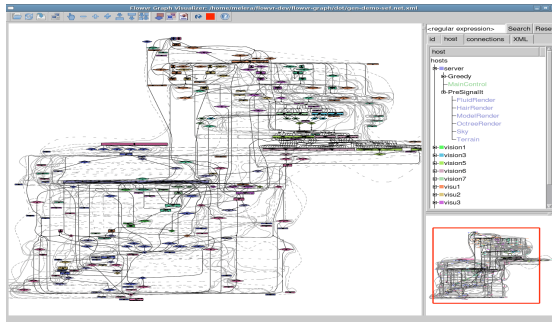
Fig. 1. FlowVR glgraph [6]

FlowVR presents a lot of tools helping the programmer to develop his applications. The most useful, the "glgraph" (see Fig. 1) is a schematic representation of the application (every components and links are detailed) automatically generated at the execution of the application. Another tool, the "gltrace", visualizes the exchange of messages between the different modules depending on the time during an execution. We chose FlowVR to lead those researches because of its development environment very well adapted to the creation of scientific visualization applications. We plan to use FlowVR to develop our future visualization programs, and choosing this software to do researches on the visualization of remote datasets will make possible to use the two systems described in this paper with those future programs, just by linking them via module interconnections.

3   Particle-Based Volume Rendering

In our rendering systems, we use PBVR to render the unstructured dataset. PBVR is based on the density emitter model of Sabella's theory, which uses grainier particles than the splatting algorithms and assumes that the particles are fully opaque. With this proposed particle model, our system requires neither alpha blending nor visibility ordering when rendering.

PBVR can be generally divided into three steps: particle generation, particle projection and ensemble averaging (see Fig. 2). Firstly, the particle generation is to generate particles from the volume data by using the given transfer function (see Fig.

2.A). The particle is generated in cell-by-cell manner. As the pre-process, this process is done by CPU. And then the generated particles are projected to the image plane in the particle projection process (see Fig. 2.B). This projection process is performed by GPU. At the same time, the calculation of the particle size and shadow processing is also performed by using normal vector. The third step ensemble averaging process means that, for the group of particles generated by using different random numbers, particle projection process is performed multiple times, and the generated image superimposes each other averagely (see Fig. 2.C). The multiple times is called repeat level.

As a result, high-quality rendering results can be obtained by setting a high repetition number. However, with the increase of the particle number, the particle data size is also increasing which would put pressure on the CPU memory and may lead to a deceleration of the rendering speed. To solve this problem, PBVR provide a LOD control which allows user to control the level of detail dynamically by changing the number of particles projected to the image plane. By reducing the repetition level, we can obtain a high-speed rendering result.
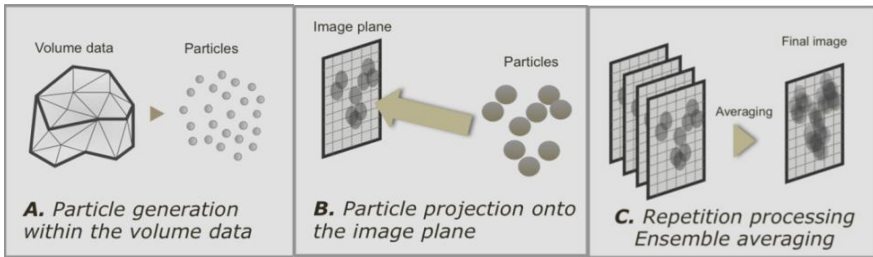
Fig. 2. Three steps of PBVR: Particle generation, Particle projection, Ensemble averaging [5]

4   Proposal Method

4.1   Pixels Read-Back and Send System

The "Pixels Read-back and Send" system consists in reading and rendering the dataset on the remote computer, copying the pixel buffer and sending it to the user's machine. For example, we name "PC1" the user's machine and "PC2" the remote machine with the datasets. We can see on Figure 3 the algorithm of the two created modules. On PC2, the "Compute" module uses KVS (library used by PBVR) to open a new window, reads the data, which is then rendered with KVS. Two versions are actually implemented: one using the PBVR renderer (see Fig. 8.A, 8.B and 8.C) and one using the polygon renderer. Each time the display is actualized, the pixel buffer of the visualization window is copied and sent via the output port "PixelOut" to PC1. The "Display" module on PC1 receives this pixels array and displays it in a new window. Each time the user reshapes his visualization window, the new size is sent to PC2 to adapt his own visualization window size.
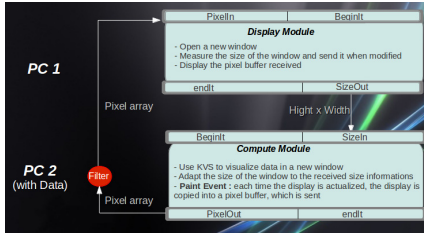
Fig. 3. Structure of "Pixels Read-back and Send" system
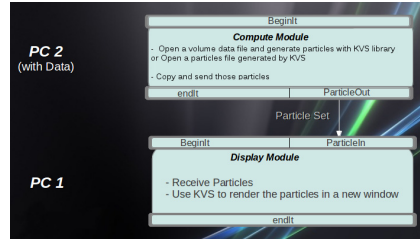
Fig. 4. Structure of "Particles Send" System

3 bytes per pixel
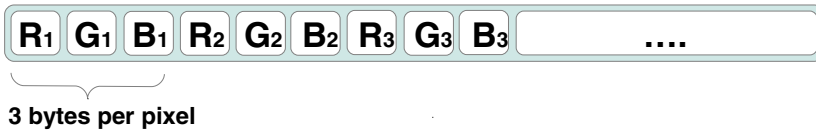
Fig. 5. Composition of the messages in "Pixels Read-back and Send" system

The structure of the message containing the pixel buffer is shown in the figure 5. Each pixel is stored on 3 bytes, one after the other. The three bytes correspond to the 3 colour components of pixels, Red, Green and Blue. We can deduce a general formula giving the size of sent messages in function of the size of the visualization window (of dimensions width x height):

$$Message_{Pixels} = (3 \text{ x Width x Height}) \text{ bytes} . \tag{1}$$

In FlowVR, the linked entry ports block the module until they receive messages. In this situation, we want the "Compute" module working 100% of the time, to be able to capture every single mouse control done by the user. So we have to set the entry port of this module to "non blocking port". Doing this, the "Compute" module is capable of sending a lot of messages in a short time. But the "Display" module has to finish its current iteration before receiving the next message. The messages waiting are stored in memory, so there is a possibility of overflow of the dedicated memory. We avoid this possibility by adding a filter between the two modules, which deletes messages which can't be treated in time, avoiding any possible overflow. Messages are sent when the display is changed, or when the window is reshaped. It comes that when the display doesn't change, no message is sent, in which case the application doesn't consume any resource.

4.2   Particles Send System

We wanted to propose a solution allowing to visualize remote data with a big display, and without sacrificing the interactivity by sending regularly huge messages. We decided to separate the PBVR processes (see Fig. 2) in two. In this second system, the rendering is done on PC1, while it was done on PC2 on the "Pixel Read-back and

Send" system (see Fig. 4). On PC2, the "Compute" module uses KVS library to generate particles from a volume data file, or directly opens a particles file, itself previously generated by KVS. Those particles are copied and are sent to PC1. On PC1, the "Display" module receives the particles and uses PBVR to render them. On this system, only one message is sent by the "Compute" module, so there is no overflow risk due to the messages frequency. This operation corresponds to a download of particles after their generation on the remote machine, whereas the "Pixels Read-back and Send" is more a real cooperation of two modules with exchanges of messages. This unique message is a good point for bad networks. The particle data being smaller than the original irregular volume data, this method is far faster than downloading the entire volume data from the remote machine.
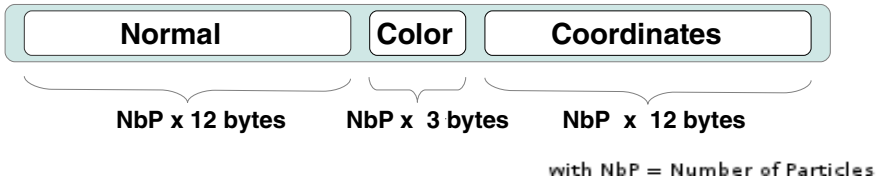
Fig. 6. Composition of the message in "Particles Send" system

Each particle has three components: a normal vector (x_n, y_n, z_n) in three floats (14 bytes), a colour array (r, g, b) (3 bytes), and a set of coordinates (x, y, z) in three floats (14 bytes). The composition of particle messages starts with the normal components, then the colour components, and finally the coordinates (see Fig. 6). We can deduce the general formula for particle messages:

$$\text{MessageParticles} = (27 \times \text{Number of Particles}) \text{ bytes .} \tag{2}$$

The normal component of particles is not essential to obtain a visualization of the data. Indeed, we can choose not to use it in exchange of a loss of image quality (see Fig 8.D and 8.C). We decided to implement this "reduced" version of the "Particles Send" system to improve the performances of the final application. The message of this version has the same structure than in the regular version of "Particles Send", except for the normal parameters which have been removed (see Fig. 7).
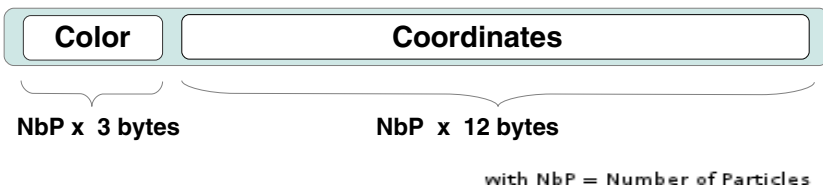
Fig. 7. Composition of the message in reduced version of "Particles Send" system

The new message size is given by the following formula:

$$\text{MessageParticles(Reduced)} = (\ 15 \text{ x Number of Particles) bytes .} \qquad (3)$$

The reduction ratio of this system is theoretically of 0.5556. This reduction should decrease the message size of 45%.

4.3   Global Application

These two systems have been unified in a global application. The user chooses between two launch files: one for local execution (all modules are executed on the local machine) and one for distributed execution (modules are executed on different machines using the .csv file as described below). Some commands are also available (compile, build, clean …). The user can determine the nature of the execution of the program (nature of the application, the data path and others) by editing a parameters text file (see Table 1).

The choice of the system is done with the "send" parameter. The "gpu" parameter allows the user to run the application only on CPU, which reduces drastically the image quality but allows a rendering on machines without GPU. The "lod" parameter enables a progressive level of details. When the user moves the camera, the PBVR rendering is done with a low repetition level, so the image is of medium quality but high interactivity. When the camera position stops from moving, the repetition level is increased automatically to the set one. It allows a high quality rendering with high interactivity (high fps). If the GPU rendering is disabled, this parameter has no role. The path of the volume data or the particle data is set with the "path" parameter. We can also note that the user can choose the "reduced" version of "Particles Send" system (if this last one has been chosen) by setting the "reduce" parameter to 1. In case of "Pixels Read-back and Send" chosen system, this parameter has no role.

A ".csv" file is also used by FlowVR to map the modules in view of a distributed execution. On the first line, the names and IP addresses of the used machines are written. On the first column, the name of each module is indicated. The user just has to choose how many iterations of each module he wants on each machine by indicating the number on the corresponding cell.

This structure allows the user to easily control the whole application with only one parameters text file and one module mapping .csv file.

Table 1. Setting parameters of the global application

Parameters	Description	Value	Condition of use
send	Sending Type	pixel or particle	
repeat	Repetition level	1, 4, 9, 16, 25, …	
gpu	GPU rendering	0:disable, 1:enable	
lod	LOD rendering	0:disable, 1:enable	gpu=1
data	Path to the data	[path to the data]	
reduce	Reduced version	0:disable, 1:enable	send=particle

5   Experimental Results

In this paper, we test the efficiency of the two developed systems with an oral airflow irregular volume dataset. First of all, we present different visualization results we can obtain with the two systems using the previously cited dataset (see Fig. 8.A, 8.B and 8.C).

We want to compare the two programs in different situations, to be able to determine what the field of application of each one is. In theory, the "Pixels Read-back and Send" program is supposed to have efficiency depending on the size of the visualization window, so we chose to experiment our project with two different sizes of display: 300x500 and 1249x934. On the other hand, "Particle Send" should have its efficiency depending on the number of particles used. We decided to test the two systems with 2 different generated particle data with different sizes (see Fig. 8.A and 8.B).
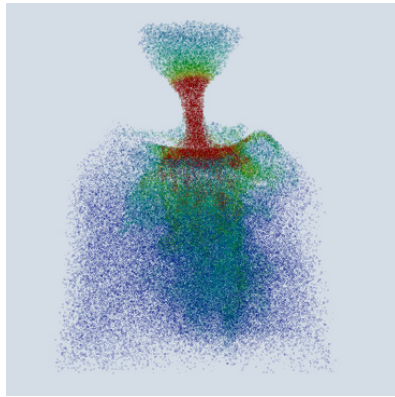
Fig. 8. A. PBVR rendering with repetition level = 9
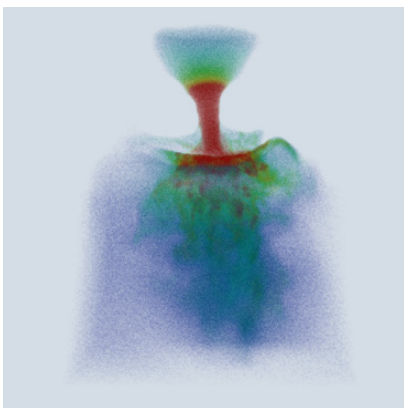
Fig. 8.B. PBVR rendering with repetition level = 144
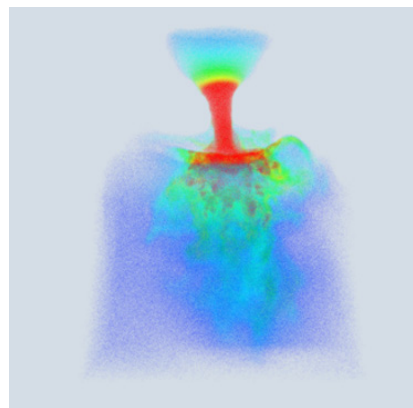
Fig. 8.C. PBVR rendering with reduced particle data and repetition level = 144

We experimented the distributed execution between two machines on a local network (1G bandwidth). As results, we observe the bandwidth occupation of the network by the messages in 10 seconds of execution (see Fig. 9). Those results confirm the theory: the efficiency of "Pixels Read-back and Send" is mainly depending on the size of the visualization window, because the pixels arrays are sent each time the display is updated. We can see that "Particles Send" system is very effective with files with a small number of particles (see Fig. 9.A and 9.B). "Pixels Read-back and Send" seems to be a good solution for visualizing remote data files with a huge number of particles (see Fig. 9.C and 9.D). Between those 2 systems, the "reduced" version of "Particles Send" should be a good compromised, as it saves 45% of the starting time of its non-reduced version, with a short loss of quality.
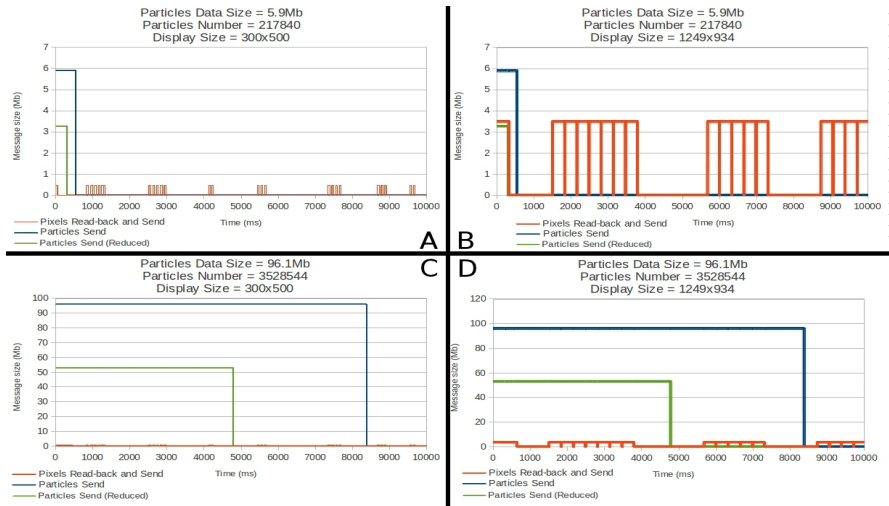
Fig. 9. Messages sizes and sending times on a 10 seconds distributed execution with 4 configurations
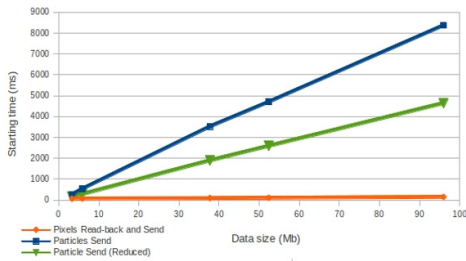
Fig. 10. Influence of particle data size on starting time of both systems with fixed display size (500x300) and repetition level = 9
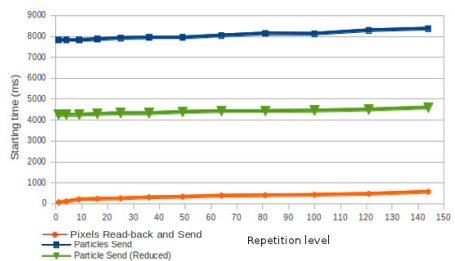
Fig. 11. Influence of the repetition level on starting time of both systems with fixed display size (500x300) and data size (96.1Mb)

To get a visualization of the data after launching the start command, each system has to send at least one message. The previous results and the theory (see formulas (2) and (3)) let us guess that there is a relation of proportionality between the number of particles used and the starting time of the "Particles Send" system (starting time is considered as being the time between the execution of the launch command of the application and the display of the visualization window). We decided to test the starting time of both systems for different particle data, to confirm this theory (see Fig. 10). We can see that the hypothesis was confirmed by this experiment. "Particles send"'s starting time presents a relation of proportionality with the data size (the data size is directly linked with the number of particles as all particles have the same size in memory).

The starting time of "Pixels Read-back and Send" system is increasing significantly with the number of particles, whereas the message size is independent from it. Indeed, the starting time corresponds not only to the time necessary to send one message, but also refers to the rendering time and the application launching time (we use in this experiment particle data, so the particle generation time is not included to the starting time). We can define the starting time of the application for both systems by the formula:

$$\text{Starting Time} = T_{FlowVRProcess} + T_{ReadingParticleFile} + T_{Rendering} \\ + T_{SendingMessage} + T_{ReceivingMessage} .$$

(4)

Using the formulas (1) and (2), we can deduce one formula for each system:

$$\text{Starting Time (Particles)} = a \times NbP + b \times Width \times Height + \varepsilon \ \mu s .$$

(5)

$$\text{Starting Time (Pixels)} = c \times NbP + d \times Width \times Height + \delta \ \mu s .$$

(6)

In these formulas, we consider NbP = Number of Particles, Width and Height the dimensions of the visualization window, and ε and δ two constants.

We know, regarding to the theory and the results of the experiments, that NbP has a great influence on the starting time of "Particles Send" system, and not the display size (see Fig. 9 and 10). On the other hand, the display size has a great influence on the starting time of the pixel-based system, and NbP has a small influence. We should have $a \gg b$ and $d > c$. In this experiment, we found by regression analysis $a = 2.44963$, $b = 0$, $\varepsilon = 45$ ms for the first system, and $c = 0.023862$, $d = 0.272719$, $\delta = 20$ ms for the second one, confirming the theory. However, it has to be noted that the coefficients δ and ε are very small in this experiment, and might have been under-evaluated, because of the absence of tests with very small data, and the precision of the measures. The coefficients a and d depend mainly on the network speed, b and c mainly on the GPU resources, and δ and ε on the CPU and GPU resources. The result $b = 0$ needs some explanations. Indeed in PBVR, the number of particles is linked to the size of the display [7]. The size of the footprints of particles is equal to the size of a pixel, so the number of particles increases with the resolution of the display. But in this system, we use only one particle data for the rendering, so the number of particles is fixed. When the resolution increases, it is the radius of particles that is increasing

(see formula (6) of paper [7] for more details). By this way, PBVR equations are respected and so the size of the display has almost no influence on "Particles Send" system's starting time.

We also tested the starting time of the developed systems in function of the repetition level of PBVR (see Fig. 11). We can observe that the starting time is increasing with the repetition level on both systems with approximatively the same slope, explained by the fact that increasing the repetition level increases the time $T_{rendering}$ needed by PBVR to render the image. This time is the same for the 2 systems under the same configuration.
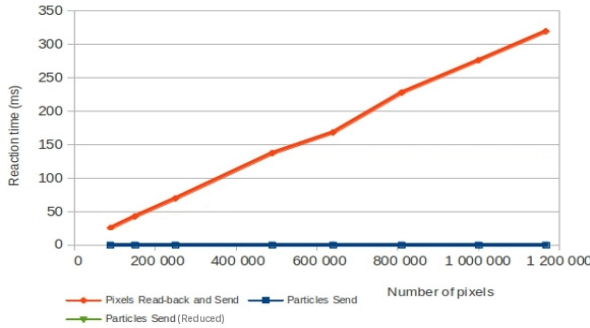
Fig. 12. Influence of the display size on the reaction time of both systems to a user stimulus with repetition level = 9
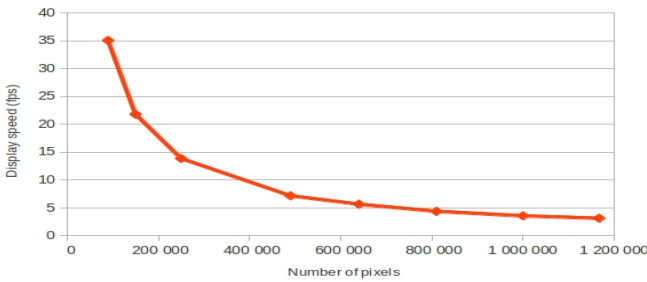
Fig. 13. Influence of the display size on the display speed of "Pixels Read-back and Send" system with repetition level = 9

One of the other important points in comparing the two systems is the fluidity and the response time of the application to a user stimulus. As we observed some variations of those two elements when the visualization window size was modified, we decided to evaluate them taking a set of different display sizes and by measuring the response time (see Fig. 12) and the display speed in frames per second (see Fig. 13). We can see that there is a relation of proportionality between the response time and the number of pixels visualized by the "Pixels Read-back and Send" system. Indeed, the visualization process induces transmissions of messages which have a size

depending on the number of pixels to display. "Particles Send" system presents no variation of response time, as the visualization process happens on the local machine, without using any message. On the Figure 13, we can observe the variation of maximum display speed of "Pixels Read-back and Send" system. Indeed, the loop in the algorithm of this system (see Fig.3) induces the presence of a maximum display speed, decreasing when the message transmission time increases (so when the pixels number becomes bigger). The maximum speed of "Particles Send" is not on this figure, because it depends on the GPU resources and the number of particles. But this maximum speed is superior or equal to the maximum speed of the showed results in Figure 13 (superior to 35 fps in this experiment).

We also were able to study the reduction ration of the "Reduced" version of "Particles Send" system. It should be noted that the reduction ration theoretical of 0.556 has been measured equal to 0.57 by studying the results presented in figures 9 to 13. These experiments showed that this version of the particles-based message system allows to save 43% percent of the message size, which allows a gain of time on the starting time which tends to 43% for huge particle datasets (the coefficients ε and b then become negligible compared to a).

The behaviours described in this study are of course more qualitative than quantitative. Indeed all time and speed indications are specific to the network speed (1G connection in these experiments), but whatever this network speed is, the behaviour of the two systems is qualitatively the same.

6   Discussion

As shown with those experiments, the "Pixels Read-back and Send" system is well adapted for huge datasets with small visualization window. The main advantage of this system is that its efficiency is very few influenced by the number of particles, and almost only depends on the size of the visualization window. Moreover, it allows the execution of any visualization program, and is not limited to a PBVR rendering. The user can define his own visualization program in the compute module with the same template, and he will get the same functionalities of window copying than the system presented in this paper. We can note that, as all calculations are done on the remote machine, the user's machine needs minimal resources and the user can obtain a high quality rendering on a low-capacities machine. The weak point of the system is that the interactivity and the reaction time of the application decreases inversely with the size of the visualization window (see Fig. 12 and Fig. 13). Moreover, the high number of messages can saturate networks with small bandwidth. One of the solutions for high resolution visualization should be to reduce the display size when the user wants to change the camera position (at this point we have a small window with high interactivity propitious to manipulations), and to enlarge it once the new camera position is set.

On the other side, "Particles Send" offers a larger scope of applications. If the execution takes generally more time than the previous system, the interactivity is maximum at any time, and the program is smooth without using any bandwidth

(the particles transmission can take a long time for big datasets, but this transmission is the only message sent in this system). It also presents a parameter allowing to reduce the particle data size down to 57% of its original size, with only a low loss of lightening quality (seeing Fig. 8.B and 8.C). Its weak point is that for very huge particle data (for example for time varying data, with one particle data file per frame), the starting time becomes important (see Fig. 10).

In this paper we propose an application which encapsulates two different systems developed on FlowVR, allowing the user to render with PBVR remote irregular datasets on his own machine, using ssh connections to communicate. Both systems have different characteristics and scopes of using, and are easy to use thanks to a set of parameters. In future, we are going to upgrade those 2 systems. For "Pixels Read-back", we plan to add mouse controls on the "Display" module and to work on compression and interpolation of pixel array to improve the interactivity and decrease the bandwidth utilization. We want to implement "Particles Send" to make it able to render irregular time varying data and want to test it with several executions of "Compute" module on different machines to read and send huge particle data more quickly. We are also planning to test the application on long distance distributed execution between France and Japan.

References

1. Sakamoto, N., Nonaka, J., Koyamada, K., Tanaka, S.: Particle-based Volume Rendering. In: Proc. of Asia-Pacific Symposium on Visualization 2007, pp. 141–144 (2007)
2. Sakamoto, N., Kuwano, H., Kawamura, T., Ebara, Y., Koyamada, K., Nozaki, K.: Distributed Particle-based Volume Rendering for Irregular Volumes. In: The First International Workshop on Super Visualization (IWSV 2008), CD-ROM (2008)
3. Allard, J., Gouranton, V., Lecointre, L., Limet, S., Melin, E., Raffin, B., Robert, S.: FlowVR: A Middleware for Large Scale Virtual Reality Applications. In: Danelutto, M., Vanneschi, M., Laforenza, D. (eds.) Euro-Par 2004. LNCS, vol. 3149, pp. 497–505. Springer, Heidelberg (2004)
4. Arcila, T., Allard, J., Menier, C., Boyer, E., Raffin, B.: FlowVR: A Framework for Distributed Virtual Reality Applications. Journees de l'AFRV (2006)
5. Zhao, K., Nishimura, J., Sakamoto, N., Koyamada, K.: A New Framework for Visualizing a Time-varying Unstructured Grid Dataset with PBVR. In: AsiaSim 2011 (2011)
6. FlowVR Documentation, `http://flowvr.sourceforge.net/index.html`
7. Sakamoto, N., Kuwano, H., Kawamura, T., Koyamada, K., Nozaki, K.: Visualization of Large-scale CFD Simulation Results Using Distributed Particle-Based Volume Rendering. International Journal of Emerging Multidisciplinary Fluid Sciences (IJEMFS) 2(2-3), 73–86 (2010)

A Fast Intuitionistic Fuzzy Support Vector Machine Algorithm and Its Application in Wind Turbine Gearboxes Fault Diagnosis

Bin Jiao and Qing Zhang

Shanghai DianJi University, Shanghai, 200240, China
East China University of Science and Technology, Shanghai, 200237, China
`jiaob@sdju.edu.cn, tsingo611@163.com`

Abstract. Support vector machine has been successfully applied to the fault diagnosis field, but there are still some problems in practical applications. In this paper we proposed an improved algorithm which reduces the number of support vectors through the reduction of the sample spae to improve the efficiency of the algorithm. As the traditional fuzzy support vector machine cannot classify the sample with the same membership, so we use intuition index to lower the probability of the sample to get the same membership. Here we improve the accuracy of the algorithm through properly redefine the fuzzy membership and intuition index. Finally, we use the improved algorithm to build a multi-classifier based on one against one principle and the voting rules, and apply the multi-classification algorithm to the wind turbine gearbox fault diagnosis. The diagnose results prove that the improved algorithm we proposed can properly resolve the problem of wind turbine gearboxes fault diagnosis.

Keywords: wind turbine gearbox, fault diagnosis, intuitionistic fuzzy, support vector machine.

1 Introduction

Gearbox is the main transmission device of the wind turbine, and it is also the component that is most probably to failure. According to the data statistics, the longest parking time is caused by the failure of gearbox. The gearbox fault also caused the greatest loss of generating capacity. Once the wind generator fault occurs, large amounts of money and a lot of human and non-human sources will be put into maintaining the equipment, which causes huge losses. So the study of gearbox fault diagnosis has great significance. Along with the development of the intelligent algorithm, more and more intelligent algorithms are introduced to the fault diagnosis field, and many intelligent diagnosis algorithms has been successfully used in gearbox fault diagnosis, such as particle swarm algorithm[1], neural network algorithm[2], genetic algorithm[3], etc.

Support Vector Machine (SVM) has a good statistical learning theory basis and excellent generalization ability, and it can solve a variety of practical problems better than other algorithms which makes it a favorite to many scholars [4-5]. The SVM

algorithm has successfully applied to different fields, such as pattern recognition, function fitting, modeling and control area and so on. But there are still some problems for the SVM algorithm in practical application. Therefore many scholars make many modifications to the SVM algorithm. In order to improve the calculation speed of the algorithm, Suykens [6] puts forward the Least Square Support Vector Machine (LS-SVM), through making some modifications to the structure of the Support Vector Machine to reduce the computational complexity, which has greatly improved the efficiency of the algorithm. But for the mass sample the computing speed of this algorithm is still very slow. Osuna [7], Yang [8], Liu [9] try to find an effective pre-extraction of SV to simplify the SVM algorithm which usually adopts the distance-based methods. But the efficiency of the algorithm for the random and arbitrary form of massive samples is still very low. The basic SVM algorithm cannot process samples containing noises and outliers accurately. In order to overcome this defect, Lin [10-11] proposes Fuzzy Support Vector Machine (FSVM), through introducing the fuzzy membership to distinguish the contribution of the different samples in classification, which can reduce the influence of noise and the outlier in classification. But this method can't distinguish samples with the same membership accurately. According to the deficiency of the traditional SVM algorithm, Ha Ming-Hu [12] comes forward Intuitionistic Fuzzy Support Vector Machine (IFSVM) based on the traditional FSVM algorithm, which obtains good results.

The above methods are only used to improve a particular aspect of the performance of the algorithm, which cannot gain a good classification accuracy and better calculation speed at the same time. In order to make the algorithm have a good performance both in accuracy and efficiency, we will propose a new method to improve the IFSVM algorithm. This method can improve the efficiency of the algorithm through reducing the fault characteristics of samples, and improve the accuracy of the algorithm through properly redefine the fuzzy membership and intuition index. In the end of this paper, the new algorithm will be applied to diagnose the failure of wind turbine gearbox, and the diagnose results prove that our method is more effective and efficient.

2   Pre-extraction Intuitionistic Fuzzy Support Vector Machine

2.1   Reduction of the Fault Features Sample

For a given training sample set $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, set the first kind of sample set as S_1, which means $S_1 = \{x \mid y = 1, x \in S\}$, set the second kind of sample set as S_2, which means $S_2 = \{x \mid y = -1, x \in S\}$. Now we define a N-dimensional minimum span vector D_{span} in the N-dimensional vector space in which the sample set stayed. The expression of the D_{span} is:

$$D_{span} = \{d_1, d_2, \cdots, d_n\} = \frac{1}{m}\{\max(x_1) - \min(x_1), \cdots, \max(x_n) - \min(x_n)\}$$

(1)

Where $\max(x_i)$ is the maximum data of each dimension in the sample set, $\min(x_i)$ is the minimum data of each dimension in the sample set, and m is a constant whose value is $1/k$ of the sample size, and usually the value of k is less than 10.

Set the sample which is nearest to the origin of the N-dimensional space as the start point, and use the length of each dimension in D_{span} to divide the sample space, so the sample space will be divided into many sample blocks (S-Block), and the side length of the S-Block is d_i. It is easy to know that the S-Blocks are many rectangles in the two-dimensional space, and many cubes in the three-dimensional space.

Every S-Block may contain the first kind of sample or the second kind of sample or them both. We use φ^+ to represent the number of the first kind of sample in the S-Block, and φ^- to represent the number of the second kind sample in the S-Block. Now we define the density of S-Block as:

$$\rho = \frac{(\varphi^+ - \varphi^-) + 1}{|\varphi^+ - \varphi^-| + 1} \bullet \frac{\max(\varphi^+, \varphi^-)}{\varphi^+ + \varphi^-}$$

(1)

According to the equation we can conclude, when $\varphi^+ \geq \varphi^-$, ρ is nonnegative value, when $\varphi^+ < \varphi^-$, ρ is negative value. Set the threshold as $\theta(0 \leq \theta \leq 0.5)$, for a S-Block, if $\rho \leq \theta$, we call this S-Block as boundary block (B-Block); if $\rho > \theta$, we call this S-Block as interior block (I-Block).

From the definition of the density of the S-Block we can know the edge sample may exist in the B-Block and I-Blocks around it, and we call these S-Blocks as candidate block (C-Block). Now we use all candidate blocks to create a new sample S^{CB}:

$$S^{CB} = \{(x_i^{CB}, y_i^{CB}) | (x_i^{CB}, y_i^{CB}) \in S, i = 1, \cdots, l^{CB}, l^{CB} > 1\}$$

(3)

$$S_1^{CB} = \{x^{CB} | y^{CB} = 1, x^{CB} \in S^{CB}\}, |S_1^{CB}| = l_1^{CB}$$

(4)

$$S_2^{CB} = \{x^{CB} | y^{CB} = -1, x^{CB} \in S^{CB}\}, |S_2^{CB}| = l_2^{CB}$$

(5)

The relations are $S^{CB} \in S, S_1^{CB} \in S_1, S_2^{CB} \in S_2, S_1^{CB} \bigcup S_2^{CB} = S^{CB}$, so we can conclude the samples in the sample center that has little contribution in classification are removed from the sample set after the sample set is divided into sample blocks and the candidate blocks are picked up, so the sample size is reduced. The reduction of the sample size can definitely improve the efficiency of the algorithm.

2.2   Intuitionistic Fuzzy Support Vector Machine

2.2.1   The Determination of Fuzzy Membership

After picking up C-Blocks, we get the new sample set S^{CB} and its subspace S_1^{CB}, S_2^{CB}. We can know the number of the first kind sample in the sample space S^{CB} is l_1, and the number of the second kind sample in the sample space S^{CB} is l_2, so we can calculate the center of the first kind sample is O_1 :

$$O_1 = \frac{\sum_{i=1}^{l_1} x_i}{l_1}, x_i \in S_1^{CB} \tag{6}$$

The center of the second kind sample in the sample space is O_2:

$$O_2 = \frac{\sum_{i=1}^{l_2} x_i}{l_2}, x_i \in S_2^{CB} \tag{7}$$

We define the length between the sample points as:

$$D(x_i, y_i) = \|x_i - x_j\| (i = 1, 2, \cdots l; j = 1, 2, \cdots l) \tag{8}$$

Now we set the fuzzy membership function to the first kind sample as:

$$\mu_i^1 = \begin{cases} 1 - \dfrac{D(x_i, O^1)}{\max D(x_k, O^1)}, & \left(D(x_i, O^1) \le \min_j D(x_j, O^1), y_k = y_i = 1, y_j = -1\right) \\ m \bullet \left(1 - \dfrac{D(x_i, O^1)}{\max D(x_k, O^1)}\right) & \text{other} \end{cases} \tag{9}$$

As the same, we set the fuzzy membership function to the second sample as:

$$\mu_i^2 = \begin{cases} 1 - \dfrac{D(x_i,O^2)}{\max D(x_k,O^2)}, \left(D(x_i,O^2) \le \min_j D(x_j,O^2), y_k = y_i = -1, y_j = 1\right) \\[4mm] m \bullet \left(1 - \dfrac{D(x_i,O^2)}{\max D(x_k,O^2)}\right) \qquad\qquad \text{other} \end{cases}$$

(10)

Where m is an adjustable parameters, the range of value allowed for m is 0 to 1.

2.2.2   The Determination of Intuition Index

In order to determine the intuition index, we must confirm the concept of the density of the similar point and the density of the disparate point. We define the density of the similar point as:

$$\rho^+(x_i,R) = \left|\left\{x_j \,\middle|\, D(x_j,x_i) \le R, y_j = y_i\right\}\right|$$

(11)

Define the density of the disparate point as:

$$\rho^-(x_i,R) = \left|\left\{x_j \,\middle|\, D(x_j,x_i) \le R, y_j \ne y_i\right\}\right|$$

(12)

Now we can use the density of the similar point and disparate point to determine the intuition index. The intuition index of the first kind of sample is defined as:

$$\pi_i^1 = \begin{cases} 0, & \left(D(x_i,O^1) \le \min_j D(x_j,O^1), y_i = 1, y_j = -1,\right. \\[4mm] \dfrac{\rho^{1+}(x_i,R)}{\rho^{1+}(x_i,R)+\rho^{1-}(x_i,R)} \bullet \dfrac{\rho^{1+}(x_i,R)}{\max_k \rho^{1+}(x_k,R)} \bullet (1-\mu_i^1), & \text{other} \end{cases}$$

(13)

The intuition index of the second kind of sample is defined as:

$$\pi_i^2 = \begin{cases} 0, & \left(D(x_i,O^2) \le \min_j D(x_j,O^2), y_i = -1, y_j = 1,\right. \\[4mm] \dfrac{\rho^{2+}(x_i,R)}{\rho^{2+}(x_i,R)+\rho^{2-}(x_i,R)} \bullet \dfrac{\rho^{2+}(x_i,R)}{\max_k \rho^{2+}(x_k,R)} \bullet (1-\mu_i^2), & \text{other} \end{cases}$$

(14)

Where the ρ^{1+} is the density of the similar point in the first kind of sample set, the ρ^{1-} is the density of the disparate point in the first kind of sample set, the ρ^{2+} is the density of the similar point in the second kind of sample set, the ρ^{2-} is the density of the disparate point in the second kind of sample set, and R is the neighborhood of the sample point.

3   The Application of PSV-IFSVM in Wind Turbine Gearbox Fault Diagnose

SVM was originally designed to resolve binary classification problem, however the gearbox have more than two kinds of fault. In order to solve this multi-classification problem, we adopt the "one against one"(OAO) method to build a multi-classifier. For an n-class classification problem, we have to build $n(n-1)/2$ classifiers to complete the classification. During the classification, all the $n(n-1)/2$ classifiers have to classify the sample set respectively. Using voting method to decide which kind of fault the sample set belongs. The typical model of OAO is shown in the following picture Figure1which is a three classifications model.
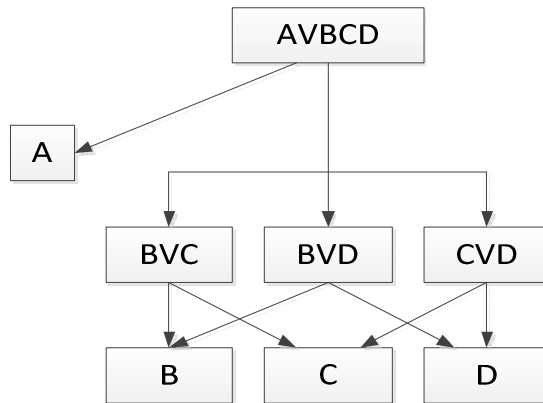
Fig. 1. Three Classifications Model Base on One Against One Principle

Now we can build the algorithm for multi-classification, and the steps of the multi-classification algorithm are as follows:

Step1: Build the pre-extraction of intuitionistic fuzzy support vector machine classifier;

Step2: According to the typical model of OAO build the gearbox fault diagnosis model;

Step3: Put the fault feature samples into the diagnosis model. By the first floor of the diagnosis model, we can judge the sample is normal or not. If the judgment is normal, then diagnostic ends; otherwise leap to step4;

Step4: Use the $n(n-1)/2$ classifiers in the second floor of the diagnosis model to classify the sample   respectively, and use the voting rules to count the votes;

Step5: Judge whether there is a largest number of votes exist, if exist, then the samples belonged to such a failure, diagnostic ends; otherwise there are fault types which have the same votes, leap to step6;

Step6: If the sample x belongs to the i fault type and j fault type has the same votes, then the result of diagnose will be given by the $IFSVM_{i,j}$ classifier directly.

The Table1 is the common vibration signal eigenvectors of the gearbox fault. We will use the following five groups of fault eigenvectors as the test sample.

Table 1. Common Vibration Signal Eigenvectors of the Gearbox Fault

No.	Real Fault Type	peak index	kurtosis index	tolerance index	Skewness index	Spectrum focus	Spectrum variance	Harmonic factor
1	Normal	0.0865	0.0697	0.0613	0.1186	0.4422	1.0000	0.4828
2	Bearing inner ring damage	1.0000	0,6001	0.7791	0.5778	0.5735	0.9175	0.1390
3	Bearing outer ring damage	0.6208	0.2622	0.4557	0.3210	0.4817	0.9842	0.0000
4	Broken teeth	0.9769	1.0000	1.0000	1.0000	0.6140	0.8732	0.0819
5	Tooth wear	0.7128	0.3986	0.5230	0.4224	0.5270	0.9585	0.1931

Now we will apply the multi-classification algorithm we build in this paper to classify the common vibration signal eigenvectors of the gearbox fault, and determine which kind fault of the gearbox it is. The result of the determination of the algorithm is shown in Table2:

Table 2. The result of the determination of the algorithm

No.	Fault type					Diagnosis
	A	B	C	D	E	
1	1	0	0	0	0	Normal
2	0	4	1	0	1	Bearing inner ring damage
3	0	2	3	0	1	Bearing outer ring damage
4	0	1	1	0	4	Broken teeth
5	0	0	2	3	1	Tooth wear

From table1 and table2, we can see that the algorithm we build could properly classify the gearbox faults; the statistical results of votes indicate that the algorithm has a good resolution. All the results show that this algorithm is of high accuracy and practical in the problem of wind turbine gearbox fault diagnosis.

4   Conclusions

In this paper we put forward a new method to improve the performance of SVM algorithm from two different aspects. As some sample points in the sample space make no contribution to classify, if we don't remove these points, the algorithm will take a long time to calculate these unhelping points, which will lower the efficiency of the algorithm, so we propose an effective method to improve the efficiency of the algorithm through effectively reducing the sample size. For the traditional FSVM algorithm cannot correctly classify samples with the same membership which will lower the accuracy of the algorithm, so the IFSVM algorithm introducing another concept of the fuzzy theory that is the intuition index to avoid the appearance of two samples with the same membership. Here we improve the accuracy of the algorithm through properly redefine the fuzzy membership and intuition index, after calculating the center of the sample, the algorithm will give a big membership to the neighborhood of the center point, and give a little membership and intuition index to the boarder part which is long away from the center point. The intuition index can effectively avoid the appearance of the same membership, so the algorithm could get a higher accuracy. In the end of this article, we use the "one against one" method and voting rules to build a multi-classifier, and apply this multi-classifier to the wind turbine gearbox fault diagnosis, the diagnose results prove that our method is more effective and efficient.

References

1. Wang, H.-L., Zhang, Z.-Q., Cui, X.-X., Song, T.: Test optimization of real-time monitoring and fault diagnosis system based on improved particle swarm optimization. Systems Engineering and Electronics 33(4), 958–962 (2011)
2. Mi, J., Ji, G.-Y.: Application of Improved BP Neural Network in Fault Diagnosis of Fans. Noise and Vibration Control 31(2), 94–98 (2011)
3. Qiao, J., Pan, H.: Gear Fault Diagnosis Based on GA-Elman Neutral Network Model. Water Resource and Power 28(6), 106–108 (2010)
4. Wang, Y., Sun, X.F., Li, M.: Training Method for Support Vector Machine Based on Chaos Particle Swarm Optimization. Computer Engineering 36(23), 189–191 (2010)
5. Ji, A.-B., Pang, J.-H., Qiu, H.-J.: Support vector machine for classification based on fuzzy training data. Expert Systems with Applications 37(4), 3494–3498 (2010)
6. Suykens, J.A.K., Vandewalle, J.: Least Squares Support Vector Machine Classifiers. Neural Processing Letters (S1370-4621) 9(3), 293–300 (1999)
7. Osuna, E., Freund, R., Girosi, F.: An Improved Training Algorithm for Support Vector Machine. In: Proceedings of the 1997 IEEE Workshop on Network for Signal Processing VII, Amelea Island, USA, pp. 276–285. IEEE, USA (1997)
8. Yang, M.H., Ahuja, N.: Geometric Approach to Train Support Vector Machine. In: Proceeding of IEEE Computer Society Conference on Computer Vision an Pattern Recognition, Hilton Head, SC, USA, pp. 430–437. IEEE, USA (2000)
9. Liu, W.-L., Liu, S.-Y., Du, Z.: Sample Decreasing Method Based on Distance in SVM 23(3), 333–337 (2008)
10. Lin, C., Wang, S.D.: Fuzzy support vector machines. IEEE Transaction on Neural Network 13, 464–471 (2002)
11. Lin, C., Wang, S.D.: Fuzzy Support Vector Machines with Automatic Membership Setting. In: Wang, L. (ed.) Support Vector Machines: Theory and Applications. STUD FUZZ, vol. 177, pp. 233–254. Springer, Heidelberg (2005)
12. Ha, M.-H., Huang, S., Wang, C., Wang, X.-L.: Intuitionistic Fuzzy Support Vector Machine. Journal of Hebei University 31(3), 225–229 (2011)

The Induced Charge Test under Thunderclouds Simulation Background

Xiaoming Ren, Jun Liu, and Qin Zhou

School of Electric, Shanghai Dianji University, Shanghai 200240, China

Abstract. In order to study induced charge under simulated thunderclouds background, a thunderclouds simulation test platform had been built and a measuring instrument of induced charge had been developed. The test platform consisted of a metal disk, whose diameter is 3 meters the change of electric field can be simulated though imposing changing voltage to the metal disk. The slow changing electric field can be simulated by applying the direct-current voltage to the disk and cut it off instantaneously, while the fast changing electric field can be simulated by applying the impulse voltage to the disk. The induced charge can be measured by the measuring instrument under these two methods. The results showed that measuring instrument of induced charge can realize measurement of simulated fast changing and slow changing electric field. So the measuring instrument can be applied to the study of induced charge.

Keywords: induced charge, simulation thunderclouds, dc voltage, impulse voltage.

When the weather is fine, the earth surface is a stable weak electric field. Potential difference of objects on the earth surface can be seldom measured. Electric charge is evenly distributed according to the terrains and ground objects, which means surface charge density is almost equal in each place. When there is a thundercloud, it will generate positive vertical electric field; positive electric charge will be inducted on the earth surface. When these electric charges are out of restriction, they will cause damage to the nearby equipments. Reference[1]-[4] studied the coupling mechanism and damage of lightning induced surge on transmission lines, Reference[5] studied the damage of lightning induced surge on the oil tank .simulated lightning in laboratory is one of the most important method because lightning is random and variable. In this paper, experiment platform was established in the laboratory to develop the measurement instrument of induced charge which can measure the charge with rapid changing and slow changing electric field, and to study the induced charge.

1   Measuring Instrument of Induced Charge

1.1   Measuring Instrument of Field Strength Change

It was generally acknowledged that the majority charge distribution of thunderclouds is positive dipole, which means positive charges are in the upper part of cloud while negative charges are in the lower part, some observation showed that there is another

positive charge area under the negative charges. When a thundercloud appears, dramatic change will occur in the surface electric field. Fig.1 showed the principle of measuring instrument of field strength change. Metal induction circular disk is connected to the operational amplifier through reversed-phase input, while phase-end earthed, this allowed induction panel been maintained up to "virtual" state. Resistance R and capacitance C cross over the input and output of operational amplifier. With the external electric field E, charge is induced on the metal disk.

$$Q = \varepsilon_0 AE \tag{1}$$

A is the area of metal induction plate, $\varepsilon_0 = 8.85 * 10^{-12} F/m$ is Vacuum Dielectric Constant.

When the electric field changed, induced charge generated by the change of induced charge would flow over R and C,

$$i + \frac{U}{R} + C\frac{dU}{dt} = 0 \tag{2}$$

Due to
$$i = \frac{dQ}{dt} = \varepsilon_0 A \frac{dE}{dt} \tag{3}$$

When $\Delta t << RC$,

$$\Delta U = -\frac{\varepsilon_0 A}{C}\Delta E \tag{4}$$

So there is a linear relationship between output of the operational amplifier and electric field change while their polar is opposite.
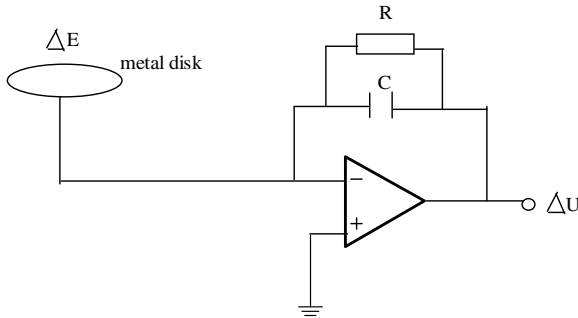
Fig. 1. The schematic of measuring instrument of field strength change

1.2   Measuring Instrument of Induced Charge

Since the induced charge of metal disk need to be measured, small changes has been made on the measuring instrument of field strength change, then it can measure the change of induced charge directly. Fig. 2 showed that R and C connected in parallel first and then connected to the reverse side of the operational amplifier, while another side of operational amplifier was grounded, so the output voltage U of operational amplifier is the capacitor voltage, induced charge can be calculated by the following formula.

$$Q = C * U \qquad (5)$$

among which $C = 1.02 * 10^{-8} F$, $R = 0.5 * 10^9 \Omega$, the output voltage of operational amplifier need to be within $\pm 2V$.
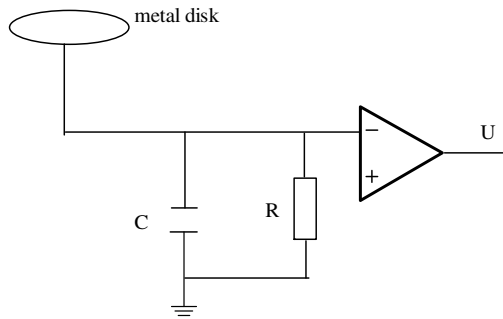
Fig. 2. The schematic of Induction charge measurement

2   Test of Induced Charge

2.1   Simulation of Thundercloud

The designed thundercloud was a mental disk with the diameter of 3m. By imposing changing voltage on the metal plate, charge can be induced on the metal disk. There are two ways of generating changing voltage, one was to impose direct-current voltage on the mental disk and cut it off suddenly, another was to impose impulse voltage on the mental disk [6-10].

2.2   Test of Induced Charge under Direct-Current Voltage cut It Off suddenly

Rectifier equipment was generally used to generate DC high voltage, and the generally used half-wave rectifier circuit was shown in the left part of Fig. 3, where R_1 was the protective resistance. h_1 was the distance between 3M diameter metal disk, and h_2 was the distance between the metal disk and the metal disk which was in the lower level. Because of the output voltage limit of the operational amplifier, the charge can be

measured by adjusting voltage and h_1. Fig. 4 showed induced charge waveform when DC voltage was cut off instantaneously. The waveform showed that the change of induced charge is relatively slow using this method. It took about 3s of current to rise to the peak value. It also indicated that when DC voltage was cut off instantaneously it can simulate slow changing electric field.
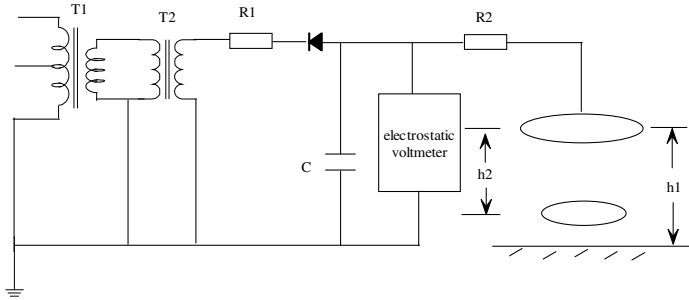
Fig. 3. Schematic of measurement of induced charge when DC voltage was cut off suddenly

Table 1. The induced charge test when DC voltage was cut off suddenly

DC voltage /kV	h_1=2m, h_2=1.7m		h_1=2.5m, h_2=2.2m	
	measure/V	Q/10^{-8} C	measure/V	Q/10^{-8} C
2.0	0.55	0.56	0.51	0.52
3.0	0.83	0.85	0.76	0.78
4.0	1.05	1.07	0.99	1.01
5.0	1.26	1.29	1.18	1.20
6.0	1.69	1.72	1.42	1.45
7.0	1.84	1.88	1.69	1.72
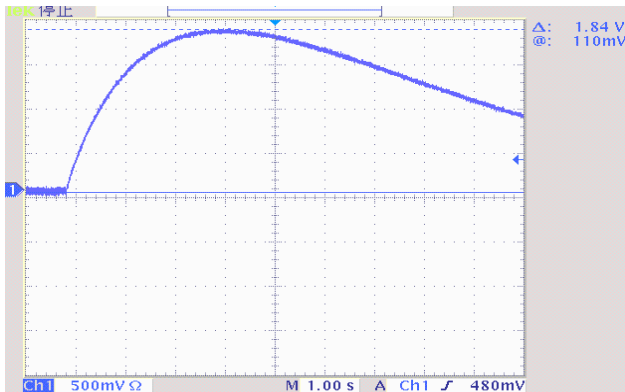
Fig. 4. Induced charge waveform when DC voltage was cut off suddenly

2.3   Test of Induced Charge under Impulse Voltage

Imposing impulse voltage on the mental disk was another way of simulating thundercloud as shown in Fig. 5.This test was to impose 1.2/50μs impulse voltage at the same height. DC voltage can be imposed at the same time when imposing impulse voltage; the results showed that DC voltage has little impact on the induced charge. Fig. 6 showed the steep waveform of induced charge under the impulse voltage. Table 2 was the induced charge test data under impulse voltage. It took only several μs to rise to the peak value, because the imposed impulse waveform is steep enough to simulate the fast changing electric field. When DC voltage was cut off, its relative slow change can simulate the slowing changing electric field.
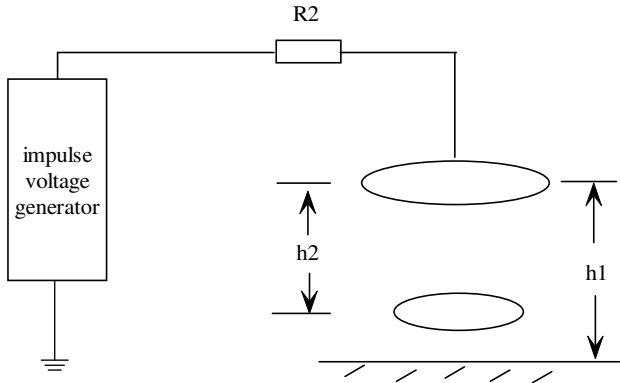
Fig. 5. Induction charge measurement schematic diagram under impulse voltage
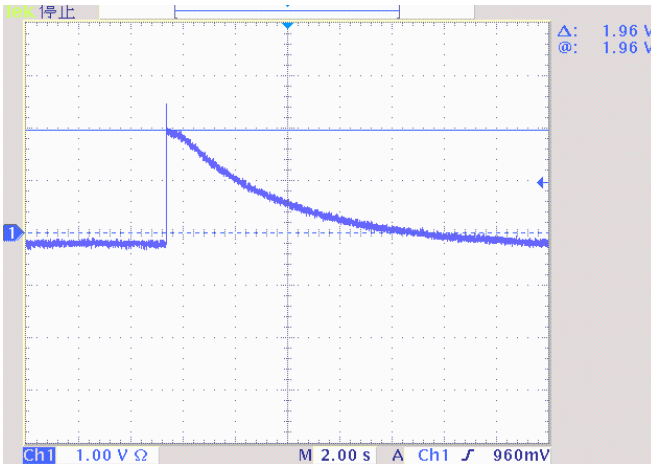
Fig. 6. Induced charge waveform under impulse voltage

Table 2. The induced charge test under impulse voltage

maximum value of impulse voltage /kV	DC voltage	measure	Q
	/kV	/V	/10^{-8}C
380	/	1.90	1.94
380	10	1.96	1.99
380	20	1.96	1.99

h_1=3.60m, h_2=3.30

3   Conclusions

When DC voltage was cut off suddenly it can simulate the slow changing electric field infected by thundercloud while impulse voltage can simulate the fast changing electric field, therefore the measurement of induced charge with simulation of thundercloud using above two ways can reflect the real situation of thunderclouds in some extent, which has provided foundation for study induced charge in the laboratory. Since the measurement range limit of the developed measuring instrument of induced charge which can only measure $2*10^{-8}$ C, the amount of voltage and height of mental disk need to be adjusted in the test.

References

1. Mo, F., Chen, Y., Ruan, J.: Analysis on coupling mechanism and calculation method of lightning induced surge on overhead transmission lines. Power System Technology 29(6), 72–77 (2005)
2. Li, L., Qi, X.: Calculation of the lightning induced voltages on power distribution line. High Voltage Engineering 37(5), 1093–1099 (2011)
3. Wang, X., Wang, S., He, J., Zang, R.: Characteristics of lightning induced overvoltage of 10 kV distribution lines. High Voltage Engineering 37(3), 599–605 (2011)
4. Wen, X., Peng, X., Xie, G.: Numerical calculation of lightning induced voltages on overhead distribution lines. Proceedings of the CSEE 18(4), 299–301 (1988)
5. Buccella, C., Feliziani, M.: A Hybrid Model to Compute the Effects of a Direct Lightning Stroke on Three-Dimensional Structures. IEEE Transactions on Magnetics 39(3), 1586–1589 (2003)
6. Silveira, F.H., De Conti, A.: Lightning overvoltage due to first strokes considering a realistic current representation. IEEE Transactions on Electromagnetic Compatibility 52(4), 929–935 (2010)

7. Ablerto, B., Fabio, N., Carlo, A.N., Mario, P.: Calculation of lightning-induced voltages on an overhead line taking into account the presence of nearby buildings. In: 7th Asia-Pacific International on Lightning, pp. 833–839 (2011)
8. Metwally, I.A., Gastli, A., Al-sheikh, M.: Comparative investigation of lightning impulse tests on electrical submersible pump motors. Electric Power Systems Research 77(8), 1065–1072 (2007)
9. Orlandi, A.: Numerical modeling of lightning induced potentials in vessels containing flammable liquids. In: Proceedings of Ninth International Symposium on High Voltage Engineering, pp. 8353.1–8353.4 (1995)
10. Yamanoto, K., Noda, T., Yokoyama, S., Ametani, A.: Experimental and analytical studies of lightning overvoltages in wind turbine generators systems. Electric Power systems Research 79(3), 436–442 (2007)

Research on Description Method of Operational Task Oriented to Operational Effectiveness Evaluation

Zenghua Li, Shen Zhang, Jingye Wang, and Shuo Liu

Department of Equipment Command and Management
The Academy of Armored Forces Engineering
Beijing, China, 100072
lzhphd08@126.com, {Zhangshen_1980,glorioushuo}@163.com,
wangjingye@yahoo.com

Abstract. Equipment of weapon is a critical factor which determines the result of information-age warfare. Operational effectiveness evaluation is an important step in equipment of weapon development and construction. In order to demonstrate the development scheme of equipment and weapon correctively and effectively, the measures of effectiveness need to be selected scientifically. Operational task is the standard factor of operational effectiveness. The concept of different types of effectiveness is given and analyzed firstly. The requirement of description of operational task is also analyzed at the same time. Then the measures of effectiveness hierarchy are put forward. Aiming to the concept and characteristic of operational effectiveness, a formalized description method is discussed. The description products of operational task are defined and specialized. Then the detailed description process is given. This is of some importance and realistic and theoretic meaning to make sure that the development of weapon and equipment is scientific and reasonable.

Keywords: operational effectiveness, mission, task, description method, effectiveness evaluation.

1 Introduction

Equipment of weapon is a critical factor which determines the result of information-age warfare. With the development of military theory and science technology, the future fight shows on the confrontment among the system of systems of equipment of weapon under the informational conditions undoubtedly. Different equipment of weapon systems constitutes a wholeness of function through information interaction to realize the strategic objective or to complete specific operational task. The operational effectiveness of each equipment is manifest by system of systems of equipment, so it is needed to aims to improve the fight capability of the whole equipment system of systems to demonstrate the construction scheme of the system of systems of equipment. The basic thought about demonstration of equipment system of systems is as follows: first, the combat conception during a future period of time is needed to build according to the development strategic and research results of military theory and

then to analyze the operational task on the basis of combat consumption. The requirement of combat capability is put up according to the requirement analyze of operational task. Finally, the evaluation of operational effectiveness of equipment system of systems is made and the advisement report about the equipment system of systems construction is submitted.

Through the evaluation of operational effectiveness of equipment system of systems, the whole combat capability under certain condition of equipment system of systems can be analyzed and the fault of the equipment system of systems construction scheme can be found, so the optimization of the equipment system of systems construction scheme can be realized. It can be concluded that operational task is the start point of equipment system of systems demonstration and is also the precondition to ensure the correctness of equipment combat effectiveness. It is necessary to use standardized mode to describe operational task in order to make sure that the equipment system of systems demonstration is correct and coherent. The most early and prosperous method of equipment requirement description is the DoDAF which realizes the effective transformation from the combat field to the capability field and to the equipment field finally through the definition of combat view, capability view and systems view. Using for reference of modeling thought of DoDAF, the description product of operational task oriented to operational effectiveness evaluation is put forward and the detail process is also given in the paper.

2   Basic Theory of Measures of Effectiveness

2.1   Conception of Effectiveness

Whether the measures of effectiveness are made certain determines the accuracy of the decision making. Selecting the measures of operational effectiveness of equipment system of systems correctively and in reason is important to provide credible decision making accordance. The basis meaning of effectiveness is the extent to which a system is used to perform some aims under certain condition. Effectiveness can be divided into three types: single effectiveness, systems effectiveness and operational effectiveness.

Single Effectiveness: it means the extent to which equipment can achieve when used for single aim. Shooting effectiveness, detecting effectiveness and command and control effectiveness of air defending equipment systems are all single effectiveness. Single effectiveness corresponding to operations of single aim, such as detection, mines laying and blocking of fire exertion and support.

Systems Effectiveness: it means the extent to which equipment system can satisfy a set of task requirements when used under certain condition. It is the synthesis evaluation about the effectiveness of equipment systems and is and important parameter when demonstrating.

Operational Effectiveness: which is also called force effectiveness, it means the extent to which equipment system can achieve when used by corresponding force to complete specific operational task under combat environments. Here, the operational tasks cover various main operational tasks the equipment systems may assumed in real combat and involving the whole combat process. So operational effectiveness is the final effectiveness of any equipment systems, and it is also the fundamental quality characteristic of any equipment systems.

2.2   The Effectiveness Measures Hierarchy

There are a number of terms used to describe system performance, while several of these terms are often used interchangeably to describe the same thing. For example, Measures of Performance (MOPs) is interchanged with Measures of Effectiveness (MOEs), the MORS＇s work recognized that there is indeed a hierarchy of effective measures. MORS identified the following key concepts: parameters, Measure(s) of Performance, Measure(s) of Effectiveness, and Measure(s) of Force Effectiveness. While the later term is not appropriate for systems in general the idea is valid and will be addressed below.

Environment

Force

System

Dimensional
Parameter

MOPs

MOEs

MOFEs
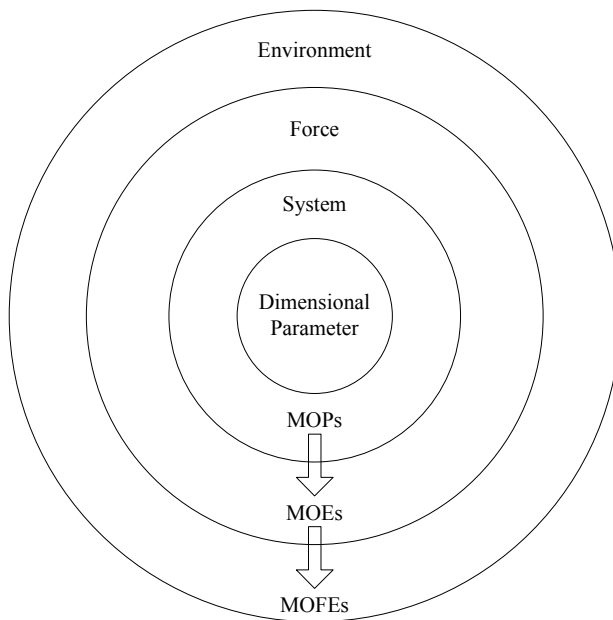
Fig. 1. The Effectiveness Measures Hierarchy

Parameters: the properties or characteristics inherent in the physical entities, whose values determine system behavior and the structure under question, even when not operating. Typical examples include signal-to-noise ratio, bandwidth, frequency, aperture dimensions, and bit error rates.

1. *Measures of Performance (MOP):* measures derived from the dimensional parameters (both physical and structural) and measure attributes of system behavior. MOPs quantify the set of selected parameters. Examples include sensor detection probability, sensor probability of false alarm, and probability of correct identification.
2. *Measures of Effectiveness (MOE):* measure of how a system performs its functions within its environment. An MOE is generally an aggregation of MOPs. Examples include survivability, probability of raid annihilation, and weapon system effectiveness.
3. *Measures of Force Effectiveness (MOFE):* measure of how a system and the force (sensors, weapons, C3 system) of which it is a part performs military missions.

3   Description Method of Operational Task

According to the existing theories, the fishing hauls have a relationship with the distribution of temperatures and vortices. When warm currents meet cold currents, warm currents lift the nutrients in cold currents up to the shallow part of sea, which makes a favorable environment for the reproduction of plank-ton, and thus the plankton-eating fish will be attracted. Besides, many kinds of fish move with currents, hence the fish density may be relatively higher at the confluence of warm currents and cold currents. Vortices also help to gather nutrients and fish. To find the relationship between these factors, some visualization methods need to be applied to our system. In this paper, we propose a system which is composed of overall and detailed visualization tools in order to verify the relationship between these factors. As for the visualization tools, we develop a parallel coordinates plot to show the overview of the high-dimensional ocean data, a streamline tool to show sea currents and a volume rendering tool to show the distribution of temperatures.

3.1   The Basic Principle of Operational Task Description

Generally speaking, in the military analyses, the judgments of the whole combat effect and the rationality of combat scheme on combat are mainly are mainly made according to the completion of every task. The task can be divided into three levels. Showed as figure 2, task can be composed by both subtask and action. Action is the basis of behavior modeling, and task is the middle section of behavior modeling, and mission is the macroscopical embodiment of behavior.
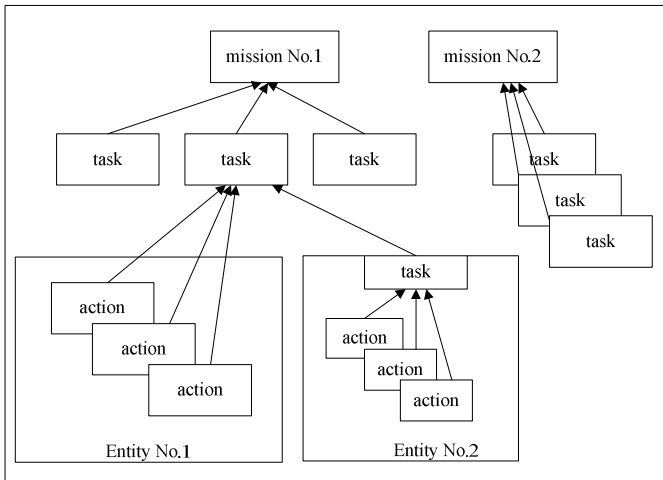
Fig. 2. Rlationship Among Mission, Task And Action

1. *granularity of mission decomposition*: in the process of mission decomposition, it needs not to decompose when the requirement of the subtask is related to the capability vector of combat resource. Thus the subtask is the meta-task.
2. *principle of targets*: in the process of mission decomposition, the subtasks are commonly decomposed according to the attack or the occupation of the targets. The realization of all targets must be one of the meta-task.
3. *phase division*: combat phase is thought about the combat process and action series. In the process of mission decomposition, it needs to take account that whether the meta-task can be obtained by the combat phase division during which the targets are realized.
4. *area and direction of combat*: the principle of area and direction of combat is thought about the route selection, selection of advancement direction and selection of combat area in the subtask realized process by the combat force.
5. *completeness principle*: any combat task can be expressed by a meta-task, so the meta-task must be complete. That is to say all the operational tasks can be expressed by meta-task. If there is one or more task that cannot be expressed by meta-task, the meta-task needs to be modified.

3.2 Products of Operational Task Description

Operational task is the objective needed to be achieved and the responsibility to be realized by the force during the combat process. The analyses of operational task are endowed with special mission in the demonstration of equipment requirement. It not only needs to describe the operational task serial and their relationship to achieve the mission, but also form the basic to realize the transformation from the combat field to the capability field and equipment field. That is to say in order to describe the task,

which combat nodes are needed and which actions are needed to take and which information are required. The questions need to be answered including:

1. What is the combat mission and how is the combat process?
2. Which operational tasks are required in order to complete the mission? How is the order among these tasks? Which information interchanges are taken on?
3. Which combat nodes are required in order to execute the task? How is the relationship among them? What are the information requirements?
4. Which combat actions needs to be taken in order to compete the task? What are the information requirements?
5. Aiming at the four questions stated above, the model of operational task is divided into eight sub-models whose basic forms are as Table 1.

Table 1. Products of Operational Task Description

Product	Name	Description
TV-1	High combat concept view	Describing the high combat process and making sure the combat mission
TV-2	Decomposition tree of operational task(activities)	Describing operational tasks and activities and their hierarchy structure
TV-3	Relationship models of operational task information.	Describing information requirements among operational tasks
TV-4	Time schedule view of operational tasks	Describing time schedule relationships among operational tasks
TV-5	Assignment matrix between operational tasks and combat nodes	Describing assignment relationships between operational tasks and combat nodes
TV-6	Decomposition tree of combat nodes	Describing the combat nodes and their hierarchy structure
TV-7	Information requirement models of combat nodes	Describing information requirement among combat nodes
TV-8	Relationship view of combat nodes	Describing command and control relations among combat nodes

3.3   The Description Process of Operational Task

The basic flow of operational task description is shown as figure 3. The first step is to analyze the combat field to fix on the higher operational mission. The next step is to analyze the executing process of operational mission to make certain the operational task order and their time and information relationships to complete the operational mission. Then analyze the combat nodes and their relationships of command and control and information requirements to complete the operational tasks. The last step is to analyze the actions the combat nodes will take and their relationships in order to complete the operational tasks. It is noted that the process of operational task description is not one straight step but a recycle process. The top level information is detailed to obtain the bottom level information and the top level information is completed continuously.
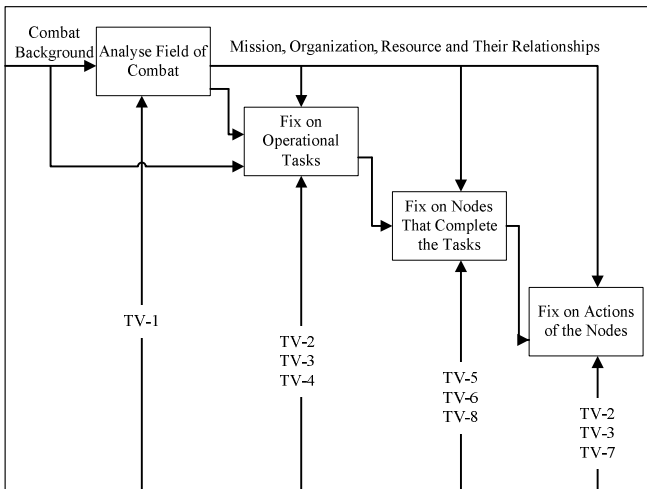
Fig. 3. The Flow Char of Operational Task Description

Analyze Operational Concepts: Operational concept is description about what and how is the combat. It characterizes the future warfare and its requirements wholly. The description of operational concepts is the start points of operational task analyze. It mainly describes how to use the resources to finish the operational tasks. Operational concepts are described by high combat concept view TV-1. TV-1 ascertains the combat mission, organization, resources and the interchanging information among them through analyzing combat background.

Confirm the Operational Task

Discompose the Operational Mission: Operational missions are macroscopical, general and fuzzy. Thus it is need to be discomposed to a set of operational task. An ordered set of combat activities is corresponding to a specific operational task and can be seen as a solution of the operational task. So the combat activities can be

confirmed through the analyzing of the operational task executed process. If information interchanging is emphasized, the model TV-3 of combat activity information relationship can be adopted. If logic time order is emphasized, the model TV-4 can be adopted. Finally, pick up the relative combat activities and build the decomposition tree of combat activity as TV-3.

Discompose the Combat Activity: The relationships between operational tasks include order relationships and information relationships. Order relationships can be described by TV-4. Information relationships are the in/out information flow of operational task and can be described by TV-3.

Confirm the Condition, Standard and Measures of   Operational Tasks: The comprehensive description of operational tasks not only includes basic information, order and information relationship, but also includes the initial conditions, final conditions, and the index adopted as measurement and evaluation standards.

Confirm the Combat Nodes Executing the Operational Tasks

Confirm the Combat Nodes: Combat nodes are entities that generate, consume or deal with data in the process of executing operational tasks. The confirmation of combat nodes can be made from the resource and organization in the high level combat view TV-1. Finally, the allocation matrix TV-5 of operational task and combat node can be generated according to the relationship between operational task and combat nodes.

Discompose the Combat Nodes: The layer characteristic of operational task determines that combat nodes also have the same characteristic. The decomposition of operational task corresponds to the decomposition of combat nodes. The decomposition of combat nodes cannot be made until it fulfills the requirement of combat. The decomposition of combat nodes can be described by decomposition tree of combat nodes TV-6.

Confirm the Relationships Between Combat Nodes: Relationships of command, control, cooperation and support are needed to build in order to finish the operational task. It can be described by relation view of TV-8.

Confirm the Activities Taken by the Nodes

Confirm the Combat Activities: Combat activities are behavior or action taken by combat nodes to complete the operational task. Combat activities are detail operational tasks. An ordered set of operational task is corresponding to a specific combat mission and can be seen as a solution of the combat mission. So the decomposition of the mission can be done through the analyzing of the combat process. If information interchanging is emphasized, the model TV-3 of operational task information relationship can be adopted. If logic time order is emphasized, the model TV-4 can be adopted. Finally, pick up the relative operational task and build the decomposition tree of operational task as TV-2.

Discompose the Combat Nodes: The layer characteristic of operational task and combat node determine that combat nodes also have the same characteristic. The decomposition of combat activity corresponds to the decomposition of combat nodes.

The decomposition of combat activities cannot be made until it fulfills the requirement of combat. The decomposition of combat activities can be described by decomposition tree of combat activities TV-2.

Confirm Requirment of Combat Information: Requirement of combat Information mainly describes which information is needed between combat nodes in order to support the complete combat activities. It can be described by combat information requirement model TV-7. Information requirement of combat nodes come from the information interchange between combat activities. The in/out information flow can be described by combat activity information model TV-3.

4   Conclusions

Using the modeling thought of DoDAF for reference, the description method for operational task oriented to operational effectiveness evaluation. The concept of effectiveness is given, meanwhile, different measures of effectiveness are put forward. The requirement of operational task description is demonstrated. The main part of the paper is about the description product of operational task and detailed describing steps are researched. Canonical description method of operational task is main guarantee of scientific and reasonable operational effective evaluation. This is of some importance and realistic and theoretic meaning to make sure that the development of weapon and equipment is scientific and reasonable.

References

1. DOD Architecture Framework Working Group. DoDArchitecture Framework Version2.0 (2009)
2. Davis, P.K.: Exploratory Analysis Enabled by Multiresolution, Multiperspective Modeling. In: Proceedings of the 2000 Winter Simulation Conference, pp. 299–301. IEEE, USA (2000)
3. Cares, J.: Distributed NetWorkOperations. Alidade Press (2005); Zhu, S., Hu, X.: Research on Quantitative Analysis Method of Networked Effectiveness of Weapon and Equipment System of System. In: Modeling and Simulation of System of System Combat Capability Based on Information System (June 2010)
4. Wang, J.: System Theory and Combat Complexity. Yellow River Publishing Company, Jinan (2008)
5. Zhang, B.: General Station on Requirment of Equipment and Weapon of the Army. People Libaration Army Publishing Company (2005)
6. Sage, A.P., Cuppan, C.D.: On the Systems Engineering and Management of Systems of Systems and Federations of Systems. Information, Knowledge, Systems Management 2(4), 325–345 (2001)
7. Tourwe, T., Mens, K.: Mining Aspectual Views using Formal Concept Analysis. In: Proceedings of the 4th International Workshop on Source Code Analysis and Manipulation, pp. 97–106. IEEE Computer Society (2004)
8. Horne, G.E., Meyer, T.E.: Data Farming: Discovering Surprise. In: Proceeding of the 2004 Winter Simulation Conference, pp. 807–813 (2004)

Moment Exponential Stability of Neutral Impulsive Nonlinear Stochastic Delay Partial Differential Equations

Lei Zhang[1,3], Yongsheng Ding[1,3,⋆], Tong Wang[1,3], Liangjian Hu[2], and Kuangrong Hao[1,3]

[1] College of Information Sciences and Technology, Donghua University,
Shanghai, 201620, P.R. China
[2] Department of Mathematics, Donghua University,
Shanghai, 201620, P.R. China
[3] Engineering Research Center of Digitized Textile & Fashion Technology,
Ministry of Education, Donghua University,
Shanghai, 201620, P.R. China
`ysding@dhu.edu.cn`

Abstract. In this paper, we consider the stability in pth moment of mild solutions to nonlinear impulsive stochastic delay partial differential equations (NISDPDEs). By employing a fixed point approach, sufficient conditions for the exponential stability in pth moment of mild solutions are derived.

Keywords: pth moment exponential stability; impulsive stochastic equations, neutral equations.

1 Introduction

In recent years, the stochastic partial differential equations (SPDEs) have been considered in [1-5], which discuss the stability of the strong solutions and mild solutions. More recently, the asymptotic stability of mild solutions of SPDEs with delays is studied by using fixed point approach which shows that some of these difficulties are rectified when applying the fixed point theory [6].

However, in real world, many dynamical systems not only depend on present and past states, but also involve derivatives with delays, and, neutral SPDEs with delays are often used to describe such systems. But there is only a few work on the stability of mild solutions to neutral SPDEs with delays, one of the reasons is that the neutral term is present and the mild solutions do not have stochastic differential, so Itô formula fails to deal with this problem. Recently, fixed point theory is applied to overcome the difficulties on the stability of deterministic differential equations, which give the points that some of these difficulties are rectified [7-10]. Subsequently, this valuable method is applied to deal with the asymptotic stability in mean square of neutral SPDEs with infinite delays [11].

⋆ Corresponding author.

Nowadays, impulsive differential equations have attracted many researchers' attention due to their wide applications such as electrical engineering, control technology, drug administration [12-16]. However, it is well known that there exist stochastic failures and repairs of the components, changes in the interconnections of subsystems, sudden environment changes, etc. The stability estimation of stochastic differential equations with delays has been discussed by several authors [17-24].

The rest of this paper is organized as follows. In Section 2, we briefly present some basic notations and preliminaries. Section 3 is devoted to the proof of the stability of mild solution by fixed point principle. An example is provided to illustrate the effectiveness of the proposed result in Section 4.

2   Preliminaries and Problem Formulation

Let $\{\Omega, \mathcal{F}, \mathbb{P}\}$ be a complete probability space equipped with some filtration $\{\mathcal{F}_t\}_{t \geq 0}$ satisfying the usual conditions, i.e., the filtration is right continuous and $\mathcal{F}_0$ contains all $\mathbb{P}$-null sets. Let H, K be two real separable Hilbert spaces and denote their inner products by $\langle \cdot, \cdot \rangle_H$, $\langle \cdot, \cdot \rangle_K$.

In this paper, we consider the following impulsive neutral stochastic delay partial differential equations

$$d[x(t) + Cx(t - \tau)] = [Ax(t) + f(x(t - \tau))]dt$$
$$+ g(x(t - \tau))dW(t), \qquad t \geq 0, t \neq t_k \tag{1}$$

$$\Delta x(t_k) = x(t_k^+) - x(t_k^-) = I_k(x(t_k^-)), \qquad t = t_k, k = 1, 2, \cdots, m \tag{2}$$

$$x_0(\cdot) = \varphi \in D_{\mathcal{F}_0}^b([-\tau, 0], H) \tag{3}$$

where $D := D_{\mathcal{F}_0}^b([-\tau, 0], H)$ is the space of all $\mathcal{F}_0$-measurable càdlàg functions from $[-\tau, 0]$ to H, equipped with the norm $|\varphi|_D = \sup\limits_{t \in [-\tau, 0]} |\varphi(t)|_H$

$$C : D([-\tau, 0]; H) \to H, \quad f : D([-\tau, 0]; H) \to H, \quad g : D([-\tau, 0]; H) \to \mathrm{L}(K, H)$$

are all given Borel measurable functions. A is the infinitesimal generator of a semigroup of bounded linear operators $S(t), t \geq 0$ in H, $I_k : H \to H$. Furthermore, the fixed moments of time t_k satisfies $0 < t_1 < \cdots < t_m < \lim\limits_{k \to \infty} t_k = \infty$, $x(t_k^+)$ and $x(t_k^-)$ represent the right and left limits of $x(t)$ at $t = t_k$, respectively. Also, $\Delta x(t_k) = x(t_k^+) - x(t_k^-)$ represents the jump in the state x at time t_k with I_k determining the size of the jump.

Definition 1. *A stochastic process $\{x(t), t \in [0, T], 0 \leq T < \infty\}$, is called the mild solution to (1) if*

(a) *$x(t)$ is adapted to $\mathcal{F}_t, t \geq 0$,*
(b) *$x(t)$ has a càdlàg path on $t \in [0, T]$ almost surely, and $x(t)$ satisfying the following integral equation.*

$$x(t) = S(t)(\varphi(0) + C(0, \varphi)) - C(x(t - \tau)) - \int_0^t AS(t - s)C(x(s - \tau))ds$$

$$+ \int_0^t S(t - s)f(x(s - \tau))ds + \int_0^t S(t - s)g(x(s - \tau))dW(s) \qquad (4)$$

$$+ \sum_{0 < t_k < t} S(t - t_k)I_k(x(t_k^-))$$

and

$$x_0 = \varphi \in D_{\mathcal{F}_0}^b([-\tau, 0], H)$$

Definition 2. *The mild solution $x(t)$ of (1) is said to be exponential stable in pth moment if there exists a constant $M > 0$ and $\gamma > 0$ such that*

$$E|x(t)|_H^p < ME|\varphi|_D^p e^{-\gamma t}, \qquad t \geq 0$$

We impose some assumptions below to obtain our main result.

(I) A is the infinitesimal generator of a semigroup of bounded linear operators $\{S(t), t \geq 0\}$ in H satisfying $|S(t)|_H \leq e^{-\eta t}, t \geq 0$, for some constant $\eta > 0$,
(II) For any $x, y \in H, t \geq 0$, there exist some constants $\alpha \in [0, 1]$ and $K_D > 0$ such that $|(-A)^{-\alpha}C(t, x) - (-A)^{-\alpha}C(t, y)|_H^p \leq K_C|x - y|_H^p$
(III) The coefficient f and g satisfy the Lipschitz condition and there exists a constant L such that $|f(t, x) - f(t, y)|_H \vee |g(t, x) - g(t, y)|_H \leq L|x - y|_H$ for all $x, y \in H, t \geq 0$. We further assume that $C(t, 0) = f(t, 0) = g(t, 0) = 0$ for all $t \geq 0$.
(IV) $I_k \in C(H, H)$ and there exists a constant q_k such that $|I_k(x) - I_k(y)|_H \leq q_k|x - y|_H$.

The following lemma is needed to consider our results.

Lemma 1. *Let the assumption (I) hold. Then for any $0 < \beta \leq 1$, the following equality holds:*

$$S(t)(-A)^\beta x = (-A)^\beta S(t)x, \qquad x \in D((-A)^\beta),$$

and there exists a positive constant M_β such that for any $t > 0$,

$$|(-A)^\beta S(t)| \leq M_\beta t^{-\beta} e^{-\eta t}.$$

3   The Main Results

In this section, we concern the existence and exponential stability of mild solution to (1). Our main method is the Banach fixed point principle.

Theorem 3. *Suppose the assumptions (I)-(IV) hold, and we further assume that*

$$5^{p-1}\left[K_C|(-A)^{-\alpha}|^p + \frac{M_{1-\alpha}^p K_C \Gamma^2(\alpha)}{((p-1)\gamma)^{p\alpha}} + \frac{L^p}{(p-1)\gamma^2} + \frac{L^p}{p\gamma} + \hat{L}\right] < 1$$

where $\Gamma(\cdot)$ is the Gamma function, $M_{1-\alpha}$ is the corresponding constant in Lemma 1. If the initial value $\phi(t)$ satisfies

$$E|\phi(t)|^p \le M_0 E|\phi(0)|_H^p e^{-\mu t}, \qquad t \in [-\tau, 0]$$

for some $M_0 > 0, \mu > 0$, then the mild solution to (1) exists uniquely and is exponentially stable in pth moment.

Proof. Define nonlinear operator $\Psi : H \to H$ by $\Psi(x)(t) = \psi(t)$ for $t \in [-\tau, 0]$ with a norm $\|x\|_H := \sup_{t \ge 0} E|x(t)|_H^p$ such that there exist some constants $M^* > 0$ and $\eta > 0$ satisfying

$$E|x(t)|_H^p < M^* E|\phi|_D^p e^{-\eta t}, \qquad t \ge 0$$

and for $t \ge 0$,

$$\begin{aligned}
\Psi(x)(t) &= S(t)(\phi(0) + C(0, \phi)) - C(t, x(t - \tau)) \\
&\quad - \int_0^t AS(t-s)C(s, x(s-\tau))ds + \int_0^t S(t-s)f(s, x(s-\tau))ds \\
&\quad + \int_0^t S(t-s)g(s, x(s-\tau))dw(s) + \sum_{0 < t_k < t} S(t-t_k)I_k(x(t_k^-)) \qquad (5) \\
&:= \sum_{i=1}^6 F_i(t)
\end{aligned}$$

To prove the pth moment stability, it is enough to show that the operator ψ has a fixed point in H. In order to get this result, we use the contraction mapping principle. First, we verify the pth moment continuity of ψ on $[0, \infty)$.

Let $x \in H, t_1 \ge 0$ and $|r|$ be sufficiently small then

$$E|\psi(x)(t_1 + r) - \psi(x)(t_1)|_H^p \le 6^{p-1} \sum_{i=1}^6 E|F_i(t_1 + r) - F_i(t_1)|_H^p$$

It can be easily obtained that $E|F_i(t_1 + r) - F_i(t_1)|_H^p \to 0, i = 1, 2, 3, 4, 6$, as $r \to 0$. Moreover by using Holders inequality, we obtain

$$\begin{aligned}
&E|F_5(t_1 + r) - F_5(t_1)|_H^p \\
&\le 2^{p-1}c_p \big[\int_0^{t_1} (E|(S(t_1 + r - s) - S(t_1 - s))g(s, x(s-\tau))dw(s)|_H^p)^{2/p}ds\big]^{p/2} \\
&+ 2^{p-1}c_p \big[\int_0^{t_1+r} (E|S(t_1 + r - s)g(s, x(s-\tau))|_H^p)^{2/p}ds\big]^{p/2} \to 0
\end{aligned}$$

$$(6)$$

As $r \to 0$, where $c_p = (p(p-1)/2)^{p/2}$. Thus ψ is continuous in pth moment on $[0, \infty)$. Next we show that $\psi(H) \subset H$. From (5), we obtain

$$
\begin{aligned}
E|\psi x(t)|_H^p &\leq 6^{p-1} E|S(t)(\phi(0) + C(0, \phi)|_H^p + 6^{p-1} E|C(t, x(t-\tau))|_H^p \\
&\quad + 6^{p-1} E|\int_0^t AS(t-s)C(s, x(s-\tau))ds|_H^p \\
&\quad + 6^{p-1} E|\int_0^t S(t-s)f(s, x(s-\tau))ds|_H^p \\
&\quad + 6^{p-1} E|\int_0^t S(t-s)g(s, x(s-\tau))dw(s)|_H^p \\
&\quad + 6^{p-1} \sum_{0 < t_k < t} E|S(t-t_k)I_k(x(t_k^-))|_H^p
\end{aligned}
\tag{7}
$$

Now we estimate the terms on the (7). Using (II), we get

$$
\begin{aligned}
E|C(t, x(t-\tau))|_H^p &\leq K_C|(-A)^{-\alpha}|^p E|x(t-\tau)|_H^p \\
&\leq K_C|(-A)^{-\alpha}|^p (M^* e^{\eta r} E|\phi|_D^p e^{-\eta t} + M_0 e^{\mu r} E|\phi(0)|_D^p e^{-\mu t})
\end{aligned}
\tag{8}
$$

By using Holder inequality and assumption (II), we obtain

$$
E|\int_0^t AS(t-s)C(s, x(s-\tau))ds|_H^p \leq E(\int_0^t |AS(t-s)C(s, x(s-\tau))|_H ds)^p
$$

$$
\leq M_{1-\alpha}^p E\left[\int_0^t (t-s)^{\alpha-1} e^{-\gamma(t-s)}|(-A)^\alpha C(s, x(s-\tau))|_H ds\right]^p
$$

$$
\leq M_{1-\alpha}^p K_C \gamma^{-\alpha} \Gamma^2(\alpha)[\frac{M^*}{((p-1)\gamma - \eta)^{(p-1)\alpha}} e^{\eta r} E|\phi|_D^p e^{-\eta t}
$$

$$
+ \frac{M_0}{((p-1)\gamma - \mu)^{(p-1)\alpha}} e^{\mu r} E|\phi|_D^p e^{-\mu t}]
\tag{9}
$$

Similarly, we obtain

$$
E|\int_0^t S(t-s)f(s, x(s-\tau))ds|_H^p \leq E\left[\int_0^t e^{-p\gamma(t-s)}|f(s, x(s-\tau))|_H ds\right]^p
$$

$$
\leq \frac{L^p}{\gamma}\left[\frac{M^*}{(p-1)\gamma - \eta} e^{\eta r} E||\phi|_D^p e^{-\eta t} + \frac{M_0}{(p-1)\gamma - \mu} e^{\mu r} E|\phi|_D^p e^{-\mu t}\right]
\tag{10}
$$

Taking into account Burkholder inequality, we have

$$
E|\int_0^t S(t-s)g(s, x(s-\tau))dw(s)|_H^p \leq \int_0^t E|S(t-s)g(s, x(s-\tau))|_H^p ds
$$

$$
\leq L^p\left[\frac{M^*}{p\gamma - \eta} e^{\eta r} E|\phi|_D^p e^{-\eta t} + \frac{M_0}{p\gamma - \mu} e^{\mu r} E|\phi|_D^p e^{-\mu t}\right]
\tag{11}
$$

and

$$\sum_{0 < t_k < t} E|S(t - t_k)I_k(x(t_k^-))|_H^p \le e^{-p\eta t} E|I_k(x(t_k^-))|_H^p$$

$$\le e^{-p\eta t}(\sum_{k=1}^{m} |q_k|_H^p)E|x(t_k^-)|_H^p \tag{12}$$

recalling (7), from (8) to (12), one can see that there exists a $M_1 > 0$ and $\eta_1 > 0$ such that

$$E|(\psi x)(t)|_H^p \le M_1 E|\phi|_D^p e^{-\eta_1 t}$$

Thus, we conclude that $\psi(H) \subset H$.

Finally, we prove that ψ is a contraction mapping. To get this, let $x, y \in H$, so for $s \in [0, T]$, we have

$$\sup_{s \in [0,T]} E|(\psi x)(t) - (\psi y)(t)|_H^p$$

$$\le 5^{p-1} \sup_{s \in [0,T]} E|C(t, x(t - \tau)) - C(t, y(t - \tau))|_H^p$$

$$+ 5^{p-1} \sup_{s \in [0,T]} E|\int_0^t S(t - s)[C(t, x(t - \tau)) - C(t, y(t - \tau))]ds|_H^p$$

$$+ 5^{p-1} \sup_{s \in [0,T]} E|\int_0^t S(t - s)[f(s, x(s - \tau)) - f(s, y(s - \tau))]ds|_H^p$$

$$+ 5^{p-1} \sup_{s \in [0,T]} E|\int_0^t S(t - s)[g(s, x(s - \tau)) - g(s, y(s - \tau))]dw(s)|_H^p$$

$$+ 5^{p-1} \sup_{s \in [0,T]} E|\sum_{0 < t_k < t} S(t - t_k)(I_k(x(t_k^-)) - I_k(y(t_k^-)))|_H^p$$

$$\le 5^{p-1}[K_C|(-A)^{-\alpha}|^p + \frac{M_{1-\alpha}^p K_C \Gamma^2(\alpha)}{((p-1)\gamma)^{p\alpha}}$$

$$+ \frac{L^p}{(p-1)\gamma^2} + \frac{L^p}{p\gamma} + \hat{L}](\sup_{s \in [0,T]} E|x(t) - y(t)|_H^p)$$

where $\hat{L} = e^{-\eta p T} E(\sum_{k=1}^{m} |q_k|_H^p)$. Therefore, ψ is a contraction mapping and hence there exists a unique fixed point $x(\cdot)$ in H which is a solution of (1) with $x(s) = \phi(s), s \in [-\tau, 0]$ and $x(t)$ is exponentially stable in pth moment stable.

4   Example

At last, an example is provided to illustrate our results. Considering the following neutral SPDE with delays of the form:

$$d[u(x,t) + \frac{1}{M_{\frac{1}{2}}|(-A)^{-\frac{1}{2}}|}u(x, t - \tau)] = [\frac{\partial^2}{\partial x^2}u(x,t) + au(x, t - \tau)]dt$$

$$+ bu(x, t - \tau)dw(t), \qquad t \ge 0 \tag{13}$$

$$\Delta u(x, t_k) = u(x, t_k^+) - u(x, t_k^-) = I_k(u(x, t_k^-)) \qquad t = t_k, k = 1, 2, \ldots$$

with initial and boundary conditions as follows:

$$u(x, s, \cdot) = \phi(x, s, \cdot) \in L^2[0, \pi], \qquad u(x, t, 0) = u(x, t, \pi) = 0, \qquad s \leq 0$$

Let $H = L^2(0, \pi)$, define $A : H \to H$ by $A = \frac{\partial^2}{\partial x^2}$ with domain
$D(A) = \{u \in H : u, u' \text{ are absolutely continuous, } u'' \in H \text{ and } u(x, 0) = u(x, \pi) = 0\}$.

It is well known that A is the infinitesimal generator of an analytic semi-group $S(t), t \geq 0$, in H and $S(t)u = \sum_{n=1}^{\infty} e^{-n^2 t} \langle u, e_n \rangle_H e_n, u \in H$. where $e_n(\xi) = \sqrt{\frac{2}{n}} \sin n\xi, n = 1, 2, \cdots$ is the set of eigenvector of $-A$. Moreover, $|S(t)| \leq e^{-\pi^2 t}, t \geq 0$. It is obviously that all the assumptions are satisfied with

$$\gamma = \pi^2, \qquad K_C = \frac{1}{M_{\frac{1}{2}}^2}, \qquad L = \max\{a, b\}$$

from the definition of $(-A)^{-\frac{1}{2}}$ we have $|(-A)^{-\frac{1}{2}}| \leq \frac{1}{\pi}$. Thus by Theorem 3, if $E|\phi(s)|^p \leq M_0 E|\phi(0)|^p e^{-\mu s}$ for $s \in [-\tau, 0]$, where $M_0 > 0, 0 < \mu < 1$, the mild solution of (13) is exponential stable in pth moment and almost surely provided by

$$5^{p-1} \left[\frac{1}{M_{\frac{1}{2}}^2} + \frac{M_{\frac{1}{2}}^{p-2}}{(p-1)^{p/2} \pi^{p-1}} + \frac{L^p}{(p-1)\pi^4} + \frac{L^p}{p\pi^2} + \hat{L} \right] < 1$$

If we set $a = 0.01, b = 0.02, M_{\frac{1}{2}}|(-A)^{-\frac{1}{2}}| = 6/\pi$, and let $\tau = 0.36, t_1 = 1, t_k = t_{k-1} + k, k = 2, 3, \cdots$, the simulation results is illustrated in Figs. 1-4, which show that the equilibrium point is globally exponentially stable.
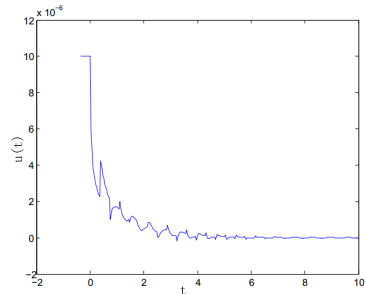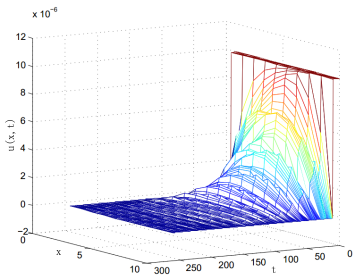
Fig. 1. The 3-D trajectory $u(x, t)$ of Eq. (13) without impulse

Fig. 2. The 2-D trajectory $u(x, t)$ of Eq. (13) without impulse
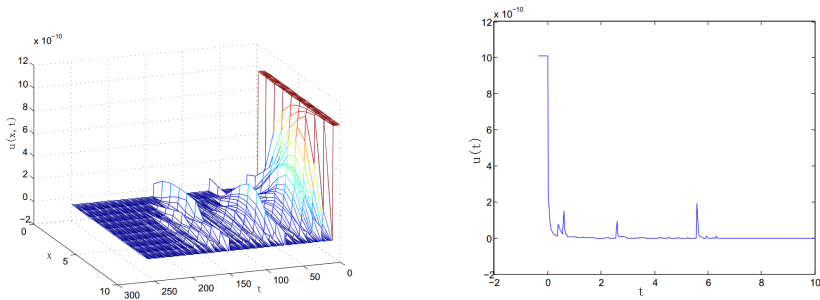
Fig. 3. The 3-D trajectory $u(x,t)$ of Eq. (13) with impulse

Fig. 4. The 2-D trajectory $u(x,t)$ of Eq. (13) with impulse

References

1. Liu, K.: Stability of Infinite Dimensional Stochastic Differential Equations with Applications. Chapman and Hall, CRC, London (2006)
2. Luo, J., Liu, K.: Stability of infinite dimensional stochastic evolution equations with memory and Markovian jumps. Stochastic Process. Appl. 118, 864–895 (2008)
3. Da Prato, G., Zabczyk, J.: Stochastic Equations in Infinite Dimensions. Cambridge University Press (1992)
4. Taniguchi, T.: The existence and asymptotic behavior of mild solutions to stochastic evolution equations with infinite delays driven by Poisson jumps. Stoch. Dyn. 9, 217–229 (2009)
5. Wan, L., Duan, J.: Exponential stability of non-autonomous stochastic partial differential equations with finite memory. Statist. Probab. Lett. 78, 490–498 (2008)
6. Luo, J.: Fixed points and exponential stability of mild solutions of stochastic partial differential equations with delays. J. Math. Anal. Appl. 342, 753–760 (2008)
7. Burton, T.A.: Stability by fixed point theory or Lyapunov theory: a comparison. Fixed Point Theory 4, 15–32 (2003)
8. Burton, T.A.: Fixed points, stability and exact linearization. Nonlinear Anal. 61, 857–870 (2005)
9. Burton, T.A., Furumochi, T.: Asymptotic behavior of solutions of functional differential equations by fixed point theorems. Dynam. Systems Appl. 11, 499–521 (2002)
10. Burton, T.A., Zhang, B.: Fixed points and stability of an integral equation: nonuniqueness. Appl. Math. Lett. 17, 839–846 (2004)
11. Luo, J., Taniguchi, T.: Fixed points and stability of stochastic neutral partial differential equations with infinite delay. Stoch. Anal. Appl. 27(6), 1163–1173 (2009)

12. Wu, S.L., Li, K.L., Zhang, J.S.: Exponential stability of discrete-time neural networks with delay and impulses. Appl. Math. Comput. 218(12), 6972–6986 (2012)
13. Zhang, X.M., Huang, X.Y., Liu, Z.H.: The existence and uniqueness of mild solutions for impulsive fractional equations with nonlocal conditions and infinite delay. Nonlinear Analysis: Hybrid Sys. 4(4), 775–781 (2010)
14. Cui, J., Yan, L.T.: Existence results for impulsive neutral second-order stochastic evolution equations with nonlocal conditions. Math. Comput. Model. (in press, 2012)
15. Lin, A.H., Ren, Y., Xia, N.M.: On neutral impulsive stochastic integro-differential equations with infinite delays via fractional operators. Math. Comput. Model. 51, 413–424 (2010)
16. Tai, Z.X., Lun, S.X.: On controllability of fractional impulsive neutral infinite delay evolution integrodifferential systems in Banach spaces. Appl. Math. Lett. 25(2), 104–110 (2012)
17. Liu, K.: Lyapunov functionals and asymptotic stability of stochastic delay evolution equations. Stochastics 63, 1–26 (1998)
18. Taniguchi, T.: The exponential stability for stochastic delay partial differential equations. J. Math. Anal. Appl. 331, 191–205 (2007)
19. Hou, Z., Bao, J., Yuan, C.: Exponential stability of energy solutions to stochastic partial differential equations with variable delay and jumps. J. Math. Anal. Appl. 366, 44–54 (2010)
20. Liu, K., Truman, A.: A note on almost sure exponential stability for stochastic partial functional differential equations. Statist. Probab. Lett. 50(3), 273–278 (2000)
21. Luo, J.: Fixed points and exponential stability of mild solutions of stochastic partial differential equations with delays. J. Math. Anal. Appl. 342, 753–760 (2008)
22. Luo, J., Liu, K.: Stability of infinite dimensional stochastic evolution equations with memory and Markovian jumps. Stochastic Process. Appl. 118, 864–895 (2008)
23. Samidurai, R., Anthoni, S.M., Balachandran, K.: Global exponential stability of neutral-type impulsive neural networks with discrete and distributed delays. Nonlinear Analysis: Hybrid Sys. 4(4), 103–112 (2010)
24. Sakthivel, R., Luo, J.: Asymptotic stability of impulsive stochastic partial differential equations with infinite delays. J. Math. Anal. Appl. 356, 1–6 (2009)
25. Taniguchi, T.: Asymptotic stability theorems of semilinear stochastic evolution equations in Hilbert spaces. Stochastics 53, 41–52 (1995)

Real-Time Visualization for Large Scale Terrain Based on Linear Quadtree

Qin Li[1,2], Xiang Wang[1,3], Yongjia Zhao[1,3], and Shuling Dai[1,3]

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University,
Beijing 100191, China
[2] Large Aircraft Advanced Training Center, Beihang University, Beijing 100191, China
[3] School of Automation Science and Electrical Engineering, Beihang University,
Beijing 100191, China
{liqin7801850,wangxiang8602,activezyj}@126.com, sldai@yeah.net

Abstract. To solve the problem of low frame rate and high memory space cost in large scale terrain visualization, a real-time rendering algorithm based on linear quadtree is proposed in this paper. It consists of three parts. Firstly, digital elevation models (DEMs) are de-sampled to get the same size blocks, terrain pyramid model is built with different scales and biases factors offline. Then linear quadtree index of terrain blocks is created for fast query, a more reasonable scheduling scheme is defined, and every frame only allowed loading one terrain block in order to avoid unstable rendering quality. Finally, GPU-based morphing method is applied to slow down vertex popping; vertical skirt is used to eliminate crack. Compared with Geomipmapping algorithm, the experimental results show the proposed algorithm could drop down memory space, GPU processing time and GPU rendering batches significantly, and the rendering rate is high and smooth.

Keywords: Terrain Rendering, Dynamic Scheduling, Linear Quadtree, GPU.

1 Introduction

Real-time visualization of large scale terrain has been an active area of research in flight simulation, virtual reality, GIS and other various fields. But due to the huge amount of raw data and photo textures, it is still a challenge for recent PC architectures because of limitations in memory and communication bandwidth. Thus, the main problem for large scale terrain real-time rendering today is how to maintain interactive frame rates for data sets that vastly exceed the available main memory [1].

2 Related Work

The early terrain rendering algorithms were executed on the CPU. They focused on how to decrease the number of triangles and improve the realism effect as well. The classic academic algorithms were ROAM and restricted quad tree algorithms [2, 3].

However, since the processing power of GPU has been improved much faster than before, the main impact factors have become CPU processing time and communications between CPU and GPU. The terrain rendering algorithms have changed to draw on GPU as much as possible. de Boer [4] proposed Geomipmapping algorithm. Compared with ROAM, it produced a higher triangle output with less optimal distribution but also less expensive execution on CPU, thus provided better result. Later, Ulrich [5] developed a GPU-based restricted quad tree algorithm, which has good detail distribution and optimally tessellated meshes. Other popular algorithms include geometric clipmap and CDLOD (Continuous Distance-Dependent Level of Detail) that also have effective performance [1, 6 and 7].

Unfortunately, these optimized algorithms work on few numbers of terrain blocks and small terrain regions. To meet real-time rendering of large scale terrain, we must make dynamic management and scheduling of terrain blocks as consideration.

3   Contribution

The primary contribution of this paper is the development of a dynamic scheduling scheme for terrain real-time rendering. Based on Geomipmapping algorithm, our optimal method has been carried out mainly in two fields. On one hand, terrain pyramid model, which consists of the same size blocks, is built with different scales and biases factors offline; On the other hand, we create linear quadtree index of terrain blocks for fast search and query, and propose more reasonable dynamic scheduling scheme. With these efforts, memory space, CPU processing time and GPU rendering batches have been dropped down significantly.

4   Geomipmapping Algorithm Overview

Geomipmapping is one kind of block LOD terrain real-time algorithm. As mipmap technique for textures, this algorithm applies this concept to three dimensional meshes [4]. It could create different LOD levels offline, which decreases CPU processing time online. However, quadtree management is only used in Geomipmap levels of each terrain block. Different terrain blocks are independent, which are distributed with the index of row and column values. This will make memory space and GPU rendering batches increase too fast when the scale of terrain extends.

5   Real-Time Rendering Algorithm Based on Linear Quadtree

To solve above disadvantages of Geomipmapping algorithm, our implementation focuses on dynamic scheduling and management. This section describes the complete algorithm into five parts. The first part will introduce optimized structure of terrain blocks; Section 2 will give the particular dynamic scheduling criterion and rule; Section 3 will improve the basic algorithm with two aspects; and the rest ones will solve the problems caused by different layers and LOD levels terrain blocks loading.

5.1 Improved Structure of Terrain Blocks

Compared with the sequential storage structure of Geomipmap algorithm, we design terrain pyramid model which has different resolutions and same size blocks; linear quadtree index of terrain blocks are created; quadtree nodes and scene nodes are brought in to store terrain blocks. These implementations are able to improve the follow-up dynamic scheduling scheme efficiency.

5.1.1 The Pyramid Model of Terrain Blocks

In traditional terrain pyramid model, vertex amount will increase by 4 times when the layer number decreases. So the lower layer blocks cover with more vertexes and larger range of terrain region. However, they are chosen to render only when the distance from the camera is very far. It seems that virtual effects are not obviously improved if more vertexes are used to create the lower layer blocks.

Thus, we change the vertex amount of all layers blocks in terrain pyramid model into the one of the highest layer blocks. The pyramid model could greatly reduce the stored vertex number, and meet LOD requirement as well.

As Fig. 1 shows, the build processing of terrain pyramid model is:

1). De-sample DEMs to get the same size blocks
2). Calculate vertexes coordinate with the scale and bias factors of terrain block
3). Write vertexes coordinate into vertex and index buffers in GPU.
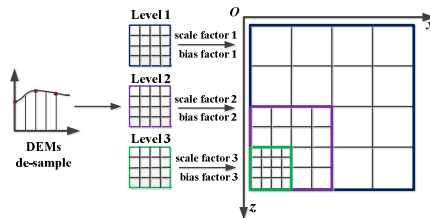
Fig. 1. The build processing of terrain pyramid model

5.1.2 The Linear Quadtree Index of Terrain Blocks

Linear quadtree uses nodes encoding for unique identification with regions related to nodes. It needs less memory space and finds neighbor faster. The most common encoding methods are quaternary and decimal Morton code [8].

To improve the speed of terrain blocks search, we create linear quadtree index of terrain blocks. In correspondence with terrain pyramid model, linear quadtree index starts with the top of terrain pyramid model, and encodes from left to right, top to down. Fig. 2 shows the quaternary and decimal linear quadtree indexes of 3 layers pyramid model. Meanwhile, to ensure linear quadtree lay-encoding consistent with layers of terrain pyramid model, each layer terrain blocks use unique one-dimensional array to store linear quadtree index.
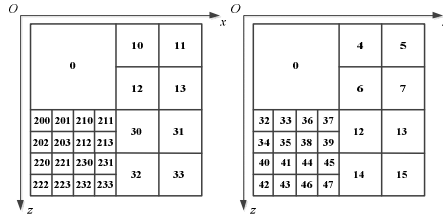
Fig. 2. The quaternary and decimal linear quadtree indexes in 3 layers pyramid model

5.1.3   Node Structure

To apply terrain pyramid model and linear quadtree index, node structure used for terrain blocks management is designed. As Fig. 3 shows, node structure includes quadtree and scene node structures.
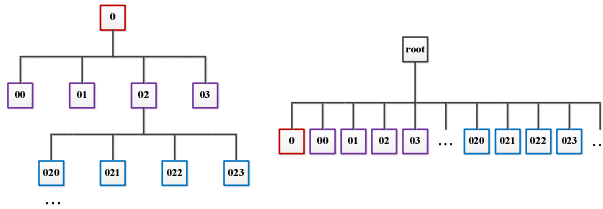
Fig. 3. Quadtree node structure is on the left, and scene node structure is on the right

Quadtree node structure works on terrain blocks search, each node of it has three parts: 1) linear quadtree index of attached terrain block; 2) parent and four children nodes indexes; 3) tree depth in quadtree node structure.

Scene node structure is in charge of quadtree nodes. It lets quadtree nodes coordinative relation, and classifies them to visible and invisible node lists. This could be easy for scene management. Scene node structure consists of two parts: 1) attached quadtree node; 2) visible flag of attached quadtree node.

5.2   Dynamic Scheduling Scheme of Terrain Blocks

Large scale terrain has so many blocks that it cannot be loaded to memory once. So these terrain blocks have to be updated on-the-fly. We design a more reasonable dynamic scheduling scheme, and accomplish efficient management.

In the aspect of scheduling criterion, geometric clipmap algorithm [6] uses two-dimensional components of distance and ignores the height component, which is obvious unreasonable. For example, when the height component of viewpoint increases, it should render lower layers and LOD levels terrain blocks; Though Chunked LOD algorithm [5] uses an LOD function that is the same over the whole chunk, it provides only approximate three-dimensional distance-based criterion [7].

To manage terrain blocks more reasonably, we choose three-dimensional distance between viewpoint and terrain block center as a basis that adopts L_1 calculation formula:

$$dis = \|camera - center\|_1 . \tag{1}$$

Next, we will describe the switching function among different layers terrain blocks. As Fig. 4 shows, to a given terrain block, l is its side length; O is its center; $c_1 \sim c_4$ stand for its children centers. We define a hemisphere zone, its center is O, and its radius is r. When three-dimensional distance between viewpoint and terrain block center is less than r, it should be divided into four children blocks; otherwise, the terrain block should be sent to rendering pipeline. The calculation formula of r is:

$$r = k \cdot \text{sqrt}(2) \cdot l . \tag{2}$$

Here, k is called adjustable ratio, which controls the visible range of terrain blocks.
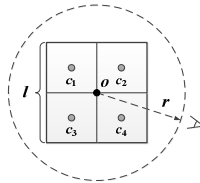
Fig. 4. The rule for switching different layers terrain blocks in terrain pyramid model

LOD level is chosen by the following formula:

$$LOD\_level = \lceil |camra.y - center.y| / stepdis \rfloor . \tag{3}$$

Here *stepdis* is step distance, $\lceil \rfloor$ stands for round down.

Dynamic scheduling scheme can be described as the following steps:

Step1. Judge whether regions covered by nodes intersect with visible view-field. If yes, execute next step; otherwise, empty visible node list, go to Step5.

Step2. Judge whether each terrain block of this layer is suitable with distance threshold. If yes, choose LOD level with Eqn. (3), and add it into visible node list; otherwise, add it into invisible node list, send its four children to test list.

Step3. If test list is empty, go to Step5; otherwise, execute next step.

Step4. Judge whether region covered by each node in test list intersects with visible view-field. If yes, delete this node and go to Step2; otherwise, go to Step3.

Step5. If visible node list is not empty, send attached terrain blocks to rendering pipeline; otherwise, empty visible and invisible node lists. The algorithm ends.

View-frustum culling could use AABB bounding box [9] for fast test. Since selecting terrain blocks from linear quadtree index arrays do not analyze spatial location, it can improve the performance of our algorithm obviously.

5.3   The Implementation of Basic Dynamic Scheduling Scheme

It is waste space memory of storing terrain blocks located below the horizon. Thus, we set the farthest three-dimensional distance where ones are visible. As Fig. 5 shows, r is mean radius of the earth; h is the height of viewpoint; the farthest distance *far_dis* is calculated in the right-angled triangle OAB, with the Pythagorean Theorem.
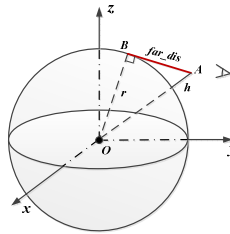
Fig. 5. The farthest three-dimensional distance (red color) from the viewpoint

If too many terrain blocks are dealt with in one frame, the frame rate will drop down all of a sudden. To avoid unstable rendering quality, we make a rule that each frame only allows loading one terrain block. It won`t send updated terrain blocks to rendering pipeline until all of them are loaded. With restricting the loading number of terrain blocks in one frame, the rendering rate could be smoother than before.

5.4   GPU-Based Geometric Morphing

When different layers and LOD levels terrain blocks are loaded, the change of vertex height will cause popping in visual effects. As Fig. 6 shows, point A will be replaced by point A_1 interpolated by point B and C. The height difference is h.
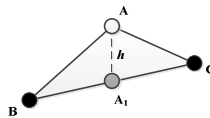
Fig. 6. Popping in visual effects between point A and A_1

To slow down popping in visual effects, we design GPU-based geometric morphing. Firstly, it needs to store height difference h in the vertex texture coordinate; then we set morphing time t. During morphing time t, height compensation is done first, and then terrain blocks are switched. With vertex shader in GPU [10], output and input vertex height coordinates transform formulas are:

$$Out\_pos.y = In\_pos.y + f \cdot h . \tag{4}$$

Here, f stands for the ratio between current and morphing time.

5.5   Crack Elimination

We adopt vertical skirt mask to eliminate crack. As Fig. 7 shows, a vertical "skirt" is created around the perimeter of each terrain block, and the top of the skirt matches the terrain block edge. With this vertical skirt, crack will be filled easily. This method could reduce CPU processing time; though a small number of triangles are added. Fig. 8 shows the effect of crack elimination where texture is stretched on the vertical skirt.
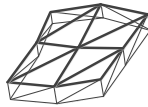
Fig. 7. Add vertex skirt around terrain block border edge

Fig. 8. Crack between terrain blocks on the left and the effect of crack elimination on the right

6   Results and Analysis

We evaluate our algorithm on a PC with an Inter (R) Xeon (R) E5420 processor, 4 GB of RAM, an NVIDIA GeForce GTX560 with 1 GB of video memory.

16-bits DEM contained with 4097 × 4097 vertexes is de-sampled and segmented as PNG (Portable Network Graphics) file type. We build 6 levels terrain pyramid model which covers 300 km × 300 km area, and each block size is 129 × 129.

The particular parameters in our experiment are shown in Table. 1. Table. 2 and 3 indicate switching thresholds among different layers and LOD levels. The main rendering quality items of our and Geomipmapping algorithms are shown in Table. 4.

Table 1. The particular parameters in our experiment

Parameter (unit)	Value
Adjustable ratio k	0.75
Morphing time t (millisecond)	3.0
The smallest edge length (m)	10500
LOD step distance (m)	2000
The height of viewpoint (m)	2000 ~ 10000
The farthest three-dimensional distance (km)	200
The movement speed (m/s)	100
Running frames	14046

Table 2. Switching distance threshold among different layers in terrain pyramid model

Layer	Distance threshold (m)
0	(50400, 200000]
1	(25200, 50400]
2	(12600, 25200]
3	(6300, 12600]
4	(0, 6300]

Table 3. Switching height threshold among different LOD levels in each terrain block

LOD layer	Height threshold (m)
0	(6000, 200000]
1	(4000, 6000]
2	(2000, 4000]
3	(0, 2000]

Table 4. The main rendering quality items of our and Geomipmapping algorithms

Quality items (unit)	Our algorithm	Geomipmapping algorithm
DEM storage (MB)	5.85	76.9
GPU batches	25	62
Average frame rate	460.3	350.5

As it is known, the efficiency of GPU is mostly restricted with the number of geometric model batches sent to API. For example, with D3D mode, one batch processing calls DrawPrimitive or DrawIndexPrimitive function once. Before GPU starting rendering, CPU should cost so much time to deal with each batch data. Thus, the reduction of GPU batches could obviously drop down CPU and GPU processing time, and improve frame rate as well.

As Table. 4 shows, our algorithm de-samples DEMs to substantially reduce vertex number, and uses linear quadtree index for dyanmic management. With the optimization of the two aspects, it could decrease GPU batches significantly, so our algorithm shows better performance than Geomipmapping algorithm.

Fig. 9 shows final terrain visualization effect based on our algorithm.
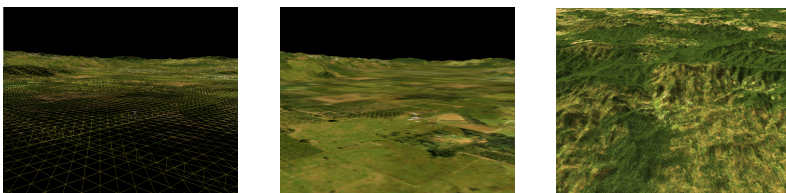
Fig. 9. Terrain visualization effect includes meshes structure on the left, texture mapping on the middle and realistic light mapping on the right

7   Conclusions and Future Work

We have presented a real-time terrain rendering algorithm based on linear quadtree. We de-sample DEMs for getting the same size blocks, and build terrain pyramid model with different scales and biases factors, which makes the number of rendering vertexes decrease obviously; Then linear quadtree index is adopted to dynamic scheduling scheme, with optimized scheduling criterion and rule, it could reduce CPU processing time and GPU rendering batches significantly. The improved algorithm, which keeps a high and smooth frame rate, could be used for real-time rendering in large scale terrain.

In the future, we will research on real-time compression of high-resolution photos used as terrain texture, and improve the visual effects of light and shadow. With the implementation of the two parts, we hope our large scale terrain visualization system more realistic while ensuring real-time quality.

References

1. Dick, C., Schneider, J., Westermann, R.: Efficient Geometry Compression for GPU-based Decoding in Realtime Terrain Rendering. Comput. Graph. Forum 28(1), 67–83 (2009)
2. Duchaineau, M., Wolinsky, M., Sigeti, D.Y., Miller, M.C., Aldrich, C., Mineev-Weinstein, M.B.: ROAM Terrain: Real-time Optimally Adapting Meshes. In: Proceedings of the 8th Conference on Visualization, pp. 81–88 (1997)
3. Röttger, S., Heidrich, W., Slusallek, P., Seidel, H.: Real-Time Generation of Continuous Levels of Detail for Height Fields. In: Winter School in Computer Graphics 1998, pp. 315–322 (1998)
4. de Boer, W.H.: Fast Terrain Rendering Using Geometrical Mipmapping (2000), http://www.ipcode.com/tutorials/geomipmaps.pdf
5. Urlich, T.: Rendering Massive Terrains Using Chunked Level of Detail Control. Technical report, SIGGRAPH 2002 Course Notes (2002)
6. Asirvatham, A., Hoppe, H.: Terrain Rendering Using GPU-Based Geometry Clipmaps. In: Pharr, M., Fernando, R. (eds.) GPU Gems 2, pp. 27–45. Addison-Wesley (2005)
7. Strugar, F.: Continuous distance-dependent level of detail for rendering heightmaps. J. Graph. GPU. Game Tools 14, 54–74 (2009)
8. Szofran, A.: Global Terrain Technology for Flight Simulation (2008), http://www.microsoft.com/Products/Games/FSInsider/developers/Pages/GlobalTerrain.aspx
9. Akenine-Moller, T., Haines, E.: Real-Time Rendering, 3rd edn. AK Peters (2008)
10. Fernando, R., Kilgard, M.J.: The Cg Tutorial The Definitive Guide to Programmable Real-Time Graphics. Addison-Wesley (2003)

Research on Modeling and Application
of Synthetic Natural Environment

Zhenhua Lv and Guanghong Gong

Aviation Key Laboratory for Advanced Simulation Technology,
Beihang University,
100191, Beijing, China
lzh_shine@asee.buaa.edu.cn

Abstract. Traditional simulation application was normally carried out with restricted consideration of environmental factors. Through comprehensive and systematic analysis of natural environment, a novel integrated environment representation method was proposed and applied to a vehicle-mounted platform simulation application system with synthesized concerns of natural factors. The comprehensive simulation experiments with diversity of natural environment elements were integrated in traditional system employing the mathematical models and the 3D rendering engine techniques to reflect the relatively close-to-nature effects of the simulation system. The simulation results are proved to be competent and offered an applicable solution for evaluating the simulation system.

Keywords: synthetic natural environment, application system, integrated representation, impact analysis.

1   Introduction

The system simulation has been intensively developed and widely used in diversity of domains. The system modeling accuracy has continuously improved. The fidelity and the credibility requirements of simulation system have become more sophisticated. Therefore, simulation has evolved from focusing only on the simulation simplex object to considering the impacts of Synthetic Natural Environment (SNE) factors (terrain, ocean, atmosphere, space, etc.). The environmental simulation fidelity takes an important effect on the fidelity and credibility of the whole simulation system. The interdisciplinary problems between single simulation entity and complicated environment have been increasingly prominent.

Terrain surface factor effect represents the basis of the virtual system in the field of environment simulation. In the modern military system simulation, the main focus is not only on the terrain visualization, but also on the impact of the topography surface characteristics to the equipment. The terrain bumping effect is a class of representative interactions, reflects the significant coupling impacts on the vehicle-mounted platform from the terrain road conditions.

In current study, it focuses mainly on the integrated environment representation issues and environment application simulation. The terrain effect on vehicle-mounted platform application system has demonstrated the necessity of the significance of natural environmental factors for the simulation system.

2 Simulation Model of the SNE

2.1 SNE Conceptual Reference Model

The SNE is a set of authoritative, integrative, polymorphic and consistent data descriptions, model representations and simulations of natural environment domains [1-2].The SNE conceptual reference model emerged in recognition and practice of the SNE modeling and simulation which implied a milestone for the SNE. In 1998, Brikel presented the SNE conceptual reference model, categorized the component of the SNE and analyzed the relations between the components [3].

The conceptual reference model provided authoritative concept for explaining the division of the research contents of the SNE modeling and simulation. It contains two major sections: synthetic natural environment and military system models. The former section provides a battlefield simulation natural environment for the latter one. Note that the entities with physical properties (vehicles, buildings, etc.) belong to the elementary parts in the environment.

2.2 Environment Influence for Entity

The conceptual reference model in section 2.1 comprehensively described the interactions between the environment and the military system. The environmental effect factors of the application system which the current study interested in are analyzed in Fig.1.
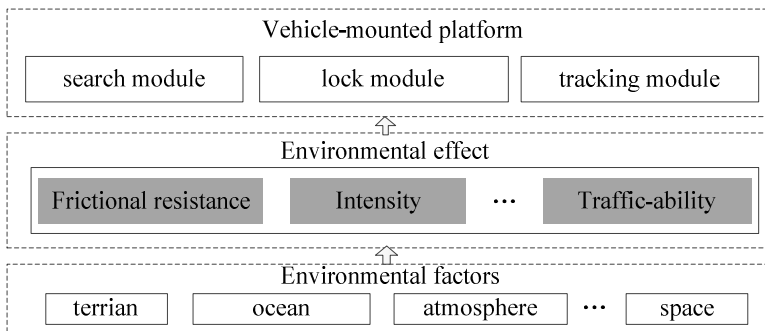
Fig. 1. Environmental effect factors of the application system

The vehicle-mounted platform is a typical entity which is influenced by the terrain factors [4]. Road frictional resistance, intensity, and traffic-ability all affect the vehicle movement properties and consequently affect the stability of the vehicle-mounted platform.

2.3   A New Integrated Environment Representation Method

Natural environment elements can be classified into two types according to the different features: (1) Elements discrete distributed in space, sharp-edged and functional complete (terrain, trees and buildings etc.), which are appropriately modeled by the object-oriented theory; (2) Elements continuous distributed in time and space (atmosphere, ocean, etc.), which are suitable for modeling by data field theory.

While constructing environmental data model based on the modeling theory combined object-oriented and data field, the natural environment from overview is divided into different environment objects. Due to different simulation applications (CGF, visualization, detector, etc.) have different needs of the environment objects the environment objects possess a variety of forms, which are expressed in different property sets, different behavior models and different object associations. Each environment object is expressed with a five-tuple by eq. (1):

$$EO_i = (ID, C, A_i, P_i, R_i), i = 1, \cdots, n .\tag{1}$$

The concrete meanings are summarized in Table 1.

Table 1. Nomenclature

Symbol	Meanings
EO	Environment object
EO_i	The i-th expression form of the EO
ID	The EO identification
C	The EO category
A_i	The attributes set of the EO_i
n	The expression form quantity of the EO
P_i	The physical model of the EO_i
R_i	The associations between EO_i and other environment objects

*Remark 1.*The identification and category of the environment object does not change with its manifestations.

*Remark 2.*For different simulation applications, the environment object can possess different attributes. For instance, visual simulation system requires surface material and texture attributes of the environment object, infrared simulation system needs the radiation and thermodynamic properties and radar simulation system calls for radar cross section reflection properties. For environment object suitable to be described as the form of field, its attributes can be expressed in a form of environmental data field.

*Remark 3.*The physical model include private dynamic model and the influencing model of weapon equipment entity to the *EO* . For different simulation applications, environment objects can possess different physical models. For example, tanks need road intensity and frictional resistance variation model, whereas infrared sensor calls for the pavement temperature change model.

*Remark 4.*The relation categories include spatial relationship, topological relationship, inheritance relationship, bearing relationship, heat exchange relationship, etc. Different simulation applications concern different associations among environment objects. For instance, vehicles need roads connectivity information for path planning, which aircrafts do not require.

Environmental data model does not include the physical model of the environment object [5], it can be written in the following form:

$$EDM_{EO} = (ID, C, A, R) \tag{2}$$

where A is the attributes set, including all forms of environment object attributes, R the associations of all *EO* behaves.

2.4   Terrain Modeling

The simulation needed terrain model is built combining with the integrated environment representation method and Synthetic Environment Data Representation and Interchange Specification (SEDRIS) [6]. The SEDRIS Environmental Data Coding Specification (EDCS) is utilized as the data dictionaries, coding for the environmental objects and the attributes. The EDCS does not define the attributes for the environmental objects. The integrated environment model associated the specific environmental object to the expected particular attributes set and determined the allowed values for each attribute.

Terrain geometric model is realized by the Digital Elevation Model (DEM) based elevation mapping. DEM data are commonly represented by 2-D array or grid in terrain elevation map with regular and irregular spatial grid. In this study, the terrain geometric model is created by Delaunay triangulation. For the simulation of large-scale topography the concept of paging, grouping and segmentation is used in organization and management of topographic data. Meanwhile, using the Delaunay algorithm to generate the triangular network also requires the paging boundary shear and alignment.

Three representation approaches are devised for terrain environment objects: geometric representation, feature representation and hybrid representation [7-8]. The representation approaches of terrain environment object are depicted in Fig.2 in the form of UML. Take ECC_ROAD object for an example, it is expressed by line features and primitive geometries and possesses surface material and surface friction coefficients and other flexible properties. The terrain environment database is mapped onto the SEDRIS in this study shown in Fig.3. For the integration of environment data exchange, SEDRIS Transmittal Format (STF) is referenced.
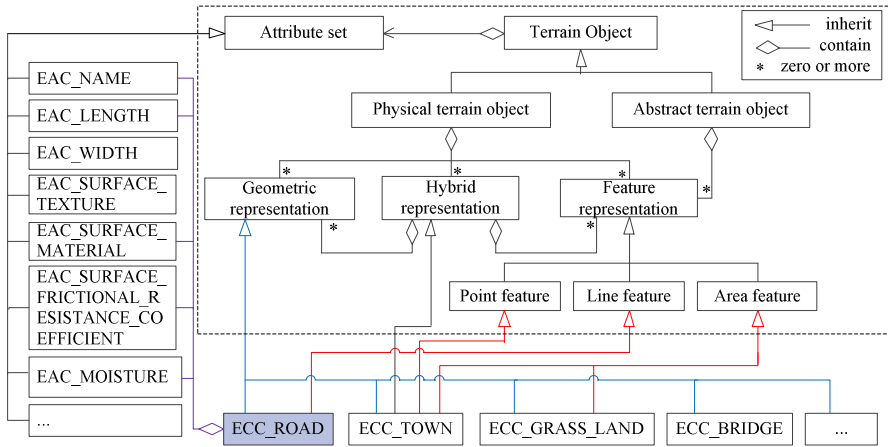
Fig. 2. The terrain environment object representation
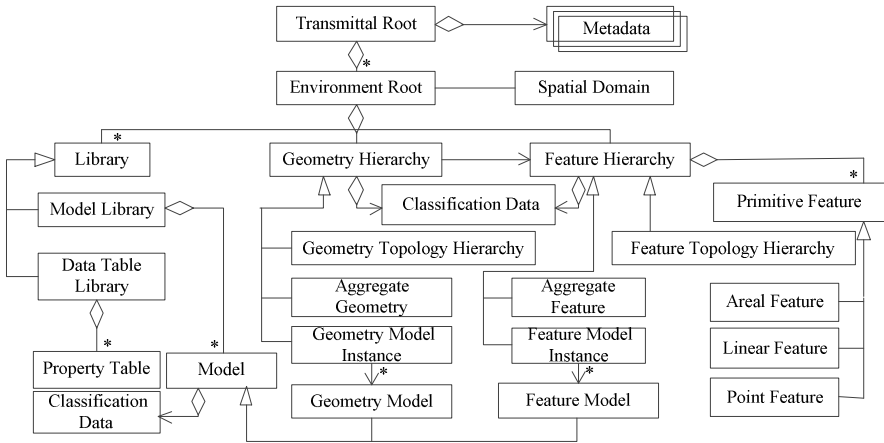
Fig. 3. The terrain environment data SEDRIS mapping

In order to satisfy the needs of the application system simulation in this study, terrain surface physical properties (soil type, prevailing conditions, surface material, ground moisture, etc.) are considered in terrain modeling. In the terrain grid data model, the storage and organization of the properties data is similar with the terrain DEM data. The index value of the properties data storage is through associating the triangle data. In order to reduce the redundancy of data storage, secondary indirect index method is used to access the properties data.

3   Application of Synthetic Natural Environment

To verify the importance of the impact of the environment, a vehicle-mounted platform simulation application system with synthesized concerns of natural factors were studied of interest specially the terrain environment effect.

3.1   Terrain Influence on Vehicle-Mounted Platform

In order to prove the effect of terrain surface factors on the vehicle-mounted platform, three-dimensional terrain data model and physical model were established. Adding the motion model of the platform to the built simulation system, the stability effect of the terrain environment will be recast in real-time.

3.2   The Core Module MATLAB Model for the Vehicle-Mounted Platform

After detailed analysis of the control circuit of the vehicle-platform, the mathematical model was built in article [9]. The core module is detailedly described in Fig.4.
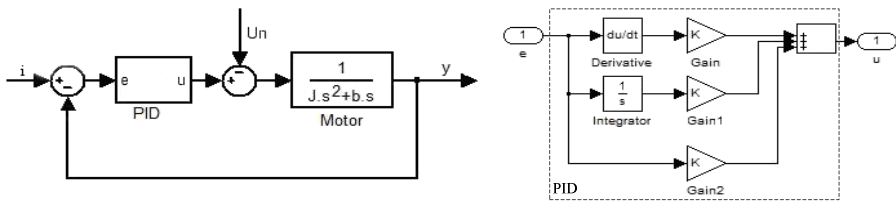
Fig. 4. Core module model of the stable platform in Simulink (i :the input of the core module; y: the output of the motor; e: the input of the PID module; u: the output of the PID module. Un: the input of the interference.)

After Z transformation, the module discrete equations are summarized as the following equations:

$$y(k)-(1+e^{-\frac{b}{J}T})y(k-1)+e^{-\frac{b}{J}T}y(k-2)$$
$$=-\frac{2J}{b^2}u(k)+(\frac{3J}{b^2}+\frac{J}{b^2}e^{-\frac{b}{J}T}+\frac{T}{b})u(k-1)-(\frac{T}{b}e^{-\frac{b}{J}T}+\frac{J}{b^2}e^{-\frac{b}{J}T}+\frac{J}{b^2})u(k-2) \tag{3}$$

$$u(k)=k_p e(k)+k_i \sum e(k)T+k_d \frac{e(k)-e(k-1)}{T} \tag{4}$$

The role of Z-transformation is to discretize the mathematical model of the stable platform and add to the real-time simulation system.

4   Simulation and Result

The environmental affected simulation system is constructed with the open source rendering engine OGRE. The influencing factors on the application system are researched from the realistic natural environment.

4.1   Simulation System Architecture

The simulation system is established on the OGRE rendering engine. The OGRE manages the objects in the scenes by the scene-graph algorithm and keeps the highly scalability and reusability by the techniques of the versatile plugins. OGRE realizes the flexible scheduling of different rendering system by packaging the fundamental 3-D APIs. The system in the current study is built up by the secondary development on the OGRE engine. The system resolves the platform angle variation using the mathematical models from the impacts of environmental factors and driving the 3-D model's motion. The above procedures render the platform motion simulation in real time and offer the error curves for the real-time environmental simulation.

4.2   The Terrain Effected Stability Simulation

The system real-time rendering and simulation effects indicating the motion simulation effect of the vehicle mounted stabilized platform on a bumping road. The stabilized platform is installed on the top of vehicles. The real-time simulation system simulates the terrain surface factors (frictional resistance, road intensity, traffic-ability, etc.). The environmental parameters are versatilely configured according to different circumstances. Therefore, the stability of the platform and environmental impacts are illustrated intuitively.

4.3   Results Comparison on MATLAB and Environmental Simulations

The stability comparison in MATLAB is evaluated by changing typical responses (step response and sinusoidal response). The stability of the platform in the real-time environment simulation system is evaluated by the real-time terrain environment response. The comparison of responses and error variation results are shown in Fig.5.

The simulation results in MATLAB are demonstrated that the error variations of the axis in difference responses inputs reach acceptable agreements.

The simulation result in real-time environment simulation system is implied that the stable platform has a good response during most of the simulation time, whereas the system also appeared three times larger fluctuations which are not observed in the simulation with typical responses (step, sinusoidal, etc.) in MATLAB. From this simulation, the researchers can verify the rationality of the control method, and improve the mathematical model of the stabilized platform control system, which overcomes the simulation limitations by MATLAB.
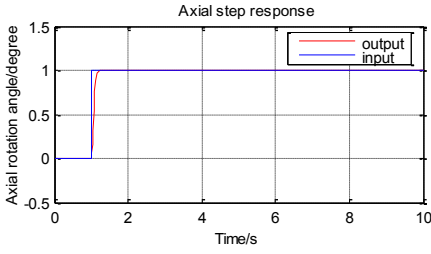
(1) step response curve
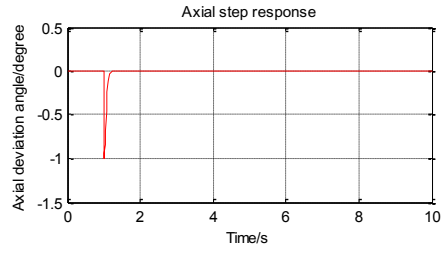
(2) step response error curve
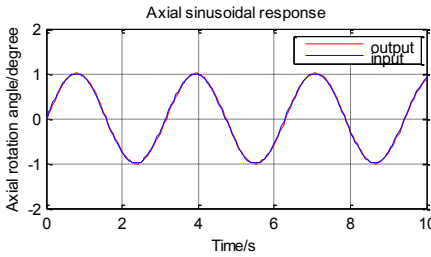
(3) sinusoidal response curve
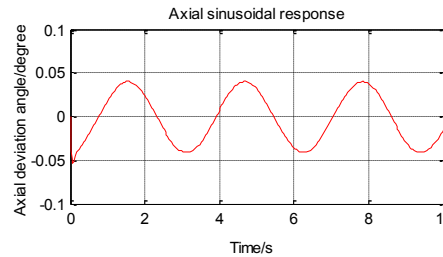
(4) sinusoidal response error curve
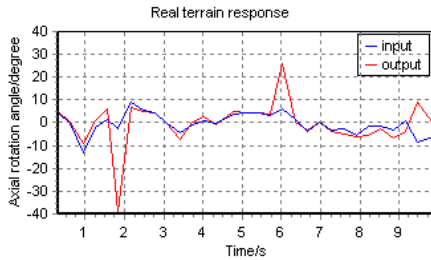
(5) real-terrain response curve
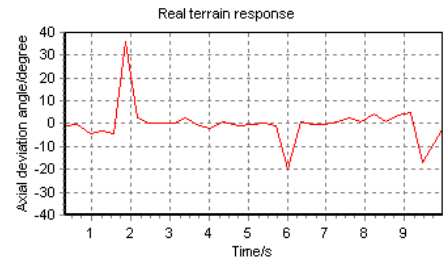
(6) real-terrain response error curve

Fig. 5. Comparison of the typical responses and the real-terrain response

5   Conclusion

Through comprehensive and systematic analysis of natural environment, a novel method of integrating environment representation in simulation system was proposed. On the basis of the idea, a new simulation attempt considering the environmental factors effect on the application system was proposed combines the three dimension rendering engine.

The current study eliminates the limitations of conventional application system, which was utilizing the representative and restrictive interference inputs by the experimental provoking signals. Comparing with the traditional simulation method, the real-time environment experiments demonstrate the novel simulation approach is relatively fidelity prone, reflects the influence factors to application system with comprehensive simulating information, evaluates the performance of the system with enhanced credibility and provides a practical solution for the application simulation system.

A contradiction exists in the simulation engineering domain that the close-to-real environmental factors assists the effective results of the application system whereas the complicated natural factors deteriorate the simulation real-time features for the heavy computational resources consumption, hence introducing improved natural models and meliorated calculating techniques are possessing promising research meanings.

References

1. Gang, G.: Research on Synthetic Natural Environment Modeling and Simulation. National University of Defense Technology (2004)
2. U.S. Army Simulation, Training and Instrumentation Command, SEDRIS Documentation Set, Orlando, Florida (March 1998)
3. Brikel, P.A.: SNE Conceptual Reference Model. Presented at the Fall 1998 Simulation Interoperability Workshop, Orlando, FL (1998)
4. Brikel, P.A.: Terrain Trafficability in Modeling and Simulation, Orlando, SEDRIS (2002)
5. Miller, D.D., et al.: An Environmental Data Model for the OneSAF Objective System. In: Simulation Inter-operability Workshop (2002)
6. http://www.sedris.org
7. Tosh, A., Campos, J.: SEDRIS'izing the CTDB Landscape. In: Proceedings of the Fall Simulation Interoperability Workshop, Paper No. 02F-SIW-094 (September 2002)
8. Donovan, K.B., et al.: Virtual Database Production Experience with SEDRIS. In: Proceedings of the Fall Simulation Interoperability Workshop, Paper No. 01F-SIW-106 (September 2001)
9. Lv, Z., Gao, D., Chen, D.: Simulation Technology Study on the Effects of Environmental Factors on the Platform Motion Performance 15(3), 301–307 (2010)

Symbolic Representation of Vector Map in Virtual Geographic Environment

Xuefeng Cao and Gang Wan

Institute of Surveying and Mapping, Information Engineering University,
450052, Zhengzhou, China
`cxfchxy@163.com`

Abstract. Those basic geometric graphs such as lines and polygons are the main representations used by 3D rendering methods of vector map at present. It is so simple that strictly limited map feature information has been transferred. Consequently, an extended rendering method based on the shadow volume stencil theory is presented, which allows real time symbolical overlay of vector map on terrain. Firstly, the basis of vector data visualization based on stencil shadow volume theory has been reviewed. Secondly, in order to improve visualization effect, the cartographic symbolization of vector map, i.e. roads, has been dedicated, which includes smooth border with rounded caps, outlines and overlaps. Finally, the experiment shows that real time displaying of vector map on terrain has been achieved, and the visualization effects have been improved by cartographic symbolical overlay.

Keywords: Vector Map, Virtual Geographic Environment, Stencil Buffer, Shadow Volume, Symbolization.

1   Introduction

Geographic environment visualization, geographic spatial analysis, geographic phenomena and process are the main functions of virtual geographic environment [1], all of them need the depiction, representation, modeling and analysis of geographic entities' spatial position and distribution, attributes and mutual relations information. As the main carrier of geographic entities' spatial information, 3D visualization of vector map has been a research hotspot.

These overlaying rendering methods of vector map on 3D terrain proposed in recent years have been divided into four kinds: geometry-embed [2~4], geometry-overlay [5~10], texture-layer [11~14], entity procedural modeling [15, 16]. The main map features representation forms of current methods are basic geometric primitives, such as point, polyline, curve and polygon, which using the color, width of line and the color, width, interior fill of polygon to distinguish different geographic features. Compare with traditional cartographic symbolic representation, the visual impact and transmitted information of these existing methods are very limited. Under the instruction of map symbol theory, and there will be great improvement of geographic feature visualization impact.

The shadow volume stencil based vector map overlaying method [17] is spirited by the shadow volume theory [18], which is a technique used in 3D computer graphics to add shadows to a rendered scene. Compared with other methods, there are two main advantages of this method, firstly it works in screen-space and can therefore be performed per-pixel exact, secondly it is independent of the underlying terrain geometry and utilized rendering engine offering high performance even for very high resolution data sets and a wide applicability.

In this paper, an extension of the shadow volume stencil (SVS) is proposed, in which vector map features will be rendered with symbols like that in traditional electric map. In order to realize symbolization of vector map in virtual geographic environment, road feature is given as an example, the rounded caps, outlines and intersections are investigated.

2   Symbolic Rendering Method of Map Features

Form the geo-spatial information visualization model [19] angle of view, the requirement of vector data visualization in VGE could not be satisfied by overlapping geometric graphs in vector map data on terrain only. For the vector map visualization in VGE, it is obligatory to be up to the mustard of cognition requirement and visual demand, because visualization mainly depends on the visual perception channel. The cognition research indicates that the visual information obtain mechanism is synoptic and topological. So that, in order to avoid falling into multifarious superficial representation and guarantee objects' scientific cognition obtainment, these contexts related to investigation aims have to be reserved, emphasized, while redundant details being omitted. The map cartographic generalization is a typical example. In the map cartographic generalization, geographic entities have been abstracted as map symbols, and then geographic entities' spatial, attribute and temporal information have been reflected by map symbols' visual variables. This abstraction and generalization methodology has been proved to be scientific and impactful in long-term practice.

In order to represent these map features with smooth vivid symbols, a symbolic rendering method of vector map features in VGE is proposed in the following.

2.1   Shadow Volume Stencil Basis

The basic idea of Shadow Volume Stencil (SVS) is to extrude the vector data to polyhedra and use them to create a mask in the stencil buffer. The generated mask corresponds to the projection of the vector data onto the terrain surface. It is applied to the scene by rasterizing geometry covering at least the entire mask with the appropriate stencil test enabled.

The approach of SVS consists of three parts:

(1) Constructing the vector data polyhedra: extrude the vector data into the polyhedra according to the bounding box of the terrain data where the vector data lie.

(2) Creating the mask: render the polyhedra to the stencil buffer to create a mask, which is consistent with the vertical projection of the vector data onto the terrain surface.

(3) Rendering the vector: apply the mask to the screen area covered by the vector data by using the appropriate stencil test method.

2.1.1   Construction of Vector Data Polyhedra

Construction is started by duplicating each vertex of the vector data. One vertex of each of the created pairs is translated towards the geocenter, the remaining vertices are moved into the opposite direction. The group of upper and lower vertices constitutes the polyhedron's top and bottom cap. The amount of translation has to be chosen such that the top and bottom cap are located completely above and below the terrain surface respectively. Applying the described construction the resulting polyhedron encloses the part of the terrain surface that is supposed to contain the vector data.

2.1.2   Generation of Mask

These polyhedra have been created, and they can be rendered into the stencil buffer. When the viewpoint is inside the shadow volume or the shadow volume intersects the near clipping plane of the view frustum, the z-fail algorithm is used; otherwise the z-pass algorithm is used. The steps of creating the mask are as follows: (1) Color, depth and stencil buffers are cleared and the terrain is rendered initializing the depth buffer with the required depth values. (2) Depth buffer writing is disabled, but the depth test still remains active. Rendering is then restricted to the stencil buffer only. (3) Rendering vector polyhedra to create the mask. The polyhedron's faces are rendered using different stencil operations depending on whether they face towards or away from the camera. To this end, face culling is enabled and the polyhedron is rendered twice, one time with back-face culling enabled, the other time with front-face culling enabled.

2.1.3   Mapping of the Mask onto Terrain

The mask created in the stencil buffer is applied to the scene. Therefore, we reactivate writing to the color buffer and activate additive blending. The stencil test is configured to pass only when the value in the stencil buffer does not equal zero. Instead of drawing a screen-sized quad which is applied the mask to the scene, we rasterize the bounding box of the respective polyhedron in order to save rasterization bandwidth. This is performed with depth test enabled and drawing only front-faces in the z-pass case and with depth test disabled and drawing only back-faces in the z-fail case.

2.2   Extension of Shadow Volume Stencil

In vector map data, road features are usually stored as a collection of 2D polylines. Currently, these visualization methods mainly convert the vector data into geometric primitives that are rendered on top of the terrain. As a result, these representation

styles only include color, linetype, linewidth of polylines. With the extrusion of polylines to a quad, these representation styles of road feature could be expanded to interior color, smooth edges, outlines and so on.

2.2.1   Rounded Cap

The naive approach is to extrude each line segment of road polylines to a quad, which results in the appearance of cracks between segments. The higher the curvature of a polyline, the more these cracks become visible. Two pragmatic solutions exist: filling the holes with additional triangles (see Fig. 1(a)) or connecting the corners of adjacent quads (see Fig. 1(b)).

(a) Extra triangle   (b) Moving corners   (c) Rounded caps
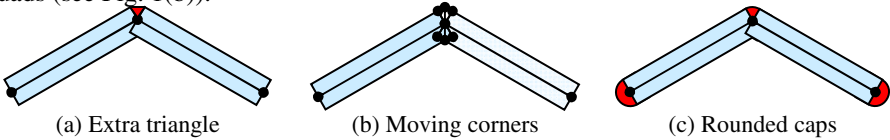
Fig. 1. Methods for the treatment of cracks between quads

A robust and elegant method has been proposed [20], which draws caps to avoid the appearance of cracks (see Fig. 1(c)) and does not require adjacency information. As following, this method has been used in the smoothly rendering of linear features.

A single quad encompassing an entire road segment is rendered, and the caps has been evaluated analytically in a shader program. This is more effective than to explicitly triangulate half-circles at the beginning and the end of each road segment. Using the endpoints P_i and P_{i+1} of each line segment and the tangent $\vec{t}_i$, a quad encompassing both capped ends is generated (see Fig. 2). These caps are cut out of the generated quad in a pixel shader. A normalized local coordinate system inside both caps is created, which allows determining those fragments of a quad that are outside the cap and have to be discarded.
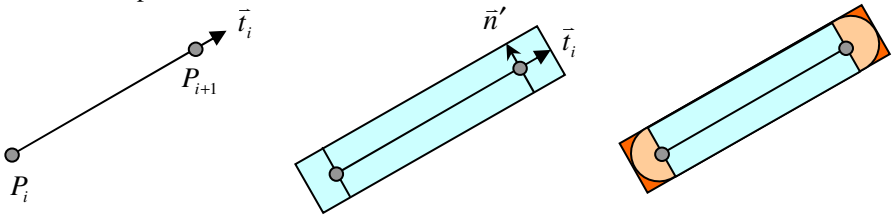
Fig. 2. Construction of rounded caps

From the camera position O, the fragment position F, and the view direction $\vec{v} = (F - O)/|F - O|$, the view ray $R = O + t\,\vec{v}$ is constructed. Given such a ray, the intersection of the ray with the cylinder spanned by the cap can be computed. Because the cylinder is always aligned with the z axis (the nadir), the 3D ray-cylinder test could be replaced by a 2D ray-circle test in the *xy* plane (see Fig. 3).

A circle with center C and radius r is defined by the equation:

$$(X - C)^2 = r^2 \qquad (1)$$

Inserting the ray R into this equation with $\bar{c} := O - C$ yields:

$$((O + t\bar{v}) - C)^2 = (\bar{c} + t\bar{v})^2 = r^2 \qquad (2)$$

Expanding this result in the quadratic equation:

$$(\bar{v} \cdot \bar{v})t^2 + 2(\bar{v} \cdot \bar{v})t + (\bar{c} \cdot \bar{c} - r^2) = 0 \qquad (3)$$

Solving for t gives the discriminant:

$$d = 4(\bar{v} \cdot \bar{c})^2 - 4(\bar{v} \cdot \bar{v})(\bar{c} \cdot \bar{c} - r^2) \qquad (4)$$

If $d \le 0$, there is none or only a single solution to the quadratic equation. This means that the ray does not hit the cap at all, or just grazes it. In this case, the fragment is discarded. Otherwise, we get:

$$t_{1/2} = \frac{-2(\bar{v} \cdot \bar{c}) \pm \sqrt{d}}{2(\bar{v} \cdot \bar{c})} \qquad (5)$$

For front faces, $\min(t_1, t_2)$ is the correct solution, for back faces it is $\max(t_1, t_2)$.
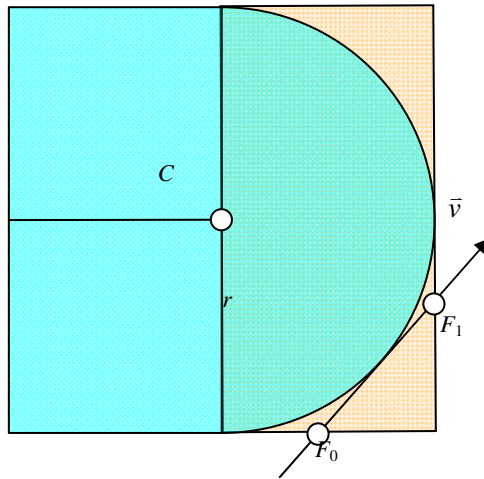
Fig. 3. Segments intersection in the local coordinates

So far, the road geometry is extruded toward infinity to generate the shadow volumes has been assumed. Since this is wasteful in terms of rasterization fill rate, the height field for limiting the extent of the shadow volumes is considered. Assuming

the terrain being partitioned into tiles, it is sufficient to extrude each line segment only within the extent of the bounding box of the tile belongs to. To accommodate this, the intersection algorithm has to be extended to handle the top and bottom sides of the extruded polyhedron: If the 2D distance between F and C is smaller than the cap radius (which can only happen for fragments belonging to the top or bottom side), F already gives the final intersection.

The algorithm as presented so far suffers from problems caused by limited numerical precision. One such problematic situation is depicted in Fig. 3: The intersection between each ray and the cylinder is computed twice, once for the front face of the bounding box (corresponding to F_0 in the figure) and once for the back face (corresponding to F_1). The ray direction is computed as $F_0\text{-}O$ and $F_1\text{-}O$, respectively. Because of small perturbations in F_0 and F_1, which are caused by the limited precision of the interpolation hardware, one of the intersection tests may report an intersection while the other one does not. This results in inconsistent output causing visible artifacts.

In order to achieve consistent results, both intersections are computed in the same shader invocation: the geometry is rendered with front face culling enabled, and analytically compute the entry point into the bounding box of the extruded road. Then both intersections between the ray and the road are computed as described above. This results in two depth values z_0, z_1 that need to be compared to the terrain depth z_t. We therefore replace the regular depth test with a custom two-sided test: z_t is read from a texture created as a secondary render target during the terrain rendering pass. If $z_0 < z_t < z_1$, then the road volume intersects the terrain geometry; otherwise, we discard the fragment.

2.2.2   Outline

To distinguish cartographic roads from the underlying terrain, dark edges have been added around roads to increase contrast. To detect edges in screen space, a 3×3 or 5×5 kernel is used to find the local maximum road intensity around each fragment. The road intensity is the road opacity for pixels which are covered by a road and 0 otherwise. The difference between the local maximum and current road intensity defines the resulting edge intensity.

2.2.3   Crossing with Overlapping

At crossroads or junctions, multiple roads overlap, resulting in visible artifacts caused by additive blending. To resolve this problem, roads are drawn into an offscreen render target without blending, in increasing order of importance.

The same approach allows for an easy integration of multi-colored roads by drawing a road multiple times with different widths and colors. This increases the geometry count proportionally to the number of colors, but since typically only a few important roads use multiple colors, this is acceptable.

3   Experiment Results

The approach in this paper has been successfully used for rendering of 3D vector data in several areas. We test the algorithm on a Core i5 (2.67 GHz, 4GB RAM) computer with Windows 7 Ultimate operation system and NVIDIA Geforce GT 435M graphic card. Fig. 4 and Fig. 5 shows that vector map have been correctly overlapped on global and regional terrain. Road features have been rendered using linear symbols with different colors (see Fig. 6), so that more information of map features, such as category and hierarchical level. Traffic features (highways and railway) have been rendered with black outlines in Fig. 7 and map features' edges have became clearer by adding outlines in contrast with Fig. 6.
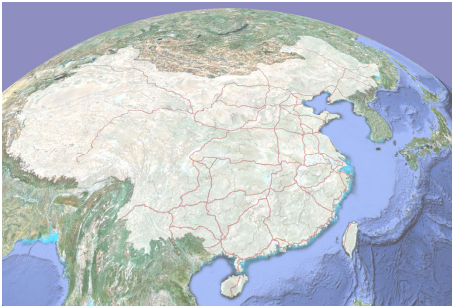
Fig. 4. Frontier and highroad   **Fig. 5.** Region boundary

Fig. 6. Traffic features with linear symbols   **Fig. 7.** Traffic linear symbols with black outlines

4   Conclusions

Under the instruction of map symbol theory, an extension of shadow volume stencil is proposed. This is a pilot study of the symbolic visualization of vector map in VGE. The smooth, outline, intersection of road feature has been address on. As a result, there are more visual forms of map feature, from lines and polygons with different colors to smooth symbols with outlines. More information of geographic entities has been transferred through different symbol parameters. According to the geo-spatial information visualization model theory, it is requisite to explore not only 3D rendering methods but also the integrated framework of vector map representation in VGE.

References

1. Nian, L.G.: Geographic analysis-oriented Virtual Geographic Environment: Framework, structure and functions. Sci. China Earth Sci. 54(5), 733–743 (2011)
2. Koch, A., Heipke, C.: Integrating 2D topographic vector data with a Digital Terrain Model – a consistent and semantically correct approach. ISPRS Journal of Photogrammetry and Remote Sensing 61(1), 23–32 (2006)
3. Ning, L.: Ground Feature and Terrain Merging Arithmetic and Accuracy Evaluation. Institute of Surveying and Mapping Information Engineering University, Zhenzhou (2004) (in Chinese)
4. Kang, L., Zhao, J., Song, H.-C., et al.: Terrain Matching for Three-dimensional Visualization of Two-dimensional GIS Vector Data. Computer Science 36(11), 262–265 (2009) (in Chinese)
5. Wartell, Z., Kang, E., Wasilewski, E.: Rendering Vector Data over Global, Multiresolution 3D Terrain. In: IEEE TCVG Symposium on Visualization, pp. 213–222 (2003)
6. Schneider, M., Guthe, M., Klein, R.: Real-time Rendering of Complex Vector Data on 3D Terrain Models. In: 11th International Conference on Virtual Systems and Multimedia – VSMM, pp. 573–582 (2005)
7. Zou, W., Fang, J.-Y., Liu, J.-G.: Research about visualization of hybrid multi-resolution terrain and vector data. Journal of System Simulation 18(s1), 324–325 (2006) (in Chinese)
8. Agrawal, A., Radhakrishna, M., Joshi, R.C.: Geometry-based Mapping and Rendering of Vector Data over LOD Phototextured 3D Terrain Models. In: 14th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, WSCG 2006, pp. 1–8 (2006)
9. Yang, B., Kang, L., Wu, L., et al.: Research on the Multi-Resolution Modeling and Three-Dimensional Display of GIS Vector Data. Computer Engineering & Science 30(9), 73–76 (2008) (in Chinese)
10. Cao, X.: The Organization Management and 3D Visualization of Vector Data based on Geographic Information Grid. Institute of Surveying and Mapping Information Engineering University, Zhengzhou (2009) (in Chinese)
11. Kersting, O., Döllner, J.: Interactive Visualization of Vector Data in GIS. In: 10th ACM International Symposium on Advances in GIS, pp. 107–112 (2002)

12. Schneider, M., et al.: Real-time rendering of complex vector data on 3D terrain models. In: 11th International Conference on Virtual Systems and Multimedia –VSMM, pp. 573–582 (2005)
13. Chen, H., Tang, X., Xie, Y., et al.: Rendering Vector Data over 3D Terrain with View-Dependent Perspective Texture Mapping. Journal of Computer-Aided Design & Computer Graphics 22(5), 753–761 (2010) (in Chinese)
14. Li, R., Zheng, W.-T.: Real–time Rendering High-Quality Vector Data on 3D Terrain. Journal of Computer-Aided Design & Computer Graphics 23(7), 1107–1114 (2010) (in Chinese)
15. Wang, L.: Study on Urban Environment Procedural Modeling Techniques. Zhejiang University, Hangzhou (2009) (in Chinese)
16. Bruneton, E., Neyret, F.: Real-time rendering and editing of vector-based terrains. Computer Graph. Forum 27, 311–320 (2008)
17. Schneider, M., et al.: Efficient and accurate rendering of vector data on virtual landscapes. Journal of WSCG 15(1-3), 59–65 (2007)
18. Crow, F.: Shadow algorithms for computer graphics. In: SIGGRAPH 1977, pp. 242–248 (1977)
19. Gao, J.: Visualization in Geo-Spatial Data. Engineering of Surveying and Mapping 9(3), 1–7 (2000) (in Chinese)
20. Vaaraniemi, M., Treib, M.: High-Quality Cartographic Roads on High-Resolution DEMs. Journal of WSCG 19(1-3), 41–48 (2011)

Modeling and Simulation of Nearshore Waves

Jianhua Xu[1,2], Hao Gu[1], Fengju Kang[1,2], Huizhen Yang[1,2], and Sunli Wang[1,2]

[1] Marine College of Northwestern Polytechnical University, Xi'an 710072, China
[2] National Key Laboratory of Underwater Information Process and Control, Xi'an
flyingcondor@163.com

Abstract. This paper presents a new modeling method for simulating nearshore waves. Waves in the process of moving to the nearshore are more complex than to the deep-sea. Nearshore wave modeling in addition to the need to consider gravity and wind effect, but also consider the impact of the seafloor topography. We construct the mathematical model of the seafloor topography. Then, the waveform data generate while constructing the seafloor topography data. Finally, we achieve real-time simulation of nearshore waves based on parametric surfaces. This experiment proved that this method can quickly realize simulation into the nearshore waves.

Keywords: nearshore waves, modeling, simulation, OpenGL.

1   Introduction

Simulation of nearshore waves has an important significance in the field such as digital coastal city, games, movies or sailing simulator. In addition to considering the wind and gravity in the nearshore waves simulation, we should also consider the impact of the seafloor topography. When the waves from deep-sea move to the coast, the wave form will be changed. The simulation model of nearshore waves have been more mature in domestic and overseas. Darwyn R. Peachey built the wave model based on height field, but can not simulate the peak of the curl; Ts'o corrected sine wave model based on B-spline function, but the realization is much more complex; Gonzato J.C modified the model of Aain Fournier by three functions, but the calculation is too complex. In this paper, we plan to simulate wave with physics-based modeling method and take into account the influence of seabed topography on the waveform. The program construct seafloor topography data and the waveform data at the same time, which is used for real-time rendering of nearshore waves. The simulation show that this method can get high-speed rendering and realistic effect.

We discussed the theory and modeling of nearshore waves in section 2. In section 3 we elaborated on the data generation method of the seafloor topography, and analyzed how the nearshore seabed topography affect the sea waveform. In Section 4 we realized the simulation of nearshore wave based on OpenGL and the conclusion and future work was followed in section 5.

2   Nearshore Wave Theory

2.1   Nearshore Waves

Nearshore waves are different from the waves of the ocean deep water area due to the impact of seafloor topography. The nearshore waves are sharp near the crest and flat near the trough. The waves also exhibit the phenomenon of refraction parallel to the coastline, showing the apparent regularity. In order to be similar to the real nearshore waves, we consider not only the wind and gravity but also the friction of the seabed. When the waves moved from the deep-sea to the coast, the wave propagation direction, wavelength and the height of wave are changed because of seabed friction.

2.2   Wave Model

For a given point on the surface of the sea, the point coordinates is $X_0(x_0, z_0)$ when t=0. At time t, amplitude of R, frequency of ω wave through the point, then the point coordinates is as follows：

$$X = X_0 + (K / k) \cdot R \sin(K \cdot X_0 - \omega t),$$
$$Y = Y_0 + R \cos(K \cdot X_0 - \omega t),$$

(1)

where K is the wave-vector, k is the wave number.

The wave number is described as follows:

$$k = 2\pi / L \tag{2}$$

where L is the wave length. We can see that the wave surface is a cycloid. Cycloid can be described as follows: A radius of a circle of $1 / \kappa$, we now start from the center of the circle to take a distance of r points, with the roll-forward of the circle, then the distance r points formed by the trajectory is a cycloid, similar to the waveshape. When the t=0, $y_0 = 0$:

$$X = -\alpha / K + (K / k) R \sin(\alpha)$$
$$Y = R \cos(a), a = -KX_0$$

(3)

where wave height $H = 2R$, wave length $L=2\pi/K$, circle $T=2\pi/\omega$, wave speed $c = L/T = \omega/K$, If the phase is 0 when $X_0 = 0$, the initial phase can be defined as $\varphi = kX_o - wt$.

The different waveforms with different different KR values are given in Figure 1. The waveform is almost consistent with sine wave when KR=0.2. Waveform is a cycloid when KR=0.7 or KR=1.0. Waveform cross when KR=1.2. In fact, the reasonable KR value should be 1, this means $H/L=KR/\pi=1/\pi$. The ratio $\delta=H/L$ is called the wave steepness.

KR=0.2

KR=0.7

KR=1.0

KR=1.2
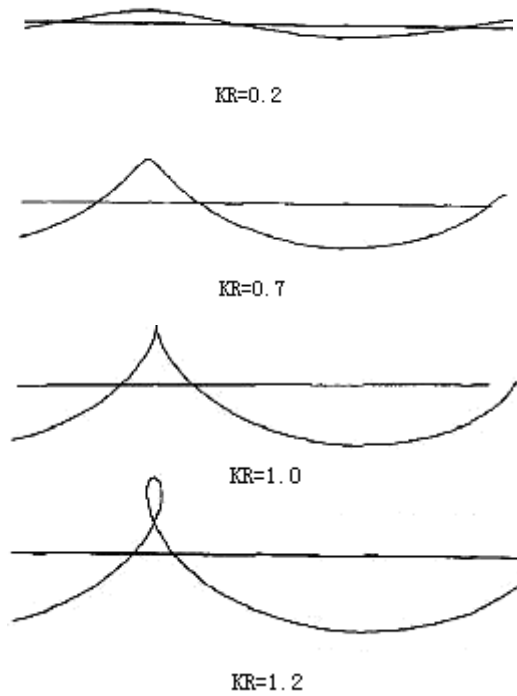
Fig. 1. With the change of KR value of waveform

Simulation wave cycle is generally from 0.5 to30 seconds. Wind is the main perturbation force, and gravity is restoring force. In order to simulate different size of the wind on the wave effects, the formula of the waveform was modified, and specific approach is to modify the waveform phase：

$$\phi = K \cdot X_0 - \omega \cdot t - \lambda \cdot^\triangle y \cdot^\triangle t \tag{4}$$

where λ is proportional coefficient，$^\triangle y = y - y_0$，$^\triangle t$ related to time. As can be seen from the type，in the wave near the wave particle rotation speed is accelerated, and in the valleys near the wave particle is slow. By modifying the λ value can be obtained with different waveforms，as shown in Figure 2. Because t has both positive and negative, so this effect will not be accumulated. Through this method can simulate the volume wave.

We can see that using the above formula created waveform is regular. In order to avoid too much rules of waveform, reflect the stochastic properties, we can learn from the linear superposition method based on ocean wave spectrum. Formula (1) is modified to form as shown below:
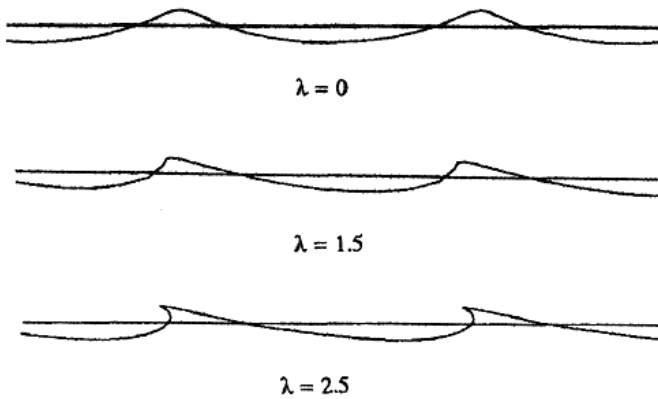
Fig. 2.   With the change of value of waveform

$$X = X_0 + \sum \frac{K_i}{k_i} R_i \sin(K_i \cdot X_0 - \omega t - \varepsilon_i)$$
$$Z = Z_0 + \sum R_i \cos(K_i \cdot X_0 - \omega t - \varepsilon_i)$$

(5)

The above basic model is what we need.

3   Implementation Details

3.1   Seafloor Topography Data Generation

The seabed topography changes the refraction and reflection of waves, so we must know the information about it. The contour of the seabed topography will directly affect the wave propagation direction, wavelength and wave height. In order to carry out the drawing of seafloor topography and waves, it is necessary to obtain three-dimensional seabed topography data and contour information.

We have constructed a 512x512 grid at xz plane,and the slope seafloor height coordinates   is from (511,-100,z) to (0,0,z). The line throughed the (512,y) point is considered the infinity, and   the seabed terrain height y is from ys=-100 to ye=0.

We sampling n points on the x-axis and y-axis respectively. So, the step size of the x-axis and y-axis can be decribed as follows:

$$X_{step} = (X_{end} - X_{start})/n,$$
$$Y_{step} = -(Y_{end} - Y_{start})/n,$$

(6)

So, each point coordinate values can be calculated as follows:

$$x= xs+i* Xstep, \quad y=ys+i*Ystep \tag{7}$$

By calculating, we can get the values of x, y, z of any point.

3.2   Wave Propagation Direction

The classical theory predicts that the motion track of wave in medium depth waters will become oval. When the waves near the coast, the final track will becomes a straight line segment. Viewing from Figure 3, we can see that wave motion orbits change with different depth. The sea waveform changes as shown in Figure 4.

Fig. 3. Wave motion orbits

Fig. 4. Waveform

3.3   Wave Length

Based on the principle of conservation of wave number, the depth of the seabed does not affect the wave cycle, only affects the wavelength, wave speed and wave propagation direction. When the depth of the sea is bigger than half the wavelength, the dispersion relation of wave is described as follows:

$$\omega^2(k) = gk \tag{8}$$

where k is wave number. When the depth of the sea is less than half the wavelength, the dispersion relation of wave is described as follows:

$$\omega^2 = gk \tanh(kd) \tag{9}$$

where k is the wave number; d is the nearshore water depth.

3.4   Wave Height

Nearshore waves follow the law of conservation of energy if there is no friction effect. Waves height changes as follows :

$$\frac{F}{F_\infty} = \frac{1}{\sqrt[4]{\tanh(k_\infty d)}} \tag{10}$$

where F is the nearshore wave amplitude; F_∞ is the wave amplitude in deep-sea ; k_∞ is the wave number in deep-sea. Taking into account the friction of seafloor topography, we add an attenuation function related to water depth closely:

$$\frac{F}{F_\infty} = \frac{f(d)}{\sqrt[4]{\tanh(k_\infty d)}} \tag{11}$$

where   f(d) is the friction function.

3.5   Simulation Flow
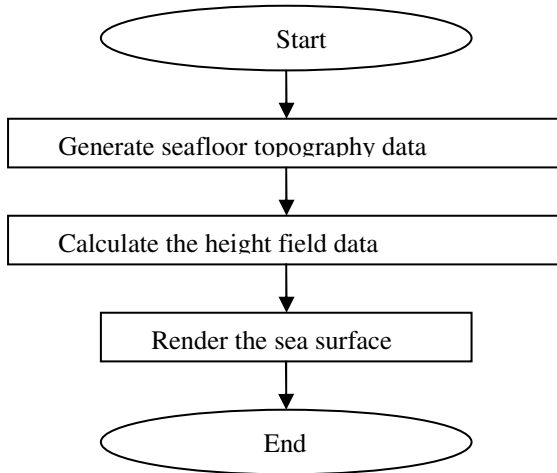
Fig. 5. Simulation flow of nearshore waves

4   Simulation Results

In order to verify the effectiveness of the proposed algorithm, we have simulated ocean surface. The experimental environment is VC++ and OpenGL running on Intel Pentium 4 with NVIDIA Geforce GT520 graphics card. The frame rate is 132 fps. Figure 6 shows the frame captured from our system.

Fig. 6. Nearshore waves

5   Conclusion and Future Work

According to the existing observations and research results, we constructed nearshore wave model based on wave functions and simulated wave with OpenGL. The results show that this method can get realistic effect and fast rendering speed. In the future work, We will research the superposition multiple waves and beach broken wave simulation.

References

1. Foumier, A., Reeves, W.: A Simple Model of Ocean Waves. In: Proceedings of SIGGRAPHICS, pp. 75–84 (1986)
2. Chen, J.X., Da Vitoria Lobo, N., Charles, et al.: Real-time Fluids Simulation in a Dynamic Virtual Environment. In: IEEE Proceedings of Computer Graphics and Applications, pp. 51–61 (1997)
3. Nie, W., Kang, F., Chu, Y., Yang, H.: Linear Ocean Wave Simulation. Journal of System Simulation 7, 1037–1039 (2005)
4. Peachey, D.R.: Modeling waves and surf. Computer Graphics 20, 65–74 (1986)
5. Jens, et al.: Towards Real-time Visual Simulation of Water Surfaces, Stutgart, Germany, pp. 211–218 (2001)
6. Tessendorf, J.: Simulating ocean water. In: SIGGRAPH 2004 Course Notes (2004)
7. Yan, L., Li, S.: Real-time generation of ocean wave surface. Journal of Computer Aided Design & Computer Graphics 12, 715–719 (2000)
8. Chu, Y., Kang, F.: Visual Simulation of Ocean Waves on the beach. In: Symposium on Navigation Technology, Xi'an, China, pp. 107–114 (2004)

3D CG Model and Virtual Space of Court Noble House "Reizei-ke"

Wang Sheng[1], Susumu Nakata[2], and Satoshi Tanaka[2]

[1] Graduate School of Science and Engineering, Ritsumeikan University,
Nojihigashi 1-1-1, Kusatsu, Shiga, Japan
`cm013076@ed.ritsumei.ac.jp`
[2] College of Information Science and Engineering, Ritsumeikan University,
Nojihigashi 1-1-1, Kusatsu, Shiga, Japan
`{snakata,stanaka}@media.ritsumei.ac.jp`

Abstract. The residence of "Reizei-ke" (Reizei family) is the only existing court noble house in Japan. This paper presents a high quality 3D CG model of Reizei-ke, and a virtual space built by a 3D game engine. Users are able to enjoy a cosmic voyage which demonstrates a scene of ancient Court nobles of Japan.

Keywords: 3D CG model, Virtual space, 3D game engine, Digital Museum.

1 Introduction

Kyoto used to be the capital city of Japan for over a thousand years, even now it is the center of Japanese traditional cultures. To bequeath the valuable tradition and culture of Kyoto to subsequent generations, government, universities, research institutions and other organizations are working on activities to record and preserve various materials related to Kyoto in digital media [1] (Fig.1). Our group has been working on one of the projects to build a digital archive of tangible and intangible cultural properties of Kyoto and Japanese cultures in Ritsumeikan University for establishing "Digital humanities center for Japanese arts and cultures". [2] (Fig.1). We mainly focus on the visualization of cultural properties of Kyoto. So far, we have

Fig. 1. The Digital Museum of Kyoto (left) and the web contents of digital humanities center for Japanese arts and cultures (right)

accomplished to develop automatic generation system of 3D ancient Kyoto city models that consist of tradesmen's house (Machiya-house), Buddhism temples, Shinto shrines, Samurai residences, etc. . Besides we are also creating high quality 3D CG models for some symbolic buildings to be embedded into the automatic generated scene [3] (Fig.2), for example, the Minamiza-theater.

This paper reports the progress of our recent modeling of the only existing Court noble house "Reizei-ke". In addition, we are constructing a virtual space controlled by a high-end 3D game engine to contribute more to the digital museum of arts and cultures of Kyoto.

Fig. 2. 3D ancient Kyoto city models created by our automatic generation system

2   Court Noble "Reizei-ke"

The history of the Court noble family "Reizei-ke" can be traced back to the Heian period in Japan. Nagaie, the sixth son of exalted and powerful Michinaga Fujiwara established the "Mikohidari-ke". Then after about 200 years, one of the descendants named Tamesuke renamed his family to "Reizei-ke". Till now, about 700 years has

Fig. 3. Court noble "Reizei-ke" and a scene of annual event

passed. Since the days of Nagaie, the family has been an expert in "Waka" (a three one style Japanese poem). "Reizei-ke" is called "the head family of versification". In the modern period, the family established the Reizei-ke Shiguretei archive which saves over one thousand collection of books, paintings and calligraphies that containing 5 National Treasures and 45 important cultural properties. The residence of "Reizei-ke" is located in the north of Kyoto Imperial Palace. It is the only existing Court noble residence of Japan and was designated as important cultural property in 1979. Now the residence is preserved well and not open to the public for most time of the year [4].

3   3D CG model of "Reizei-ke"

Fig. 4. A blueprint of "Reizei-ke" repair work reports
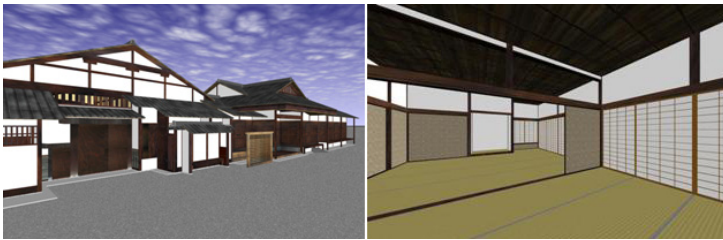
Fig. 5. The old version model of "Reizei-ke"

Table 1. Properties of the 3D CG model of "Reizei-ke"

Item	Value
Form	*.mb(Default Maya form)
Size(No textures)	2115 KB
Triangles	35340
Vertices	118275
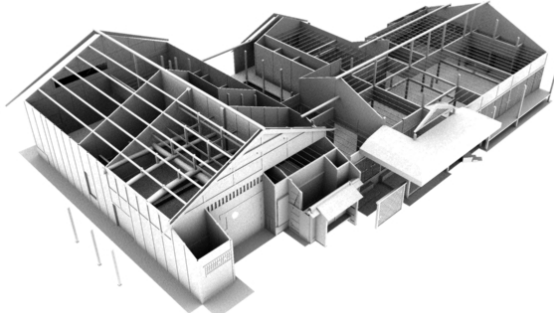
Fig. 6. Structure of the model of "Reizei-ke"

The 3D CG model of "Reizei-ke" was created by our group in accordance with a group of construction report of repair work from 1994 to 2000 (see Fig.4) [5]. The modeling tool was Form-Z (a 3D CG software developed by AutoDesSys). The model is accurate and includes most of the parts described in the repair work report. However, the materials and textures of the model are too simple to be used as contents of Digital Museum (see Fig.5). So we refined the materials and textures of our existing model. We also added some models which are not described in the repair work reports (exp. Courtyard wall, decorations, vegetation and so on). We used Maya 2011(a high end 3D CG software) developed by Autodesk as the modeling tool.

3.1   Materials and Textures

The purpose of our current project is to make the virtual model as real as possible. So we used the photographs of the real "Reizei-ke" (Fig.7) wherever possible. The work of photographing was carried out at special opening of "Reizei-ke" in Nov. 2011. However, since access to the internal area was not allowed, proximate textures are used for internal scenes, assuming that the Japanese style houses have common features.

Fig. 7. Textures are extracted from the photographs

For materials, currently we have not done excessive adjustment. Observing the final rendering results, use of the default settings of the rendering software seems to be satisfactory. Since the old-time Japanese-style houses are made of wood, there is no need to incorporate effects of high light reflection.

3.2   Additional Models

The old version of our model does not contain any decorations. The new version is enriched by some ornaments (see Fig.8). The photographs of "Reizei-ke" show a lot information on decorations. But there are still many decorations that have not been modeled especially in the private areas of "Reizei-ke".

The final rendering results are showed in Fig.9.

Fig. 8. A part of Newly created decorations
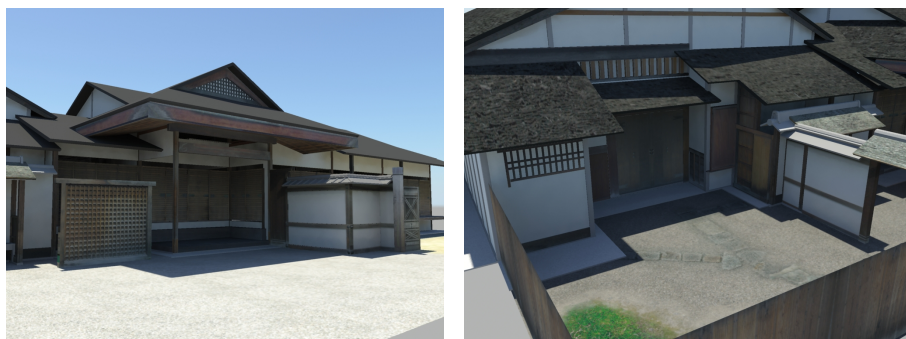
Fig. 9. 2 CG pictures of the new model rendered by Maya 2011

4   Virtual Space Created with 3D Game Engine

Real-time walking through with vivid lighting/shadowing effects requires high performance of 3D CG engine. Therefore, we adopt UDK (Unreal Development Kit) developed by Epic Games. UDK is a popular 3D game engine with superior performance of light and shadow, And it also has good compatibility with models created with Maya 2011.

In order to facilitate the partial adjustment, the 3D CG model of "Reizei-ke" was loaded to the virtual space in several parts. There is only one-directional light used to
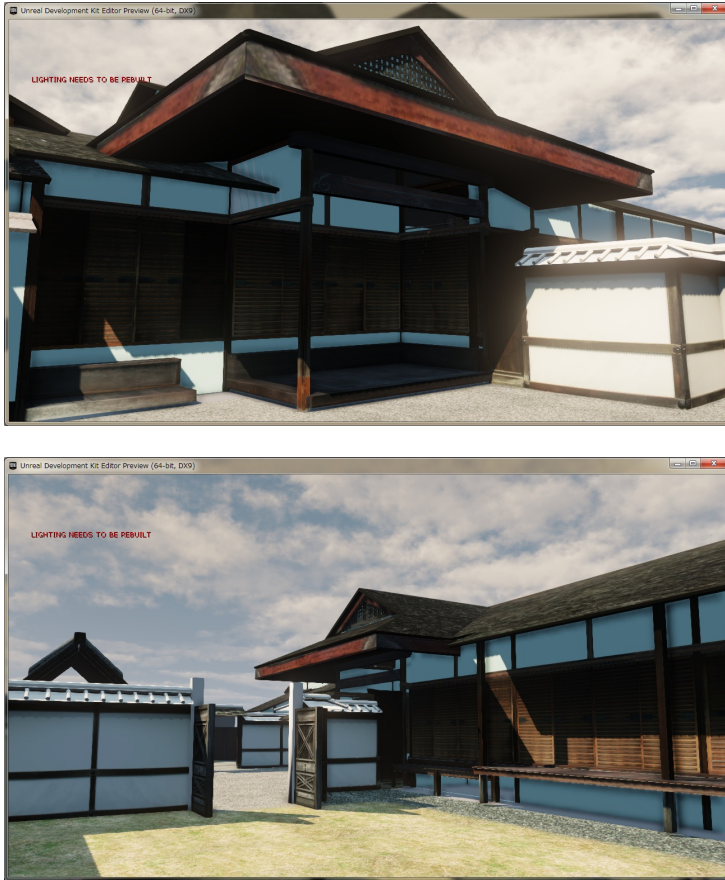
Fig. 10. Snapshots of the virtual space of "Reizei-ke"

simulate the sunlight. Although 2 or 3 light sources are general in recent CG scenes, the UDK library can express sky light well to make up for insufficiency of light sources. At present, our virtual space employs a first-person perspective and the viewpoint moves according to user's requirement through mouse operation. Sunny daytime with a little cloud is implemented. Collision detection also works. Average performance in the walk through maintains above 60 FPS (Intel i7 3.4GHz with AMD Radeon HD 6670).

5   Conclusion

This paper has reported status of our modeling 3D virtual space of "Reizei-ke", which is the only existing Japanese Court noble residence. Our model enables realistic and/or quick virtual tour of the Court noble residence, implementing decorations of

rooms, texture mapping with real photographs, etc. A 3D game engine enables real-time walking through in the created scene.

In our work, the development of the virtual space is still in a preliminary stage. In the future, we plan to add our original user interface for walking-through operation and to display information on the Court noble residence. Besides, a human characters and more decorations will be added to simulate daily scenes of Court nobles in ancient days. Better performance in the walking through in cases of higher-quality rendering will also be achieved in our future work.

References

1. Kyoto Digital Museum, `http://www.pref.kyoto.jp/it/10500040.html` (accessed May 18, 2012)
2. Digital humanities center for Japanese arts and cultures, `http://www.arc.ritsumei.ac.jp/lib/GCOE/` (accessed May 18, 2012)
3. Isoda, Y., Tsukamoto, A., Kosaka, Y., Okumura, T., Sawai, M., Yano, K., Nakata, S., Tanaka, S.: Reconstruction of Kyoto of the Edo Era Based on Arts and Historical Documents: 3D Urban Model Based on Historical GIS Data. International Journal for Humanities and Arts Computing 3, 21–38 (2009)
4. Reizei, T.: -Reizei-ke No Rekishi- (The history of "Reizei-ke"). Newspaper publishing company of Asahi (1981)
5. Honjo, A.: The modeling of Court noble house for Kyoto street generation (2009)

Fluid Motion Vector Calculation
Using Continuity Equation Optimizing

Maomao Wu and Hongyan Quan

No. 3663 Zhongshan North Road, Shanghai,
East China Normal University Science Building B219
`hyquan@sei.ecnu.edu.cn`

Abstract. It is very important to calculate fluid motion vector for natural landscape modeling of virtual reality interaction. This paper presents a new method of landscape fluid motion vector calculating. First, we use Plessy operator to extract feature points from two images and to calculate the match points using the area correlation matching method. Then the linear interpolation method with the shortest distance is used to interpolate the calculated motion vector to obtain dense fluid motion vector result. At last, we use the fluid continuity equation to optimize the dense fluid motion vector field to obtain dense and more accurate fluid motion vector calculation results. Further experimental results show that this method has the characteristic of simple and accurate. It is a valid method of fluid motion calculating and be used in the application of fluid simulation and virtual reality study.

Keywords: fluid, motion vector, continuity equation, virtual reality.

1   Introduction

It has a very wide range of applications for natural landscape simulation. With technique developing further, people proposed the modeling method of argument reality. The video contains more abundant texture information of the fluid particles and this can ensure the realism of virtual scene. In order to achieve the seamless integration and interactivity with objects of the scene, fluid motion vector calculation of fluid in the natural landscape is essential. The fluid in the natural landscape can be seen everywhere. Fluid motion vector calculation is to obtain the magnitudes and directions from a microscopic perspective and obtain the law of fluid motion from a macro perspective. Due to the special nature of fluid motion, there is occlusion, illumination changes in the procedure, certain difficulty exist in the study, and the techniques using video has not yet studied thoroughly.

Compared to the existing methods of fluid motion vector calculating, the proposed method has the following characteristics:

- It calculates the motion vector with a simple method with character of accuracy and efficiency. It takes video by an ordinary capturing device.

- It is practically efficient taking the measure of continuity equation optimizing, and this can ensure the motion vector having physical characteristics of fluid motion.
- Due to its simple and efficiency, it can be used in the virtual environment modeling and other practical applications.

This paper proceeds as follow. Section 2 briefly summarizes work related to ours. Section 3 describes the initialization methods of fluid motion vector, describes the linear interpolate method with the shortest distance of fluid motion vector, and describes the optimization methods using the fluid continuity equation. In section 4 we will show experimental studies for verifying the proposed method and the performance of comparing with existing method. It is concluded in section 5.

2   Related Works

With the development of computer vision techniques, there appear several methods of calculating fluid motion [1,2]. Lucas and Kanade [3] applied the pyramid model which reduces the limitation of location window and promoted the accuracy of estimation. But slight brightness change would decrease the accuracy of method based on brightness consistency. Thomas analysis the law of weather maps by adding the continuity equation to motion model [4]. Ashish use the Navier-Stokes partial differential equations of fluid mechanics and the continuity equation to analysis complexity of fluid motion [5,6]. In the study, some people regard the fluid as particles and obtain the motion laws by extracting the motion of particles. Zhang, et al use irregular movement of the gas molecules to study and obtain the motion vector results [7]. Li, et al made the guidelines of cell division and obtain the motion vector field by evolution from the criteria [8]. This method has the character of time and space efficiency, but the results of the calculation is prone to multi-peak problems [9]. In order to obtain the more accurate results, other additional physics constraints are often being considered, but they prone to be more complex physical model.

   This paper proposes a new method of calculating the motion vector of the fluid in natural landscape. First, Plessy operator is employed to extract feature points from the two images and the area correlation matching method is used to obtain the sparse correspondence. And then use the shortest distance to interpolate to calculated dense motion vector. In order to obtain more accurate results, use fluid continuity equation to optimize the dense fluid motion vector field.

3   Calculation of Fluid Motion Vector

3.1   Initializing of the Fluid Motion Vector

Since fluid belong to rich texture, the fluid motion can bring out occlusion problems. It is difficult to keep the consistent brightness for fluid particles in the movement of the adjacent two frames. The motion vector calculation has certain difficulty. There

are several kinds of correlation methods, area based method and feature based method are two common used methods. In the study we take advantage of the measure of density correlation between two areas to initialize the motion vectors. The method of density correlation has the character of simple and validity.

Zhang has proposed a correlation based method of standard variance [10]. In the study we use this method to study. There are three steps in the initializing. We first extract the feature points using Plessy operator. And then use correlation calculating method between two areas to initialize the motion vectors. At last we eliminate the ambiguous results and obtain the more accurate initializing results.

In order to obtain more initializing results, we adopt the loose threshold to study. In the correlation we adopt the window size of 7×7 to calculate and obtain the satisfied initializing results.

3.2   Linear Interpolation within the Shortest Distance

After initialization we can obtain sparse fluid motion vectors, in order to get dense motion vectors results, we adopt the measure of interpolation. There is the characteristic in fluid motion that particles of certain range have the similarity movement. We use linear method to interpolating motion vector within the shortest distance.

The main idea of the linear interpolation within the shortest distance is that for any two particles which have been calculated the motion vectors we use linear interpolation method to obtain a series of motion vectors of the particles that on the line connecting the two particles.

The linear interpolation with the shortest distance algorithm is described as follow:

(1)   For any particle which motion vector has been calculated, set its flag to 1. Then establish a set A including all the particles which flag is 1.

(2)   To any particle element c of A, set up a new set S including the particle elements surrounding 100×100 region and their flags is already set. If S is not empty, go to step (3), else go to step (4).

(3)   If b is the nearest particle element to c in the set S. If bc has not been interpolated, then interpolate it and remove b from S, remove c from A, go to step 4.

(4)   If set A is empty, go to step (5), else go to step (2).

(5)   To determine whether to meet the end conditions? If yes, go to step (6), else go to step (1).

(6)   End of algorithm.

In the study, in order to obtain more dense motion vector calculation results, we use the condition that the particles number of having been calculated motion vector up to ninety percent of the total numbers as the condition to terminate the interpolation algorithm. If the resolution of the image is $Width \times Height$, the total number of particles is $Width$ by $Height$, and other fewer motion vectors of particle are average by the motion vectors among local area of 5×5.

Further experiment results verified that the validity of the linear interpolation of the shortest distance method. After motion vector initialization and linear interpolation with the shortest distance, we get dense fluid motion vectors.

3.3   Optimization Method of the Fluid Motion Vector

It can be measured in two ways whether the results of fluid motion vector are accuracy or not. The first way is to measure from consistency of the appearance. The second one is to test with complying with the continuity of the movement. In order to meet these two requirements and to ensure the accuracy of the fluid motion vector calculation, we use the continuity equation to optimize the fluid motion, which can reflect the characteristics of fluid movement. The continuity equation can reflect the physical laws of fluid movement. We use the physics-based approach to optimize the fluid motion vectors can get more accurate results [11]. We use the continuity equation to optimize fluid motion vector, which can reflect the movement of fluid characteristics. The continuity equation of fluid motion shows in formula (1) [12].

$$\int_{\Omega} \partial_t I dA + \int_{\partial\Omega} If \cdot ndL = 0 \ . \tag{1}$$

$\partial_t I$ denotes the intensity derivative of image I on the time direction, $\partial\Omega$ denotes the boundary of Ω, n denotes the external vector of $\partial\Omega$. According to the divergence theory and mathematical derivation, formula (1) can be written in the form of the following equation [12].

$$I_t + uI_x + vI_y + Iu_x + Iv_y = 0 \ . \tag{2}$$

I denotes the intensity of the particle , u and v denotes the component of motion vector along x-axis and y-axis respectively, I_x denotes intensity derivatives of the particles along x-axis, I_y denotes intensity derivatives of the particles along y-axis, u_x denotes derivatives of u along x-axis, v_y denotes the derivatives of v along y-axis. I_t denotes the derivatives of the intensity on time direction.

Since the error of fluid motion vector, the value of right side in formula (2) cannot be accurate to 0, some errors can be allowed, so we use formula (3) to optimize.

$$\left| I_x u + I_y v + Iu_x + Iv_y + I_t \right| \leq T \ . \tag{3}$$

T denotes a threshold. If the particles in the two adjacent frames satisfy formula (3), we regard it satisfying the continuity equation.

Fluid motion vector optimization algorithm can described as follows:

（1） For any fluid particle, if the motion vectors of particles surrounding have been calculated. We use formula (3) to test, if it is satisfied formula (3), then set flag to 1, else set to 0.
（2） Whether all particles are tested, if yes, go to the next step, else go to step (1).

(3) For any particle, if its flag is 0, then calculate its motion vector using the results of surrounding particles, according to the fluid equation of formula (2), the calculation is divided into four cases (described below). After motion vector calculating, set its flag 1.

(4) Whether meet the end condition, if yes, go to the next step, else go to step (3).

(5) Obtain the motion vector result of the particle which flag is 0 using the weighted average interpolation method.

(6) End of algorithm.

In the study, we use the threshold 10 in formula (3) and carry out five times in optimization and obtain more satisfied results, termination condition is the iterative number is 5. In step 3, if the motion vector of the particle has not been calculated, we use continuity equation to calculate the motion vectors. There are four cases, just as shown in Figure 1.
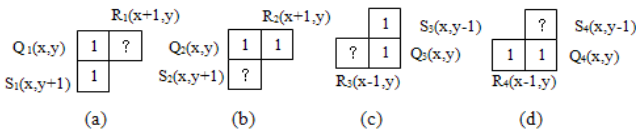
Fig. 1. Four cases in optimization of fluid motion vector

In figure 1 (a) to (d), l denotes the particle has satisfied the continuity equation and the flag value is 1. '?' denotes the particle does not obtain the motion vector and the flag value is 0.

For the case of 1(a), the motion vector along the x-axis direction $u(x+1, y)$ of the particle $R_1(x+1, y)$ can be calculated by the formula (4).

$$u(x+1, y) = u(x, y) - (I_t + I_x u + I_y v + I v_y)/I . \qquad (4)$$

I_x, I_y, I_t and v_y can be calculated by forward differential.

For the case of 1(b), the motion vector along the y-axis direction $v(x, y+1)$ of the particle $S_2(x, y+1)$ can be calculated by the formula (5).

$$v(x, y+1) = v(x, y) - (I_t + I_x u + I_y v + I u_x)/I . \qquad (5)$$

I_x, I_y, I_t and u_x are calculated by using the method of forward differential.

For the case of 1(c), the motion vector along the x-axis direction $u(x-1, y)$ of the particle $R_3(x-1, y)$ can be calculated by the formula (6).

$$u(x-1, y) = (I_t + I_x u + I_y v + I v_y)/I + u(x, y) . \qquad (6)$$

I_x, I_y, I_t and v_y can be calculated by backward differential.

For the case of 1(d), the motion vector along the y-axis direction $v(x, y-1)$ of the particle $S_4(x, y-1)$ can be calculated by the formula (7).

$$v(x, y-1) = (I_t + I_x u + I_y v + I u_x)/I + v(x, y) \ . \tag{7}$$

I_x, I_y, I_t and u_x can be calculated by backward differential.

Because image intensity in the local area changes a little in the two adjacent frames of video, the motion vector magnitude and direction are similar in the area. After optimization the motion vectors are more accurately and the number of particles absent from motion vector is relatively small, so we can calculate the motion vector of them using weighted average method.

The main idea of weighted average method is that given a particle $N(x, y)$ that does not obtain the motion vectors, consider its $n \times n$ local area, assuming that the components of motion vector of the particle $N(x, y)$ along the x-axis direction and y-axis direction can be calculated by the formula (8) respectively.

$$V(x, y) = (\sum_{i=1}^{n \times n} weight(i) * V(i)) / \sum_{i=1}^{n \times n} weight(i) \ . \tag{8}$$

$weight(i)$ denotes the weight of particles in local $n \times n$ region. In the study, we take the reciprocal of the distance between center particle $N(x, y)$ and current particle i as the $weight(i)$, the distance is closer and the weight is higher. $V(x, y)$ denotes the components of motion vectors along the x-axis or y-axis directions. In study we take the region size of 5×5 to obtain a more satisfied motion vector results.

4   Experiments

4.1   Validity of New Method

In order to verify the validity of the algorithm, we implement the experiments by using the videos in DynTex database [13]. The hardware environment is micro-computer with 2.1GHz CPU and 2GB Memory. The operating system is Windows XP.

The result is shown in figure 2. In each result of figure 2, the first one is the result of initialization motion vectors; the second one is the result of the linear interpolation; the last one is the results of optimization. Due to the large number of motion vector, we just display one per 7 particles for the second and the last one.

It can be seen from each result in figure 2, the result of initialization have gain relatively sparse particles and motion vectors are relatively correct that consistent with the trend of fluid motion. The result of the linear interpolation of the shortest distance shows the fluid motion vectors are dense, and it is highly consistent with the law of fluid motion and it clearly reflects the details of fluid motion. The motion
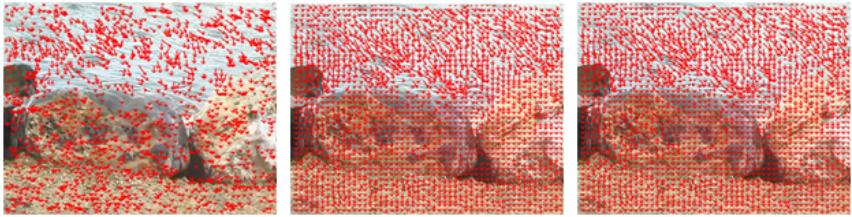
(1)The 80th frame of "64adg10"
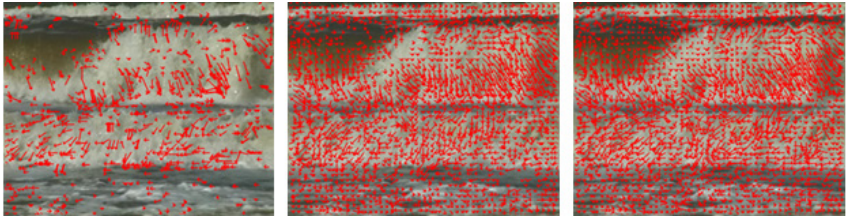
(2) The 143th frame of "649dd10"

Fig. 2. Results of motion vectors

vector optimization results can reflect the characteristics and trends of the fluid motion. It can be seen that the magnitude of motion vectors in calm fluid area are relatively small, the direction of motion vectors in local area change gently. But the magnitude of motion vectors in turbulent fluid area are relatively large, the direction of motion vectors in local area change acutely. These characteristics highly consistent with the law of motion of the fluid and keep the trends of the fluid movement.

4.2   Comparing with the Existing Methods

For further prove the effectiveness of our approach. We compare the existed method with the proposed method. In this paper, we take advantage of image in Yosemite [14] to the experimental work. Comparing with the Liu [15], we set the normalize parameter α as 1; down sampling rate r as 0.5. Comparing with the Nilanjan [16], the parameter of non-negative weights α is 0.15; the layer number of pyramid is 6.

(1) Comparing with the existed method
    The first one is the result of Liu [15], the second one is the result of Nilanjan [16] and the last one is from our method.
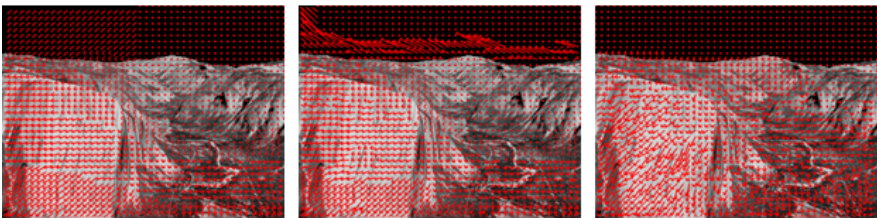
Fig. 3. The motion vector result of 1st frame in Yosemite cloud-free image sequence

In figure 3, the background cloud of exist methods region motion vector calculation is not in zero value, does not comply with the actual situation, while the use of our method, the results obtained more realistic. From the experimental results of the comparison that our method for complex fluid motion vector calculation is more accurate and this demonstrates the effectiveness of the proposed method.

(2) Error analysis

The standard motion vector of cloud-free image sequence in Yosemite has been proposed in article [14]. In order to compare the accuracy, we calculate the direction error *Errangle* and the magnitude error *Err* respectively. The direction error *Errangle* measures the difference of direction between the calculated results with the standard vector of the particle motion vector. We use formula (9) to calculate it. The magnitude error *Err* measures the difference of magnitude between the calculated results with the standard vector of the particle motion vector in the image. We use formula (10) to calculate it.

$$Errangle = \frac{1}{n}\sum_{i=1}^{W}\sum_{j=1}^{H} | \arccos(vec_{ij}, std_{ij}) | \quad . \tag{9}$$

$$Err = \frac{1}{n}\sum_{i=1}^{W}\sum_{j=1}^{H} \sqrt{[(|vec_i| - |std_i|)^2 + (|vec_j| - |std_j|)^2]} \quad . \tag{10}$$

vec_{ij} is the motion vector result of image particle $P(i, j)$, and $std_i(\,std_i, std_j\,)$ is the accurate vector of $P(i, j)$. vec_i and vec_j are the components of motion vector result along x-axis and y-axis. W and H are the width and height of the image respectively, and n denotes the total numbers of image pixels, that is $n = H \times W$.

We calculate the error of the 1st frame in Yosemite cloud-free image sequence, and the result is shown in table 1. The unit of direction error is in radians.

Table 1. Error of 1st frame in Yosemite cloud-free image sequence

error	Liu [15]	Nilanjan [16]	Our method
Errangle	0.64343	0.76996	0.21366
Err	0.41059	2.31158	0.32107

It can be seen that use this method to calculate the direction of the motion vector error and the magnitude of error is relatively small, and this method of experimental results near the standard motion vector. The comparing results can further show the accuracy and effectiveness of the proposed method.

5   Conclusion and Future Works

The reconstruction and interactive technology on the fluid motion is a hot topic of research in computer vision field. In the study, calculation and analysis of fluid motion vector is one of the difficulties in this field. For the problem in this article a new method of motion vector calculated in the fluid of a natural landscape is proposed. The method takes advantage of intensity of fluid motive processing and the

character of continuity equation, and the initialize of motion vector, short distance linear interpolation and continuity equation to optimize the result. From the experiments, we validate the accuracy of our method. From the comparison, we further validate the performance. Our results are consistent with the fact and have a lower complexity, and it can be used in the further construction of virtual reality.

References

1. Jin, H., Gao, W.: Movement Analysis and Application of Facial Expression Based on the Characteristic Flow. Journal of Software 14, 2098–2105 (2003)
2. Horn, B.K.P., Schunck, B.G.: Determining Optical Flow. Artificial Intelligence 17, 185–203 (1981)
3. Lucas, B.D., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In: 1981 DARPA Imaging Understanding Workshop, pp. 121–130 (1981)
4. Corpetti, T., Memin, E., Perez, P.: Dense Estimation of Fluid Flows. Pattern Analysis and Machine Intelligence 24, 365–380 (2002)
5. Ghosh, A., Adrian, G.B.: Anisotropic Fluid Solver for Robust Optical Flow Smoothing. In: 10th Workshop Image Analysis for Multimedia Interactive Services, pp. 117–220. IEEE Press, London (2009)
6. Ghosh, A., Adrian, G.B.: Robust Processing of Optical Flow of Fluids. IEEE Transactions on Image Processing 19, 2332–2344 (2010)
7. Zhang, Z.W., Liu, G.Z., Li, H.L., Li, Y.L.: Using Energy Flow Information for Video Segmentation. Journal of Electronics 33, 177–180 (2005)
8. Li, M.G., Du, H.: A New Method Based on Cell Segmentation Theory for Particle Image Velcimetry. Journal of Electronics 36, 767–771 (2008)
9. Lu, Z.Q., Liao, Q.M., Pei, J.H.: A PIV Approach Based on Nonlinear Filtering. Journal of Electronics and Information Technology 32, 400–404 (2010)
10. Zhang, Z.Y., Deriche, R., Faugeras, O., Luong, Q.T.: A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry. Artificial Intelligence 78, 87–119 (1995)
11. Li, F., Xu, L.W., Guyenne, P., Yu, J.Y.: Recovering Fluid-type Motions Using Navier-Stokes Potential Flow. In: Computer Vision and Pattern Recognition, pp. 2448–2455. IEEE Press, San Francisco (2010)
12. Amiaz, T., Fazekas, S., Chetverikov, D., Kiryati, N.: Detecting Regions of Dynamic Texture. In: 1st International Conference on Scale Space and Variational Methods in Computer Vision, pp. 848–859. IEEE Press, Heidelberg (2007)
13. Peteri, R., Huskies, M., Fazekas, S.: DynTex: A Comprehensive Database of Dynamic Textures, http://www.cwi.nl/projects/dyntex/
14. Michael, J.B.: Requently Asked Questions, http://www.cs.brown.edu/people/black/
15. Liu, C.: Beyond Pixels: Exploring New Representations and Applications for Motion Analysis. Doctoral Thesis, Massachusetts Institute of Technology (2009)
16. Nilanjan, R.: Computation of Fluid and Particle Motion from a Time-Sequenced Image Pair: A Global Outlier Identification Approach. IEEE Transactions on Image Processing 10, 2925–2936 (2011)

Optimization of Space Color Mapping Using Compactly Supported Radial Basis Functions for Color Reproduction

Ladys Rodriguez[1], Luis Diago[2], and Ichiro Hagiwara[1,2]

[1] Institute for Advanced Study of Mathematical Sciences (MIMS), Meiji University,
1-1-1, Higashi-Mita, Tama-ku, Kawasaki, 214-8571, Japan

[2] Department of Mechanical Science and Engineering, Tokyo Institute of Technology, 2-12-1
Ookayama, Meguro-ku, Tokyo 152-8550, Japan

Abstract. Colors play an important role for customers to find their preference. The perception of the color depends on the devices used to show the colors and it changes with the color transformation between one device and another. This paper proposes an optimization of the Compactly-Supported Radial Basis Functions (CSRBF) space mapping to minimize the error in the color conversion between the system and the printer color spaces. A clustering k-means method is used to select the representative data in the printer color space to reproduce the whole space with high accuracy. The calculation of optimized CSRBF parameters using the representative data is proposed to minimize the color difference between the predicted CSRBF color value and the printed color value of all data in the printer color space. Proposed optimization method finds the optimized CSRBF parameters values and the optimal weighting parameters for color differences evaluation.

1 Introduction

The image color perception depends on the devices showing the image and it changes with the color conversion between one device and another. The color-matching methods are required to obtain the mapping function for the reproduction of colors in different devices. Cheol-Hee [1] proposed a gamut-mapping algorithm using linear equations to map the minima and maxima of the two gamuts onto each other. Tominaga [2] describes a method for solving the mapping problem for 3-dimensional color space of color stimuli to the higher dimensional color space of printer signals. The mapping from the Lab color space to the printer CMYK color space is constructed using an Artificial Neural Network (ANN). The ANN predictions are used as the interpolation function to transform one color space to another, but it is not clear the parameters (as network typology, count of hidden layers, learning procedure) to produce optimal results for other application domains. Qiao approach [3] is focused to design approximation models to minimize the error in the color conversion. Radial Basis Functions (RBF) are used for color conversion of image data to printer color space, based on the numerical model. The numerical model in Qiao approach [3] compares the color difference between the simulated color and the printed color using the Euclidean distance. A clustering technique and error approximations are employed to form the RBF into a linear programming

algorithm. The fundamental problem with RBF is its computation cost in terms of the need for solving large linear systems that is proportional to the size of the point data set. In Wendland's approach [5], Compactly Supported Radial Basis Functions (CSRBF) have been constructed for the purpose of getting finite-element type approximations. As the linear system becomes sparse, the computational resources for solving the linear system are reduced.

In this paper, the optimized color space mapping in printer devices is proposed. CSRBF interpolation function with optimized parameters is used to map the system colors to the printed colors. Proposed approach uses the ΔE_{00} color difference formula [7], instead of Euclidean distance formula, which provides an improved procedure for the computation of industrial color difference. In the following, section 2 explains the proposed CSRBF space mapping using ΔE_{00} color difference formula. Next, section 3 proposes the optimization of the CSRBF parameters to minimize the error in the space mapping function. In section 4, the calculation of optimized CSRBF parameters the space mapping function using 656 color samples provided by industrial manufacturer are illustrated. Finally, conclusions and future works are given in section 5.

2   Creating CSRBF Color Space Mapping

Compactly supported radial basis functions (CSRBF) [5] have been constructed for the purpose of getting finite-element type approximations [6]. The compact support automatically ensures that Φ is strictly positive definite. They give rise to sparse interpolation matrices and can be used to solve numerically partial differential equations. In Wendland approach [5] a popular family of compactly supported radial functions is constructed by starting with the truncated power function strictly positive Φ. The interpolated function Φ has different formulas according to the C^0, C^2, C^4, and C^6 continuity as listed below, in which r denotes the distance between two arbitrary points and r_0 denotes the ranges of influences of a point x_i. The CSRBFs is $\Phi_{r_0}(r) = \Phi(\frac{r}{r_0})$. $\Phi(r)$ is defined to be 0 if $r > 1$.

Table 1. Continuities of CSRBF

C^0	$\Phi(r) = (1-r)^2$
C^2	$\Phi(r) = (1-r)^4(4r+1)$
C^4	$\Phi(r) = (1-r)^6(35r^2 + 18r + 3)$
C^6	$\Phi(r) = (1-r)^8(32r^3 + 25r^2 + 8r + 1)$

The International Commission on Illumination (CIE) proposed the ΔE_{00} color difference formula [7] which provides an improved procedure for the computation of industrial color difference. The ΔE_{00} formula is based on CIELAB color space. Given a pair of color values in CIELAB space L_1, a_1, b_1 and L_2, a_2, b_2 the formula is denoted in Equation 1.

$$\Delta E_{00}(L_1, a_1, b_1 : L_2, a_2, b_2) =$$
$$\sqrt{(\frac{\Delta L'}{K_L S_L})^2 + (\frac{\Delta C'}{K_C S_C})^2 + (\frac{\Delta H'}{K_H S_H})^2 + R_T \frac{\Delta C' \Delta H'}{K_C S_C K_H S_H}} \tag{1}$$

S_L, S_C and S_H are the compensation for lightness, chroma and hue respectively. The term R includes a hue rotation to deal with the problematic blue (hue angles in the neighborhood of 275). K_L, K_C, and K_H are application-dependent parametric parameters (assumed to be 1 in [7]). Some approaches define $\Delta E_{00} = 1.0$ as the smallest color difference the human eye can see depending of the application domain and the color format [8].

Proposed approach finds the CSRBF space mapping function between the input (system) and output (printer) CIELab spaces. The ICC printer profiles help in the mapping process but the results are not 100% accurate, because the printer drifts over time taking account the ink state, the paper, etc. The 3-dimensional space mapping function $f(L, a, b)$ is proposed to represent the current state of one printer. This approach creates a custom printer profile using CSRBF. Fig. 1 shows a flowchart of the construction of CSRBF space mapping function $f(L, a, b)$ between system color space and printer color space. Function $f(L, a, b)$ models an space mapping between two set of colors, in which the input is $L_i, a_i, b_i, (1 \leq i \leq n)$ values of the system space and the output is the L'_i, a'_i, b'_i values of the printer space. The CSRBFs are used to construct the mapping function f:

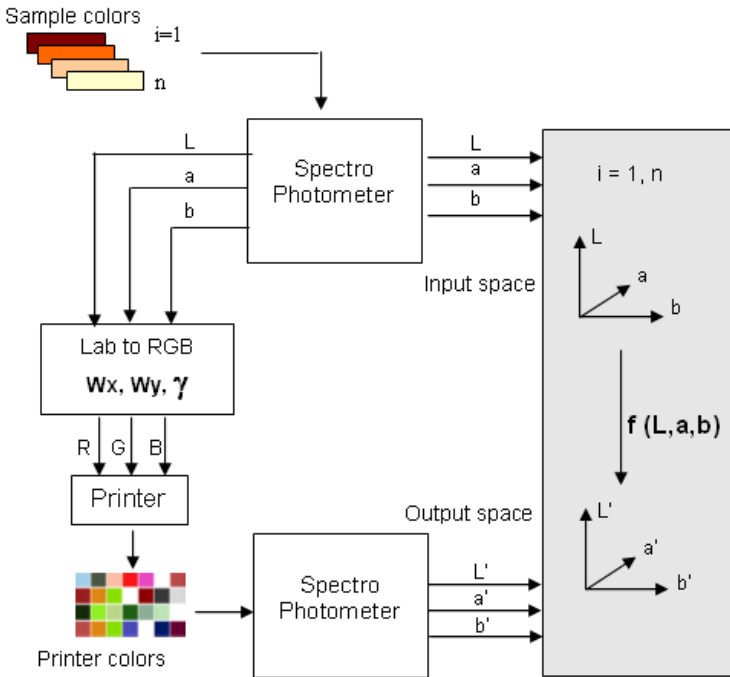$$f(\bar{x}) = \sum_{i=1}^{N} \lambda_i \Phi(|\bar{x} - \bar{x}_i|_{\triangle E_{00}}) + p(\bar{x}) \tag{2}$$

Fig. 1. Flowchart of the mapping process

for L, a and b dimensions, where λ_i are the CSRBF coefficients and p is the polynomial function of degree one. Here $\bar{x} = (L, a, b)$, $\bar{x}_i = (L_i, a_i, b_i) : 1 < i < n$ and $|.|_{\Delta E_{00}}$ denotes the color distance computed with Equation 1.

The Equation 3 has to be solved 3 times to construct the space mapping function f_L, f_a and f_b for L, a and b color dimensions:

$$\begin{pmatrix} \mathbf{\Phi} & \mathbf{P} \\ \mathbf{P}^{\mathrm{T}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{\lambda} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} \mathbf{F} \\ \mathbf{0} \end{pmatrix} \tag{3}$$

where $\Phi_{i,j} = \phi_{r_0}(|xi - xj|)$ and $P_{i,j} = p_j(xi)$ with $i, j = 1, \ldots, n$. For interpolating in 3D, the monomial basis $\{1, x, y, z\}$ is used for polynomials of degree one with $\mathbf{c} = (c_1, \ldots, c_4)^{\mathrm{T}}$ as the coefficients that give $p(x)$ in terms of this basis. The matrix composed by Φ, P, P^{T} and matrix with 0 is the same for the three dimensions. The distance between the input L_i, a_i, b_i and the output L_i', a_i', b_i' values are used as the right hand side vectors F_L, F_a and F_b respectively: $F_L = L_i - L_i'$, $F_a = a_i - a_i'$, $F_b = b_i - b_i'$ to calculate λ_L, λ_a and λ_b vectors and polynomial coefficients c_L, c_a and c_b.

However, the optimal value of the CSRBF parameter r_0, that denotes the ranges of the point influences in the prediction function, is not defined for color applications. Also, ΔE_{00} do not perform well for large color differences using proposed values for K_L, K_C, and K_H parameters. In the case of non-homogeneous surface, for example in the textile industry, optimization technique are applied to find the appropriated values for K_L, K_C, and K_H parameters [10]. In this paper, the optimization of the CSRBF space mapping parameters r_0, K_L, K_C, and K_H is proposed to minimize the error in the color conversion.

3   Optimization of CSRBF Parameters

3.1   Clustering of CIELab Color Data

To create the current state of one printer, all points in the printer color spaces have to be consider in the mapping function. As Lab color space is a continue space, it is important to select which data reproduce the whole space with high accuracy. Similar to Quiao approach [3], data clustering is used to identify the locations where basis function should be place in the model. In general, CSRBF algorithms are focused on placing functions such that the mean square error is minimized.

The clustering using k-means method is performed on the CIELab data so the characteristics of the color data can be described by each cluster. k-means clustering is a efficient method of cluster analysis. Given a set of observations $(x_1, x_2, , x_m)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the m observations into n sets $(m \leq k)$, $S = \{S_1, S_2, , S_k\}$ so as to minimize the within-cluster sum of squares, where μ_i is the mean of points in S_i.

$$\min \sum_{i=1}^{k} \sum_{x_j \in S_i} (||X_i - \mu_i||)^2 \tag{4}$$

The random center selection algorithm is used as the first step for cluster center selection. K-means produces a partitioning of the data space into Voronoi set(i.e., all points in the data closest in the color space to this center relative to other centers). K-means clustering tends to find clusters of comparable spatial extent. The set of cluster centers $C = \{c_1, c_2, , c_k\}$, $(c_i = Lab_i)$ are used to represent the system color space. The representation of the printer color space C' is obtained from the measure of the printed Lab colors in C. The CSRBF parameters λ_L, λ_a and λ_b vectors and polynomial coefficients c_L, c_a and c_b are calculated using the algorithm explained in previous section 2, to create the mapping function $f(L, a, b)$ between C and C'.

The CSRBF mapping function predicts the representative colors in the printer color space with 100% accuracy. The optimization of the parameters r_0, K_L, K_C, and K_H is proposed to minimize the error in the CSRBF prediction of all data in the printer space.

3.2   Finding Optimized r_0, K_L, K_C, and K_H Parameters

An optimization method is proposed to minimize the error in the mapping space function for printer color space. Generally, LevenbergMarquardt minimization algorithm is used in many applications for solving generic curve-fitting problems. LevenbergMarquardt algorithm (LMA) [9] provides a numerical solution to the problem of minimizing a function over a space of parameters of the function. The LMA interpolates between the GaussNewton algorithm (GNA) and the method of gradient descent. The LMA is more robust than the GNA, which means that in many cases it finds a solution even if it starts very far off the final minimum. Equation is used in LMA to find minimum of general form "sum-of-squares" function: $F = F(x[0]^2, ..., x[n-1]^2)$. Using the iterative LMA, the optimization function in the Equation 5 is proposed to minimize the sum of color difference ($\triangle E_{00}$) in the mapping process of two set of Lab colors x_i and x'_i. The objective is to find the optimized parameters r_0, K_L, K_C, and K_H involved in the CSRBF space mapping function to minimize the error in the color conversion. $E_{min} = E_{r_0, K_L, K_C, K_H}$

$$E_{r_0, K_L, K_C, K_H} = \sum_{i=1}^{k} \triangle E_{00}(f(x_i, r_0, K_L, K_C, K_H), x'_i)^2 \tag{5}$$

To start a minimization, the values $(1, 1, 1, 1)$ are used as initial values for the parameter vector. Boundary constraints for the parameters r_0, K_L, K_C, K_H have to be consider to control the influences of the high values in the $\triangle E_{00}$ color difference formula.

4   Experimental Results

Lab color values of the 656 samples provided by one industrial manufacturer were measured using the Spyder photometer to illustrate the performance of proposed approach. The experiment was made to find the optimized CSRBF parameters for 656 samples to minimize the color difference between predicted CSRBF color values and printed color values. For the construction of non-optimized CSRBF interpolation function, the Lab values of 656 colors are entered to the system and they are printed. The printed sample

a) L-a dimensions

b) L-b dimensions

c) a-b dimensions

d) $\triangle E_{00}$(clusters center, closest sample)
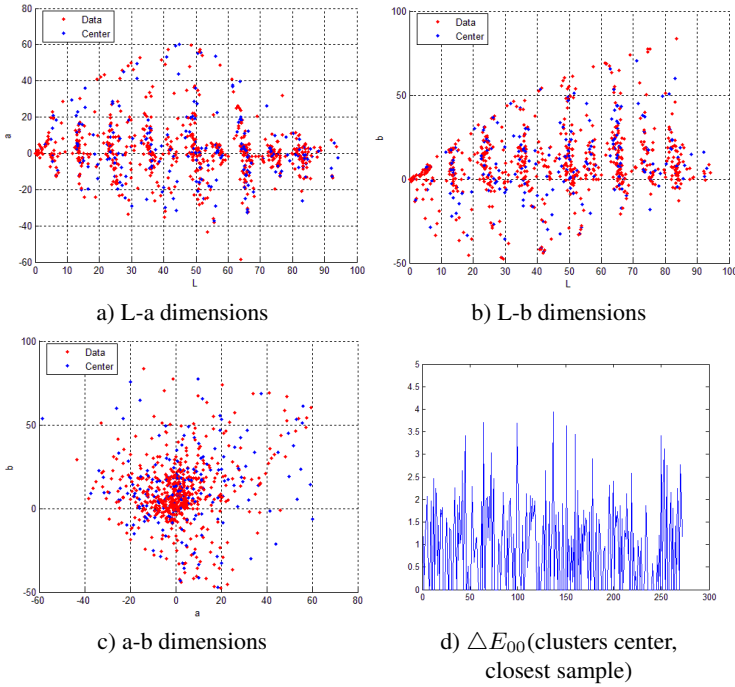
Fig. 2. a,b,c: Distribution of the 656 Lab sample colors(red points) and the 272 cluster centers(blue points)

colors are measured using the Spyder photometer obtaining the printed Lab values. The printed Lab values are used for the clustering algorithm to identify which printed colors are representative for printer color space.

Fig. 2 a,b,c shows the distribution in the L, a, b dimensions of the 656 samples colors(red points) and the 272 cluster centers(blue points) obtained by the k-mean clustering algorithm. For some cases, the Lab color values of the center of the clusters can be different of 656 the samples values. To avoid the printer impression of the 272 Lab value of the centers, we replace the center of each cluster with the closest sample value. Noted in the Fig. 2 d), the average of $\triangle E_{00}$ color differences between the center of each cluster and the closest sample value is less than 1 (Ave=0.91). In the experiment, the error introduced by the center replacement is not significant for this application, taking account the reduction of the time and the user effort in the printing and measure procedure. Using the 272 samples values (Lab system values and Lab printed values), the optimized parameters r_0, K_L, K_C, and K_H are calculated using the boundary constraint ranges for the parameters. Table 2 shows the results of the optimization function with three different boundary constraint ranges. In Table 2 is observed that proposed optimization method obtains the optimized parameters values highly dependent of the boundary constraint ranges.

Table 2. Optimization results of the mapping space function using the 272 samples values

Optimization	Value	r_0	K_L	K_C	K_H
Opt.1 Iter=197 $\min \sum \triangle E_{00}^2 = 236$	Initial	1	1	1	1
	Min.	0.1	0.1	0.1	0.1
	Max.	5	5	3	3
	Optim.	2.75	4.6	3	3
Opt.2 Iter=239 $\min \sum \triangle E_{00}^2 = 531$	Initial	1	1	1	1
	Min.	0.1	0.1	0.1	0.1
	Max.	5	12	2	2
	Optim.	4.13	3.07	2	2
Opt.3 Iter=251 $\min \sum \triangle E_{00}^2 = 2124$	Initial	1	1	1	1
	Min.	0.1	0.1	0.1	0.1
	Max.	10	20	1	1
	Optim.	8.26	1.53	1	1

5   Conclusion

This paper proposed the creation of a custom printer profile using the optimized CSRBF mapping function with $\triangle E_{00}$ color differences formula to minimize the error in the color conversion between the system and the printer color spaces. First, a clustering method is used to select the representative data in the printer color spaces to reproduce the whole space with high accuracy. Second, an optimization method is proposed to minimize the error in the mapping space function. Experiments were made to find the optimized CSRBF parameters for 656 samples color provided by manufacturer company to minimize the color difference between predicted CSRBF color value and printed color value. Proposed optimization method obtained the optimized CSRBF parameters values with high dependency of the parameters boundary constraint ranges. However, it can be extended to the other fields of applications by creating smaples profile and optimizing parameters as proposes.

LevenbergMarquardt algorithm is used in many applications for solving generic curve-fitting problems, the local convergence of the algorithm produced results depending on of the initial and constraint values. For future works, global optimization algorithm will be considered to avoid constraints dependency.

References

1. Cheol-Hee, L., Yeong-Ho, H.: Parametric Gamut Mapping Algorithms Using Variable Anchor Points. Journal of Imaging Science and Technology 44(1), 68–89 (2000)
2. Tominaga, S.: Color control of printers by neural networks. Journal of Electronic Imaging 7 (1998)

3. Qiao, Y.: Linear processing in color conversion. United States Patent Application Publication No. US2010/0157397 A1 (2010)
4. Savchenko, V., Pasko, A., Kunii, T.L., Savchenko, A.: Function representation of solids reconstructed from scattered surface points and contours. Computer Graphics Forum 14(4), 181–188 (1995)
5. Wendland, H.: Piecewise polynomial, positive definite and compactly supported radial basis functions of minimal degree. Advances in Computational Mathematics 4, 389–396 (1995)
6. Brenner, S., Scott, L.: The mathematical theory of finite elements. Springer, New York (1994)
7. CIE, Improvement to industrial colour-difference evaluation, CIE Publication No. 142-2001, Central Bureau of the CIE (2001)
8. CHROMiX, Delta-E: The Color Difference, CHROMiX ColorNews Issue No. 17 (2005)
9. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. The Quarterly of Applied Mathematics 2, 164–168 (1944)
10. Sijie, S.: An investigation of texture effects on instrumental and visual colour difference evaluation, pp. 1–331. Hong Kong Polytechnic University, Hong Kong (2006)

A Method of Integrating Simulation with C^2 System

Wei Chu and Xing E. Yan

Science and Technology on Information Systems Engineering Laboratory
MailBox #1406-69, Nanjing, China, 210007
`wicked_1979@163.com`

Abstract. In the military simulation area, providing environment for C^2(Command and Control) systems is a main application of simulation. The interoperation technique between simulation and real system is increasing attention. This paper analyzes the demand of integration between simulation and C^2 system, and then provides the interactive models. In the meantime, this paper brings forward a method of integration which is proved by practical application that this method can eliminate the drawbacks of traditional methods.

Keywords: Simulation, C^2 system, Integration.

1   Introduction

In the military simulation area, simulation can be an instrument or tool to construct the combat environment for C^2 system which just like some researchers use the technology of M&S (Modeling and Simulation) to simulate C^2 system's combat environment. Enhancing interaction and interoperability between simulation and C^2 system has become a hotspot and is taken common attentions by some military research organizations and military simulation field[1] because simulation will support C^2 system's experiments, such as T&E(Test and Evaluation) and simulated training, etc[2]. The main method which we often used to integrate simulation with C^2 system is called simulation agent. But overall, simulation agent has many shortcomings. This paper will provide a new integration method called Dual-Net which has been proven to be efficient.

2   Demand Analysis

The simulation of military equipments can be applied to many aspects, such as demonstration and analysis, design and validation, simulated training, etc[3]. We will take two types of main simulation applications including integration experiment and T&E as examples to analyze the integration demands between simulation and C^2 system.

2.1   Integration Experiment

In the integration experiment of C^2 system/equipment, we need to construct its runtime environment realistically, which includes simulating its superior/subordinate/adjacent systems and its intelligence/recon/scout devices and weapon platforms, etc. Thus, for validating real system's ability and capability, it commonly needs simulation devices to substitute some other real equipment. These simulation devices should be embedded into the experiment environment of real system and make information alteration with real

equipments. This type of information alteration in the real system's integration experiment is just only a relationship of information processing and does not need the interoperation between simulation and real system. The alteration sketch map of system's integration experiment is showed in Figure 1.
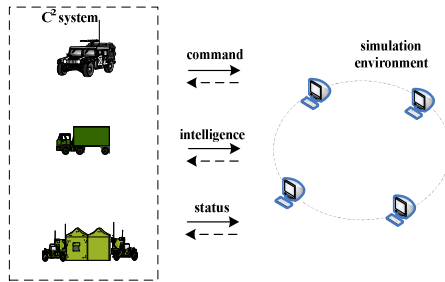
Fig. 1. Alteration Sketch Map of Integration Experiment

2.2   Test and Evaluation

For the simulation application of system's test and evaluation, we should firstly construct the simulation testing environment by using emulation techniques and embed the real system to this environment. Thus, we can utilize simulation to inspire the operation of real system and calculate the results of test and evaluation by data recording and processing. Such simulation experiment sometimes forms a distributed architecture and has its own interactive simulation protocols and interfaces, such as DIS(Distributed Interactive Simulation)and HLA(High Level Architecture), etc. The experiment of T&E is commonly carried out by the way of closed loop. That is to say, after the real system accepting the simulated data, it will make data processing and feed back the results to the simulation environment. So it will form a closed loop of C^2. In this condition, the contents of interaction not only include the simulated intelligence and command data, but also include the orders, intelligence and situation which are fed back by the real system. These data then take effect on the simulation system and cause new change. This kind of simulation application requires the interaction between simulation and real system and the information processing is bidirectional. The alteration sketch map of system's T&E is showed in Figure 2.
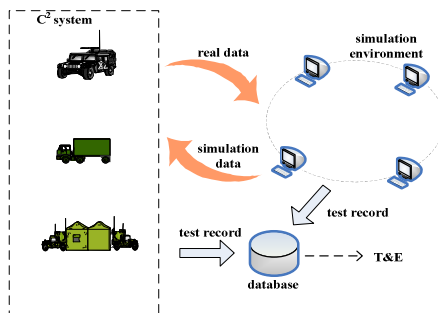
Fig. 2. Alteration Sketch Map of T&E

3   Interoperability Model of System Integration

3.1   Interactive Model

If there is only a simple relationship of information processing between simulation and real system, we call this weak interaction. The job in weak interaction is the protocol conversion between simulation and real system. The direction of data stream is unilateral and is relatively simple to realize. But for the strong interaction which needs closed loop, it needs not only realizing protocol conversion, but also resolving the information interoperability problems between simulation and real system, such as intelligence, commands and feedback results. The realization process is relatively complicated. The interactive model[4] between simulation and C^2 system is shown below. The accessorial functions of strong interaction than weak interaction are showed in the broken line frame. (ISR: Intelligence, Surveillance, Reconnaissance; EW: Electronic War)
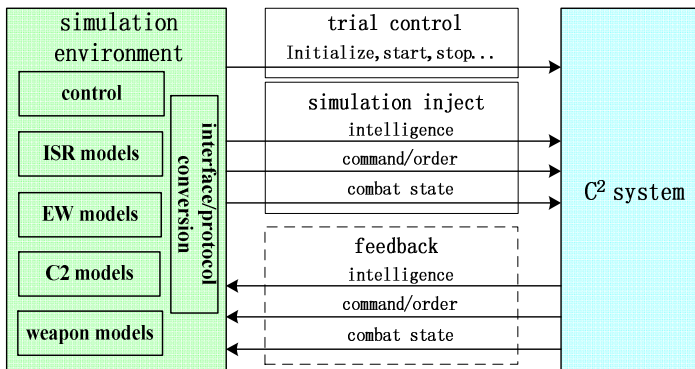
Fig. 3. Interactive Model between Simulation and C^2 system

3.2   Technical Reference Model

According to the interactive model of simulation and C^2 system, we bring forward the technical reference model as showed in Figure 4. This model defines some concerned conceptions, functions and interfaces to settle the problem of interacting with each other.

The interactive data between simulation and real system include: mutual information with simulation services, permanent data and impermanent data.

- The mutual information with simulation services is used to control, manage and maintain the process of simulation experiments. The mutual information with simulation services includes simulation control information, data gathered and simulation results, etc. C^2 system does not handle such data at present, but it will and should in answer to this kind of data in the future.
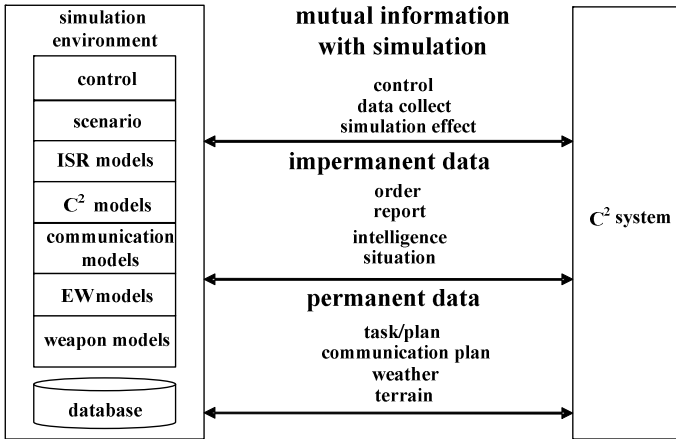
Fig. 4. Technical Reference Model of Interaction between Simulation and C^2 system

- The permanent data is the data that is always reserved and maintained in the process of simulation experiment. It includes information of tasks and plans, communication plans, weather information and terrain data, etc.
- The impermanent data is the temporary data which is transferred between simulation and C^2 system in the process of simulation experiment. It includes command, report, intelligence and situation, etc.

4   Integration Mode

4.1   Integration Based on Dual-Net

Dual-Net integration mode means the simulation environment is designed by the mode of two buses. One is called simulation inspire bus, and another is called system operation bus. Simulation inspire bus is used for passing various kinds of simulation inspire information, such as scenario information, running state, recorded data, etc. System operation bus is used for passing various kinds of application messages, such as intelligence, commands, fight status, etc. Such design pattern is able to avoid simulation inspire information occupy the bandwidth of System operation bus and enhance the test fidelity and the credibility of the test results. The integration mode of Dual-Net is showed in Figure 5.

Dual-Net structure has these functions as follows:

- Transfer Simulation Inspire Information

There are two types of equipments in the environment of simulation experiments. The first type is simulation management device which includes simulation control devices, data collect devices and experiment monitor devices, etc. These devices accomplish the functions such as generation of objective situation, management and
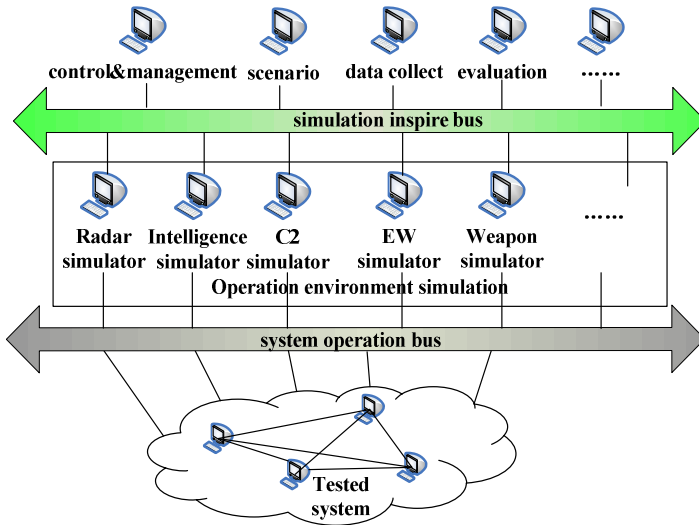
Fig. 5. Integration Mode based on Dual-Net

control of trial operation, etc. The second type is kinds of C^{4}ISR•Command, Control, Communication, Computer, Intelligence, Surveillance, Reconnaissance•simulation devices. These devices simulate various kinds of ISR, C^2, electronic war, weapons platform, etc. Both types should follow the standards and protocols of simulation interaction to transfer simulation inspire information for realizing scenario injection and trial management. In the structure of Dual-Net, there exists special physical communication link to ensure the transmission of simulation inspire information.

- Transfer Combat Application Information

When a system is integrated or tested, the C^{4}ISR simulation devices in the simulation environment are usually act as its superiors, subordinates, neighbors or rival systems to construct its combat environment. On the one hand, these devices connect the simulation inspire bus to accept and handle various simulation inspire information. On the other hand, these devices connect the system operation bus to handle the combat application information. The tested system is deployed according to real physical configure, so system operation bus also should be designed and realized according to real system's communication requirements, includes physical transmission links, interfaces, protocols, etc. In the structure of Dual-Net, the transmission of combat application information follows the real physical communication links of the tested system. Every C^{4}ISR simulator in Dual-Net integration mode has two interfaces. One is called simulation inspire interface and another is called system operation interface.

It is not hard to see that in the integration mode of Dual-Net, every C^2 simulation device includes two types of interfaces. The first type is provided according to the requirements of simulation inspire bus and is used to interact with the simulation management devices. The other one type is realized according to real system's protocols and is used to interact with tested system/equipment for driving the tested

system's operation. These two types of interface are mutually independent and not affect each other. With the help of trial scenario, they accomplish various experiments, tests and evaluations. The job of format conversion of information processing is completed by simulation devices themselves.

4.2   Improvement

Due to the advantage of simulation agent, we may deploy several simulation agents in the Dual-Net integration mode which can isolate the simulation environment and real system to ensure the independence and reusability preferably for simulation environment (Figure 6).
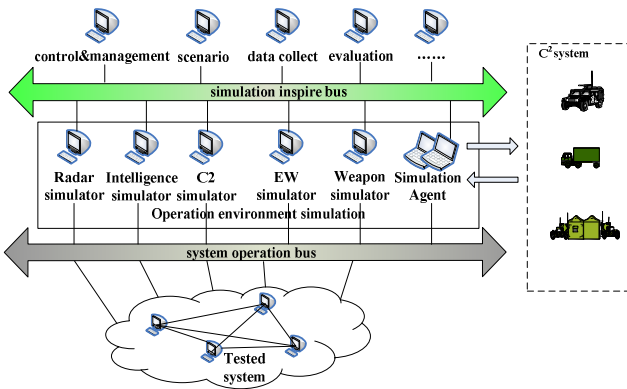
Fig. 6. Integration Mode based on Dual-Net

On the one hand, simulation agent is a component of the simulation environment itself. It can interact with other simulation components according to the criterion of simulation environment. On the other hand, simulation agent owns the interfaces of C^2 system and can make information alternation with C^2 system. Its functional structure is showed in Figure 7.
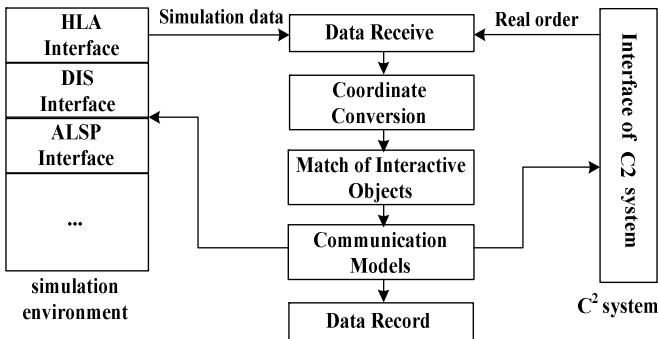

Fig. 7. Functional Structure of Simulation Agent

Simulation agent should have these functions as follows at least:

- Realize various simulation interfaces

At present, the simulation environment which has simulation agents follows the mainstream architectures of distributed simulation systems essentially as before, such as HLA、DIS、ALSP(Aggregate Level Simulation Protocol), etc. Therefore, the simulation agent should operate under the corresponding architectures of simulation technique system and alternates with other simulation members as a component of the simulation environment.

- Provide the interfaces with real system

Simulation agent should realize the interface with real system too. There should be an open physical interface between simulation agent and real system and the interface can be configured. That is to say, in order to connect with the real system or single equipment, it should provide several different physical interfaces, such as serial interfaces and Ethernet interfaces.

- Coordinate Conversion

The projection coordinate is commonly different between simulation environment and real system. When they are making information alternation which correlates with coordinate position, simulation agent should make coordinate conversion. Only in this way, it can ensure the consistent understanding between simulation environment and real system.

- Protocol Conversion

The message formats used by real system and simulation environment are different. In order to connect and understand with each other, simulation agent should provide the function of protocol format conversion between simulation environment and real system[5].

- Match of Interactive Objects

Match of interactive objects means the process of confirmation of interactive objects when the real system and simulation environment are making alternation. For example, a target's number which is appointed to hit in the real system is different from the same target's number in simulation environment. Therefore, when the target which is appointed to hit is a member of simulation environment(such as a fighter plane simulated by computer), the fight command ordered by real system needs using the method of match of interactive objects to confirm which one is appointed to hit in the simulation environment. The method of match of interactive objects has the ability to establish contact with each other automatically.

- Communication Simulation

Communication simulation means modeling the process of communication. It simulates the communication situation of the nodes between simulation and real system. The communication models include shortwave, ultra short wave, microwave, etc.

- Data Record

The function of data record means recording the data which is injected from simulation environment to real system. The data include simulated targets, intelligence, command, etc. At the same time, it also needs to record the data which is fed back from real system to simulation environment, such as command, order, intelligence, situation, etc.

4.3   Comparison

We have applied Dual-Net integration mode to lots of integration experiment between simulation and C^2 system. Practical applications prove that this integration mode is effective and efficient.

For the integration mode based on simulation agent, it is commonly adapted to the test or evaluation of single tested system or equipment. It can not support the trial of large-scale system especially when the tested system is deployed according to its real deployment and configure because there need multi simulation agents accordingly. Due to the complex relationships in the tested system, the information relationship and communication support for each simulation agent are extraordinary complex. This sometimes will cause the information relationship of tested system, simulation agents and simulators are hard to define and describe.

Analyzing from the mode of simulation experiment, Dual-Net integration mode belongs to the trial mode of simulation system embedded to real system which one may provide favorable test support ability. It can simulate the real deployment and organization of the tested system realistically. Its disadvantage lies in that it needs more work to develop, integrate and configure the simulators because the C^4ISR simulators need to realize two types of interfaces which include simulation inspire bus and system operation bus. Meantime, due to the different interfaces of various tested systems in various experiments, the interfaces between C^4ISR simulators and system operation bus need to rebuild and update to satisfy the requirements. Thus, it will affect the agility and reusability of simulation environment to some extent. However, we can combine Dual-Net with simulation agent to overcome these drawbacks which is already used at all.

5   Conclusion

This context analyzes the integration requirements between simulation and C^2 system, research on the interoperability model, provide an integration mode which called Dual-Net and make a comparison. When simulation environment is integrated with C^2 system, we should take adequate consideration on the characteristics and operating requirements of the system and adopt a proper integration mode. At the current point for the possible usage requirements, the combined integration mode of simulation agent and Dual-Net is better.

References

1. Yu, B.G., Jie, M.S.: The Research of Interoperability between Simulation and C4I System. In: The Fourth C^2 System Top Seminar: Beijing, System Simulation College (2007)
2. Yan, W., Ye, W., Wei, H.: Research of Integration Mode between Simulation and C^2 system. Journal of System Simulation 20(1), 235–238 (2008)
3. Jie, M.S., Qi, J.Z., Yu, B.G.: Simulation Technology of C^2 system and Application Summarization. Modern Electric Engineer 6, 1–6 (2008)
4. Layman, G., Weatherly, J.: New Navy Solutions: Developing Simulation Based C4I Applications. In: Command and Control Research & Technology Symposium. ONR, San Diego (2004)
5. Zach, F., Michael, R.H.: A Generalized Unified Model for C^4ISR/Simulation Interoperability. In: Simulation Interoperability Workshop. IEEE Press, Orlando (2001)

An Implementation of FlexRay Bus Data Communication

Tingyao Liu[1], Yueli Hu[1,2,*], Longjie Wang[2], and Chao Yu[2]

[1] School of Mechatronics Engineering and Automation, Shanghai University,
Shanghai, P.R. China
[2] Key Laboratory of Advanced Display and System Application,
Shanghai University, Shanghai, P.R. China
`huyueli@shu.edu.cn`

Abstract. The traditional bus in vehicle, like CAN and LIN, can not meet the requirement of some control systems in reliability and high speed. In order to realize the reliable and High-speed FlexRay bus communication, a FlexRay bus data communication implementation scheme is put forward in this paper, and a specific Brake-By-Wire system based on the scheme is designed. The experimental results show that no matter in high-speed or in reliability the scheme have shown a good performance. It can be predicted that the implementation scheme of FlexRay bus communication also can be applied to other X-By-Wire systems like Steer-By-Wire and so on.

Keywords: FlexRay, X-By-Wire, brake motor, S12XF512.

1 Introduction

With the development of bus in vehicle technology and the integrated circuit technology, more and more control systems in vehicle began to use electronic technology to replace the original machine control. Using bus technology can increase the space in vehicle and reduce fuel quantity by decrease weight. However due to the lack of synchronicities, certainty, fault tolerance and the speed is relatively low, the widely used CAN，LIN bus can not satisfied some control systems which requires high speed and reliability, such as brake system, clutch system. And the new FlexRay bus using redundancy backup methods for dual communication, when a fault occurs in one of the communication lines, the other line can complete communicate independently. In the speed aspect, FlexRay can achieve a communication speed of 20Mbps. Therefore, the FlexRay bus can meet the reliability and the high speed required in these kinds of control systems. At present,some foreign car manufacturers such as BMW, have begun toapply FlexRay to some high-end models. As for the domestic, FlexRay is currently still at the laboratory stage of research and development.

The implementation scheme presented by this paper realized the FlexRay bus communication by using a 16-bit microcontroller called S12XF512 which is designed by Freescale.TJA1080T designed by NXP was applied as the FlexRay bus driver.

---

[*] Corresponding author.

2   FlexRay Protocol

FlexRay media access scheme is based on communication cycle as depicted in Figure 1.FlexRay offers two different media access scheme in static segment and dynamic segment: static time division multiple access scheme in Static Segment and Dynamic flexible time division multiple access scheme in dynamic segment. FlexRay allows the user to configure the length of the static segment and dynamic segment according to the actual situation flexibly.

The communication cycle is the fundamental element of the media access scheme within FlexRay. It is defined by means of a timing hierarchy. The timing hierarchy consists of four timing hierarchy levels as depicted in Figure 1. From the highest level to the bottom followed by communication cycle level, arbitration grid level, macrotick level and microtick level.
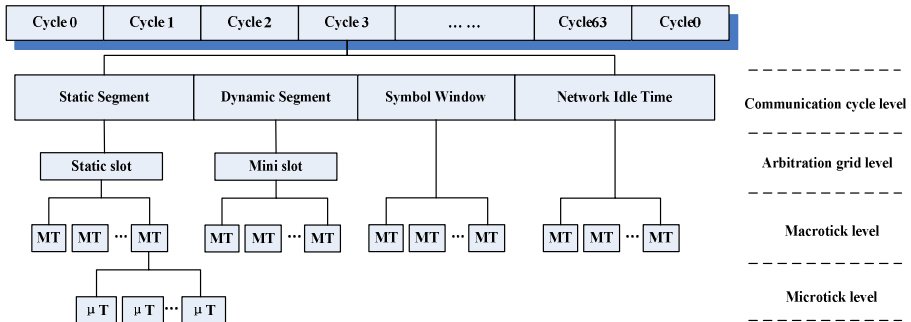
Fig. 1. Timing hierarchy within the communication cycle

A static segment, for example, is composed of N static slots. In a static segment, the number of slots in the static segment is affected by the size of data that transmitted and the communication cycle, as shown in Figure 2. Of which:

•AP（Active Point）:The action point offset, consists of a certain number of MT;
•Static Frame: Contains Frame Start Sequence, Frame End Sequence, etc;
•CID（Channel Idle Delimiter）: The continuous high level expressed that the communication ended;
•Channel Idle：Channel idle time.
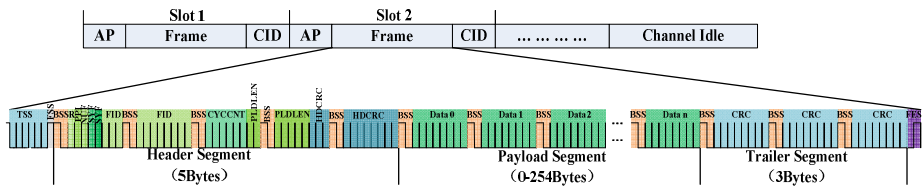
Fig. 2. The frame format of FlexRay bus

Static Frame sent after the end of the Active Point was sent .First, a 3- 15gdBit Transmission Start Sequence(TSS) was sent, followed by 1gdBit high level Frame Start Sequence(FSS) , and then 2 gdBit Byte Start Sequence(BSS) was sent before every Byte (8gdBit) data. After all the data was sent, Frame End Sequence(FES) was sent.

After transmitting the Frame Start Sequence, stat to transmit data frame which is consisted by N bytes Byte Start Sequence and the N bytes data. The data frame is composed of 3 parts: Header Segment(5Bytes), Payload Segment(0-254Bytes) and Trailer Segment(3Bytes).

HeaderSegment is consisted by some indicators, such as the Sync Frame Indicator (SYF) is used to indicate whether the frame is a synchronous frame .In addition, the Frame ID (FID) in HeaderSegment is used to define the frame in which slot sent .The FlexRay payload segment is used to convey application data which contains 0 to 254 bytes (0 to 127 two-byte words) of data.The Frame CRC field in trailer segment contains a cyclic redundancy check code (CRC) computed over the header segment and the payload segment of the frame. The computation includes all fields in these segments.The CRC is computed using the same generator polynomial on both channels.

3   An Implementation of FlexRay Bus

In the proposed implementation scheme, the FlexRay node use Freescale 16-bits microcontroller S12XF512 and NXP FlexRay transceiver TJA1080T, specific node scheme shown in Figure 3 .
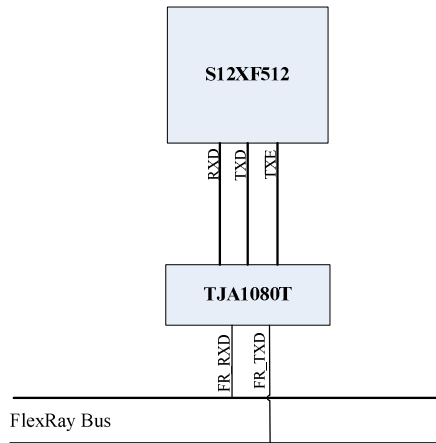
Fig. 3. FlexRay node scheme

S12XF512 microcontroller integrated 2 independent FlexRay communication controllers, and supports data rates of 2.5, 5, 8 and 10MBit/s on each of the two channels. FlexRay protocol implementation was according to FlexRay V2.1 Protocol

Implementation document. In addition, the microcontroller supports 16 channels of analog to digital converters (ADC) within just 3 μs conversion time, which could perform well in terms of time and speed when designing a FlexRay system requires signal acquisition, such as Brake pedal, steering wheel. On the other hand, the microcontroller integration of six channels of Pulse Width Modulator with Fault protection(PMF) optimized for electrical motor control. Therefore, it can be expected that an implementation scheme of FlexRay communication using this microcontroller could get good system performance, and could be applied to many application areas.
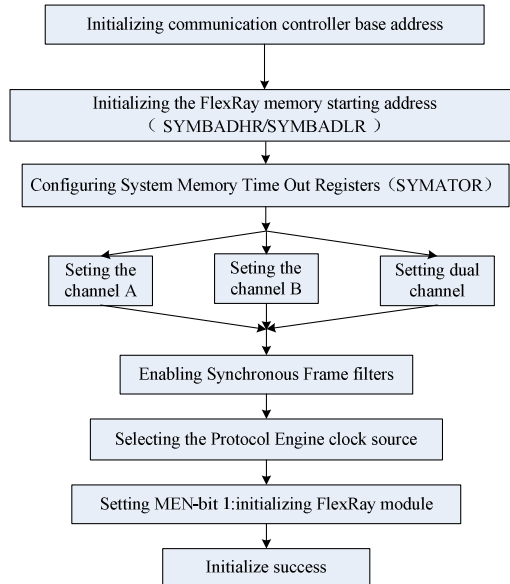
Fig. 4. FlexRay module initialization process

Before transmitting and receiving data FlexRay node need to initialize the FlexRay module in S12XF microcontroller.The initialization is shown in Figure 4.

4   Application in Brake-By-Wire System

Brake-By-Wire system can make good use of the high-speed and stability of FlexRay bus. And the microcontroller S12XF in the FlexRay communication scheme proposed in this paper has a convenient A/D converter and PWM modules. It is very suitable for the design of Brake-By-Wire system.

4.1   Hardware Architecture Design

The architecture of the Brake-By-Wire system using the FlexRay communication scheme present by this paper is shown in Figure 5.
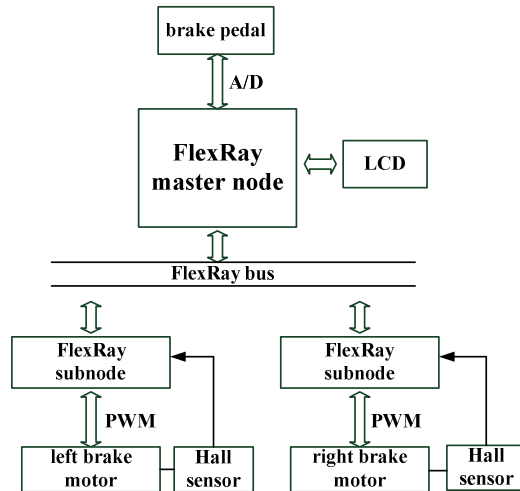
Fig. 5. System Architecture

As shown in figure 5, S12XF512 microcontroller in FlexRay master node collected the information from the brake pedal. Then, after the analog to digital conversion and encoding based on the angle of the pedal stepped on, the collected information of brake pedal was displayed on the LCD screen and sent to every subnode through FlexRay bus. The subnodes received the information of the brake pedal from master node and adjusted the PWM sent to the left and right brake motor correspondingly for brake control. On the other hand , Hall sensors was assembled on the brake motors to collect the speed of the motors, and the subnodes would transmit the speed information to the master node through FlexRay bus, so that the microcontroller in master node could display the speed information on the LCD. Once when the system fails, it could be detected and maintained according to information on the LCD.

4.2   Software Design

The software design of the system includes the FlexRay module initialization(as shown in Figure 4), FlexRay transmit subroutine , the FlexRay receive subroutine , brake-motor-control subroutine , the LCD-display subroutine and so on. The FlexRay master node and subnode software flow chart are shown in figure 6 and figure 7.

Master node and subnodes all have data to transmit and receive , so what should be done in the first place is the FlexRay module initialization .For the master node , when the initialization was successful and there is no data to transmit and receive , the master node will transmit a test frame to test the respond of FlexRay bus. When the brake was stepped on, the master node transmitted the brake information via FlexRay subroutine to the subnode , and the LCD displayed speed information of the motor from subnode by LCD-display subroutine. As for the subnode, it also started to communicate with the master node after the success of the FlexRay module initialization and completed brake function by PWM-motor-control subroutine.
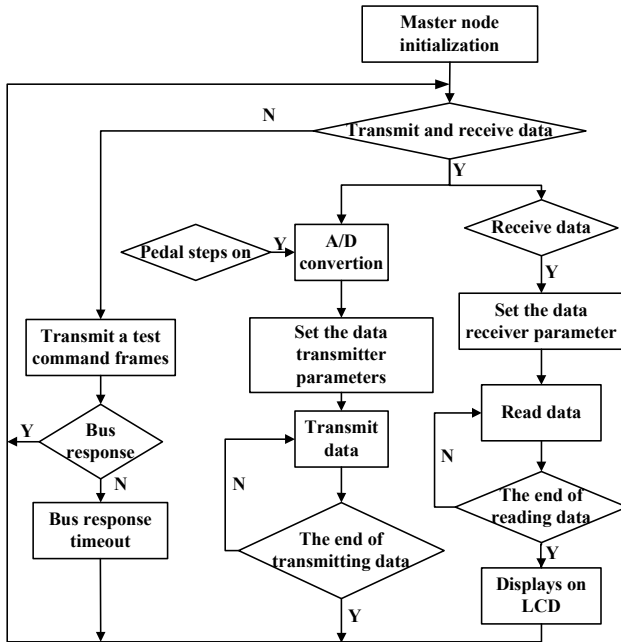
Fig. 6. Master node software flow chart
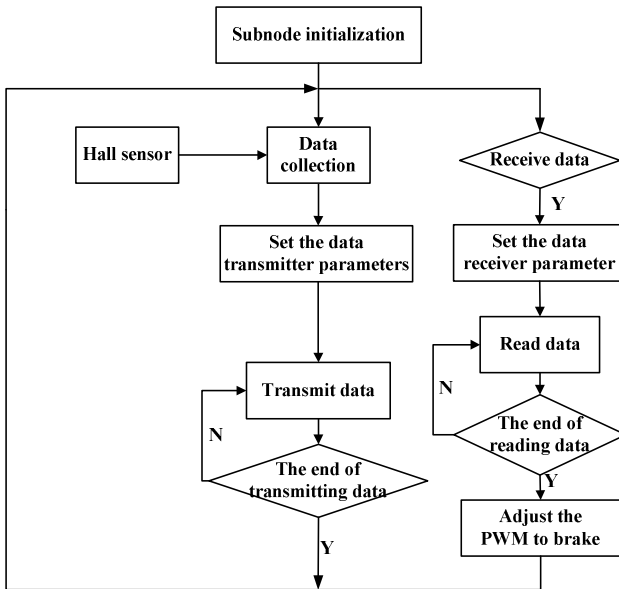
Fig. 7. Subnode software flow chart

Whether in terms of structure or function the entire system could achieve a very high flexibility, convenience and reliability.

5   Experimental Results

Hardware environment should be built before the experiment. According to the schema shown in Figure 5, the hardware connection mainly including : connection to LCD and the master node, brake motors and the two subnodes, and then each node should be connected to the bus hub. In order to obtain experimental data and waveform, an oscilloscope was connected to the two FlexRay bus signal lines.

As for the software, it should be download to the master node and subnodes by Freescale's BDM programmer. After the PC prompts the write is successful, the master node and subnodes could be power-on .

When the brake pedal was stepped on, there was signal on the FlexRay bus and the signal waveform can be measured by the oscilloscope. 1bit signal waveform was shown in Figure 8.
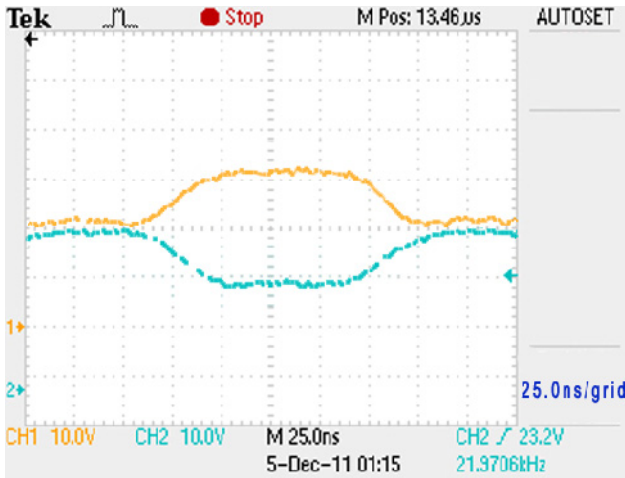
Fig. 8. The Signal waveform on FlexRay bus

As the oscilloscope screenshots shows, abscissa represents time and vertical axis represents the voltage. There are four grids in 1bit data, and it is 25ns for 1 grid. It can be achieved by calculating that the bit rate on FlexRay bus of the system is:

$$\text{Bit Rate} = \frac{1}{25ns \times 4} = 10 Mbit / s .$$

In the practical operation, when the brake pedal is stepped on the bottom, the brake motor can instantly stop turning.

During tests, some simulated vehicle environmental factors, like electromagnetic interference and high temperature, was added to the signal line and the master node and subnodes, brake motors can still work properly. And the system can still perform well on the brake function when one of the signal lines was cut during the braking process. It shows that the system has a good reliability.

6   Conclusion

This paper presents an implementation of a FlexRay bus communication, and designed an application in Brake-By-Wire system based on the scheme. The experimental results show that the bit rate of the FlexRay bus can reach 10Mbit /s on each channel in the scheme, and it can meet the speed requirements of Brake-By-Wire system. It is worth to mention that the scheme proposed by this paper can be applied not only to Brake-By-Wire system, but also to other X-By-Wire system, such as Steer-By-Wire, suspension-By-wire, Throttle-By-Wire and so on.

References

1. Sun, B., Hu, Y., Wang, K.: Diagnosis design based on KWP2000 in body control module. In: ICMTMA, vol. 1, pp. 1071–1074 (2001)
2. Sun, B., Hu, Y.: Automatic body control model experimental platform design based on CAN-bus. In: ICEEAC, vol. 2, pp. 341–344 (2010)
3. Bin, S., Hu, Y.: Improved AUTOSAR based on Multi-Core architecture and its application in the body control. Research Journal of Applied Sciences, Engineering & Technology 3, 1197–1202 (2011)
4. FlexRay Consortium: FlexRay Communication Systems Protocol Specification Version 2.1. Revision A (2005)
5. Shaw, R., Jackman, B.: An Introduction to FlexRay as an Industrial Network, pp. 1849–1854. IEEE (2008)
6. Cena, G., Valenzano, A., Vitturi, S.: Advances in automotive digital communications. Computer Standards & Interfaces 27, 665–678 (2005)
7. Peti, P., Timmerberg, A., Pfeffer, T., Muller, S., Ratz, C.: A quantitative study on automatic validation of the diagnostic services of Electronic Control Units. In: Emerging Technologies and Factory Automation, pp. 799–808 (2008)
8. Lundelius Welch, J., Lynch, N.: New Fault-Tolerant Algorithm for Clock Synchronization. Information and Computation 77(1), 1–36 (1988)

The Design and Simulation of a Two-Layer Network Protocol for Industrial Wireless Monitoring and Control System

Zhile Yang[1,2], Minrui Fei[1,2], Weiyan Hou[1,3], and Bingchen Wang[1,2]

[1] Shanghai Key Laboratory of Power Station Automation Technology
[2] School of Mechatronics Engineering and Automation
Shanghai University, Shanghai, 200072, China
[3] School of Information Engineer
Zhengzhou University, Zhengzhou, 450000, China
{yangzhile,wc04122833}@126.com
mrfei@staff.shu.edu.cn
houwy.zzu@gmail.com

Abstract. As the vital compliment of existed wired monitoring and control network, wireless systems, which share the 2.4GHz channel, are gradually introduced into the industrial application field. A previous research WICN (Wireless Industrial Control Network)-a token ring based network is limited by its single hop short range. A novel protocol named WICN-TL (Wireless Industrial Control Network-Two layers) is designed to deal with the distance problems of its predecessor. The paper introduces the topological structure, the stack model, the data format and the communicating procedure of the protocol. Moreover, the network protocol stack is implemented on the NanoNET hardware platform based on the IEEE802.15.4a standard and tested in a sewage plant and a turbine power simulation platform. The result shows that the network based on the protocol performs well in communication and the covered range is sufficient to apply to the normal industrial application demand.

Keywords: wireless network protocol, industrial network, monitoring and control, token ring, 802.15.4a.

1   Introduction

On the process industry site, field-level networks develop fast and have emerged three generations—the fieldbus system, the industrial Ethernet and the wireless networks in automation.[1] With well real-time performance and communicating reliability, wired networks including both fieldbus and Ethernet systems have been widely used and will continuously be the major style of industrial communication. However, under the circumstances of measuring rotating or mobile devices as well as in some hostile environment, wireless network can be introduced as the compliment to realize the flexible and convenient message acquirement. As a result, the wired and wireless network will undoubtedly co-exist for quite a long time.[2]

There are three industrial wireless specifications at present and all of them are proposed in the past few years-they are ISA100.11a, Wireless-HART, and WIA-PA. Not altogether incidentally, they share the same physical layer standard IEEE802.15.4 with 2.4GHz channel considering the energy cost [3]. Though the three specifications have different topological structures, all of them introduced Ad-hoc routing scheme respectively to enable the cover range. Nevertheless, the complicated routing methods of mash networks result in difficult management and time unreliability. In order to meet the need of simplicity and reliability on process industry scene, the previous design WICN[4] was created. But its single hop structure makes it hard to meet the normal demand. A novel wireless network protocol named WICN-TL is introduced to solve this problem. The protocol takes reference of PROFIBUS standard and utilizes a two layer network structure-a token ring-based network as the backbone and a polling based network as the subnet. In this paper, the topological structure, the stack model, the data format and the communicating procedure of the protocol is proposed. It is also implemented on the NanoNET platform based on the IEEE802.15.4a and tested in a sewage plant and a turbine power simulation platform.

2   Topological Structure

WICN-TL has two layers of networks as figure1 shows and it is composed of totally four kinds of wireless stations or nodes. Each of the nodes plays a different role in the network protocol.
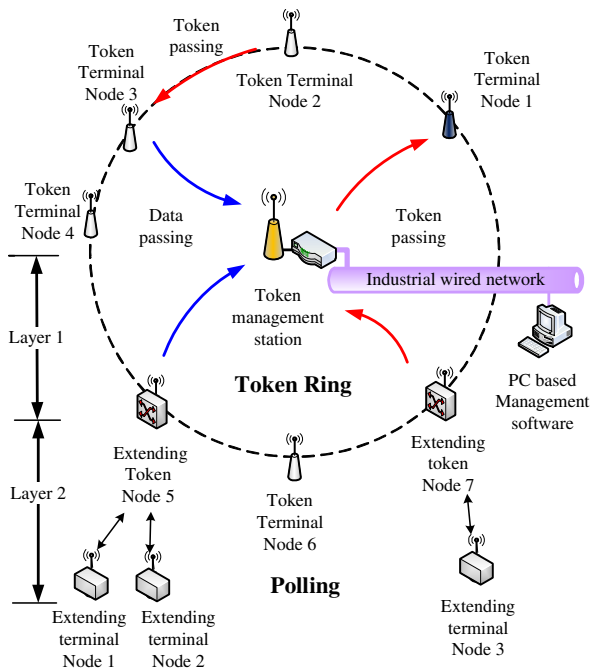

Fig. 1. Topological structure of WICH-TL

A.   Token Ring Based Network Layer

The token ring based network is the backbone and control the wireless channel by passing a token. Stations or nodes whoever get the token may use the channel to send data to the management station. There are three kinds of network stations in the token ring-the token management station, token terminal nodes and extending token nodes. The token management station is a master station which is in charge of maintaining the token ring network operating by producing and listening to the token, creating and refreshing the connection list and sending data packet. The token terminal nodes join in the token ring as a normal slave node. When the token arrive, it sends data to the token management, and passes on the token to the succeeding node. The other type of slave node in the token ring is called token extending node. This kind of node plays two roles in the whole network system. In the token ring based network, it acts as the slave node. The only difference from the token terminal node is that it sends all data from the subordinating network instead of just its own.

B.   Polling-Based Network Layer

The polling-based subnet is the extending layer including token extending nodes and extending terminal nodes. When a token extending node gets the token, it firstly polls and collects data from extending terminal nodes which share the same segment, and then sends data to the token management station. At last, it delivers the token to the succeed one in the token ring. While extending terminal nodes, the other type who acts as slave nodes, keep listening the polling message, and response data when needed.

3   Stack Model

Compared with the ISO/OSI seven standard layers, the protocol stack model covers two of them-data link layer and application layer. The data link layer is divided into two parts; the lower part involves the traditional CSMA/CA protocol to improve the communicating reliability. And the upper part introduces a token scheme to control the occupying each station or nodes. Moreover, a master/slave polling protocol layer is embedded into the upper token layer as the extending part.

4   Frame Format

In order to realize the protocol and manage the network, two kinds of Frame format is designed for management and data exchanging. The figure 2 shows how the devices use the frame to operate the protocol stack.
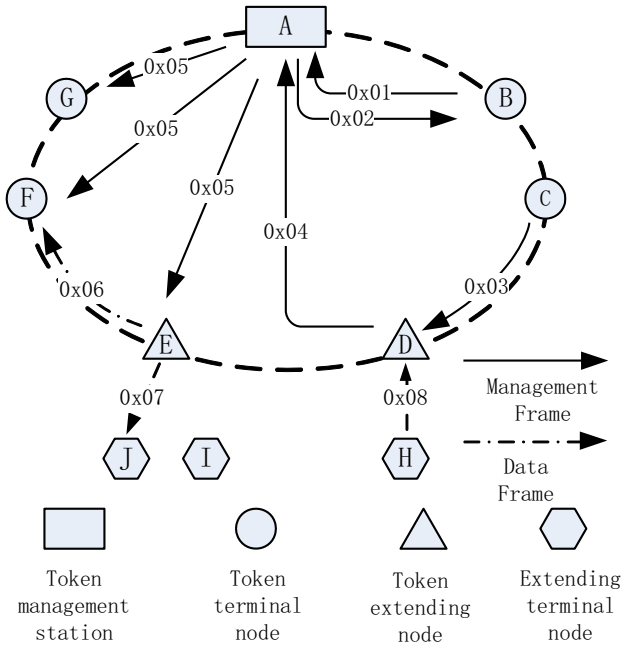
Fig. 2. The component of network protocol stack

A. Management Frames

There are five types of management frames used:

a) **ReqtoNet Frame :** only sent from token nodes. The byte source records the node's address. The byte sequence records the node's sending sequence of this frame. While the byte extend is to distinguish the token terminal and the token extend node.

b) **Connection list Frame :** contains the connection messages and sent from the token management station to the nodes in the token ring.

c) **Token Frame :** sent by each of the devices in the token rings. The byte TRC records the cycle time of the token so that to prevent from receiving the repeat token.

d) **Token Node Frame :** sent only by the management   station. It broadcasts to the related nodes to change their connection list when a new node joins or an old node drops.

e) **Extending terminal Frame :** sent from exteding token nodes to the management station node to informs the management the number and the address of its subordinating extending terminal nodes.

ReqtoNet Frame

8bit	8bit	8bit	8bit
0x01	Source	Sequence	Extend

Connection List Frame

8bit	8bit	8bit	8bit	8bit
0x02	PreAddr	SucAddr	Source	Destination

Token Frame

8bit	8bit	8bit	8bit
0x03	Source	Destination	TRC

Token node Frame

8bit	8bit	8bit	8bit	Number*8bit
0x04	BroadCast	Source	Number	StationAddr

Extending Terminal Frame

8bit	8bit	8bit	8bit	8bit
0x05	BroadCast	Source	Destination	Number

Extend	Terminal

Number*16bit
8bit   8bit

Fig. 3. Management frames

B.   Data Frames

There are three types of data frames used to carrying the measuring data.

Token data Frame

8bit	8bit	8bit	8bit	Length*8bit
0x06	Source	Destination	Length	Data

M/S senddata Frame

8bit	8bit	8bit	8bit	8bit	Length*8bit
0x07	CanSend	Source	Destination	Length	Data

M/S returndata Frame

8bit	8bit	8bit	8bit	Length*8bit
0x08	Source	Destination	Length	Data
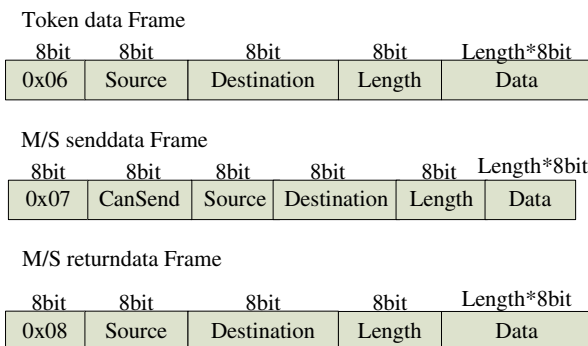
Fig. 4. Data frames

f) **Token Data Frame :** sent by token terminal nodes and extending token nodes.

g   **M/S senddata Frame :** sent from extending token nodes to their slave extending terminal nodes. The byte Can send informs an extending terminal node whether it need to return data or not.

h   **M/S Returndata Frame :** as a return sent from extending terminal nodes to extending token nodes carrying the collected data.

As the figure 3 and 4 show, the first byte of each frame is designed as the frame type byte and is numbered as above. Considering data carrying demand of the short-range wireless network, the Length byte is limited in one byte.

4   Communication Procedure

With the usage of two kinds of network frames mentioned above, the protocol stack operates as three stages. Each of the stages is discussed as follow:

A.   Establishing Stage

The network system first steps into the establishing stage to form a token ring. Token terminal nodes and extending terminal token send Reqtonet Frame to the management station. The management station establishes the connection list and sends the Connection List Frame down to the nodes once the list changes. Token nodes receive the Connection List and form a self-list.

B.   Operating Stage

The station and nodes one by one get the token, occupy the communication channel, send data, and pass on the token. Extending token nodes first poll extending terminal nodes, then send up the data when possess token. The management station sends relist Frame when topology changes. The nodes keep listening to passing process of the token, starts to send the Reqtonet Frame to the management station. The station listens to the request and refreshes the communication list once in a certain period.

C.   Recovering Stage

The management station keeps listening to the passing process of the token, starts the recovering function to rebuild the network when times up (token lost or disconnect with the network).The recovering function may repeat the establishing stage and build a new network.

5   Implementation and Simulation

The WICN-TL can be realized upon different physical layer specifications. This paper chooses the IEEE802.15.4a standard-the first international standard that specifies a wireless physical layer to enable precision ranging.[5] As an amendment to the popular IEEE 802.15.4-2006, the IEEE 802.15.4a-2007 standard introduces two implementations to enable more precise node localization and robust communication.[6]

The network system is designed as two types of devices. For one, the management station, also being the wireless access gateway, takes an industrial level ARM9 core AT91RM920 to be the processor and a real-time operating system VxWorks to realize the multitask scheduling. The wireless standard IEEE802.15.4a is implemented based on the wireless protocol chip NA1TR8 from German NanoTron Company. In order to access to the wired fieldbus, the wireless management station acts as a slave of PROFIBUS –DP fieldbus. For the other, the wireless nodes, including token terminal nodes, token

extending nodes and extending terminal nodes, take a low energy cost single chip ATMEGA128L as the controller, complete A/D converting, thermocouple acquisition and wireless communicating control. The hardware design of devices is illustrated as figure 5.

As for the software design, the VxWorks operating system loaded in access gateway operates two protocol stack tasks and a data converting task. VxWorks has 256 priority levels and assigns the levels for each task respectively. While the nodes do not have operating systems, they normally operate the communicating control function and jump to the A/D converting function by time interrupt.
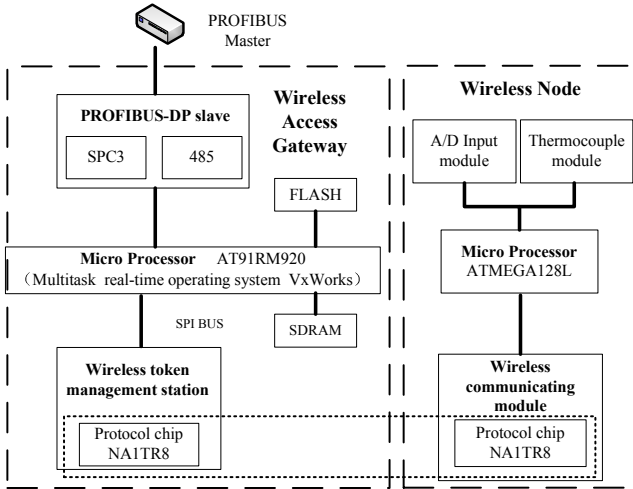

Fig. 5. hardware structure

The network system has been tested in a sewage plant site and topologically illustrated as figure 6. The site test included a token management station A, a token terminal node B2, two extending token nodes B1, B3, and two extending terminal nodes C1 and C2. The three terminal nodes were connected with three different meters respectively. The WinCC software ran on the monitoring PC showed the data from the three meters transmitted by the WICN-TL network system, which verified the network design.

Moreover, the designed network system has been installed on a turbine power simulation platform[7] as figure 7. It is a simulate power plant model which is able to generate electricity by burning the coal gas. The WICN-TL system transmits the measuring data of the platform including pressure and temperature with some other kinds of network system, and it communicate well and the data could be read on the configuration software interface.

Finally, the communicating range of the whole system is measured on the scene. Two circumstances are tested and the communicating range is listed in the table1. One is that the station and nodes are installed on a wide open space scene and no barriers between them. The other case involves a row of trees and wire netting between two devices. To be compared, a previous design, one layer token ring based network protocol (WICN), is also implemented and tested to present the improvement of communicating range. Under the circumstances of some specific distances, the packet losses was recorded and computed.
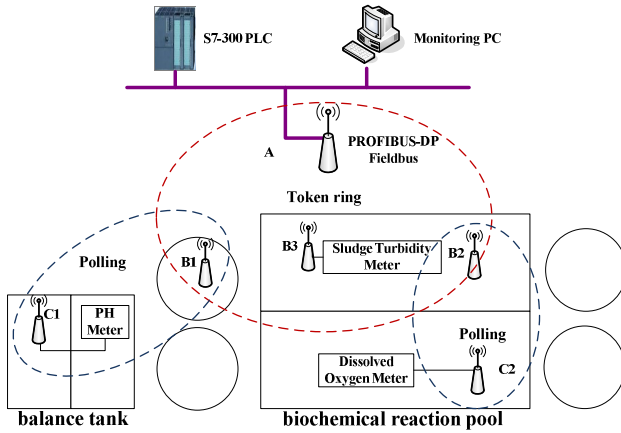
Fig. 6. Sewage Plant Test Topology
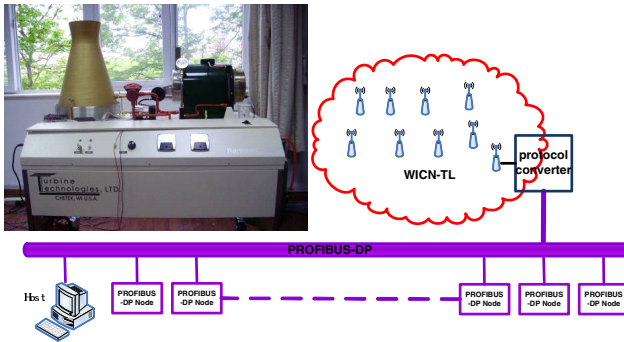
Fig. 7. Turbine Power Simulation Platform Test

The result shows that the wireless network system can cover as far as 200 meter distance at most and pass through normal barriers well.

Table 1. Wireless Network Communicating Quality

Test environment	Network system Component	Network description	Maximum Communicating range	Packet Loss
Wide open space	A management station A token terminal node	WICN	104m	3.43%(50m)
	A management station A token extending node A extending terminal node	WICN	103m	3.86%(50m)
		WICN-TL	207m	8.05%(100m)
A row of trees and wire netting	A management station A token terminal node	WICN	68m	10.91%(50m)
	A management station A token extending node A extending terminal node	WICN	68m	12.63%(50m)
		WICN-TL	135m	18.72%(100m)

6   Conclusion

This paper proposed a two layer network protocol WICN-TL specialized in industrial application. The comparatively simple protocol stack strikes a balance between the complex routing scheme and communicating distance. It is implemented and tested on the industrial scene, and the result verified that it performs better than its predecessor WICN. However, limited by the laboratory condition, both nodes scale and environment scenes are limited so that the performance still needs to be measured. Furthermore, sleeping strategy, security mechanism and time synchronization are all crucial for a wireless industrial network system and need to be added in the next step. At last, the protocol stack should be implemented on more popular standards such as IEEE802.15.4 for a more widely use.

References

1. Sauter, T.: The Three Generations of Field-Level Networks: Evolution and Compatibility Issues. IEEE Transactions on Industrial Electronics 57, 3585–3595 (2010)
2. Wang, H., Hou, W., Qin, Z., Song, Y.: Integration Infrastructure in Wireless/Wired Heterogeneous Industrial Network System. In: Li, K., Fei, M., Jia, L., Irwin, G.W. (eds.) LSMS/ICSEE 2010. LNCS, vol. 6328, pp. 129–135. Springer, Heidelberg (2010)
3. Al Agha, K., Bertin, M.H., Dang, T., Guiiton, A., Minet, P., Val, T., et al.: Which Wireless Technology for Industrial Wireless Sensor Networks The Development of OCARI Technology. IEEE Transactions on Industrial Electronics 56, 4266–4278 (2009)
4. Hou, W.-Y., Liu, W.-C., Fei, M.-R.: A Token-based MAC oriented Wireless Industrial Control Networks. In: 2006 IEEE International Conference on Information Acquisition, pp. 22–25 (2006)
5. Karapistoli, E., Pavlidou, F.N., Gragopoulos, I., Tsetsinas, I.: An overview of the IEEE 802.15.4a Standard. IEEE Communications Magazine 48, 47–53 (2010)
6. Lewandowski, A., Putzke, M., Köster, V., Wietfeld, C.: Coexistence of 802.11b and 802.15.4a-CSS: Measurements, Analytical Model and Simulation. In: 2010 IEEE 71st Vehicular Technology Conference (VTC 2010-Spring), pp. 1–6 (2010)
7. Wei, L., Wang, H.-K., Fei, M.-R.: Integration Design of Measurement and Control System Based on Heterogeneous Network and Its Application in Power Generation Experiment. Process Automation Instrumentation 32, 06–10 (2011)

Application of STM32 Microcomputer
in the Design of Pressure Sensor Compensation

Jianmin Wang[1,*], Yongxin Mou[2], Junqin Huang[1], and Xiaodong Liu[1]

[1] School of Automation Science and Electrical Engineering, Beihang University, Beijing
100191, China
wjm123121@126.com
[2] Huayou Beijing Service Corporation, Beijing 100011, China

Abstract. The use of STM32 microcomputer in the design of sensor compensator is implemented based on the principle of pressure sensor compensation. According to the characteristics of small size, low power consumption and high performance, the smallest system of STM32 microcomputer is designed to install in the compensation link of pressure sensor after debugging successfully. The results of experiment presented the dynamic response of the pressure sensor stepped up, band was widened, and stability was enhanced after adding this compensator, which reached the desired effect and proved the feasibility of this method.

Keywords: Pressure Sensor, Compensation, STM32 Microcomputer, Smallest System.

1   Introduction

Pressure sensor is a measurement device that converts pressure signal into electrical signal. Its output signal determines the control order of the controller, so it plays a crucial role in the whole control system. In the actual pressure test environment, pressure is not fixed, but a changing dynamic quantity with time. Especially in all kinds of control system of the flight vehicle and power systems, the speed of change is more rapid, and its duration time is in a short time of μs. Therefore, it puts forward higher request for the pressure sensor, namely the pressure sensor needs wide frequency band and fast response time[1].Dynamic characteristics of pressure sensor directly impacts on control system for the response ability, so searching for a sensor with an appropriate dynamic performance is particularly important. Because of the limit of sensitive materials, production process, cost and other factors, it is not always easy to find the right sensor to meet the control system dynamic requirements, and improving the dynamic performance of the sensor is the scheme often to choose [2-3].In the aspects of dynamic performance improvements in sensor, Huang Jun qin made outstanding contribution and filled the blank in the field of China. His monograph "Test System Dynamics"[1] brought forward many sensor modeling and compensation methods, which provided practical available skills and methods for the scientific and technical

---

workers, engineering and technical personnel. Based on this, the paper applied STM32 microcomputer in sensor compensation design, and achieved satisfactory results.

STM32 series microcontroller is the products of high performance, low power consumption, low cost by STMicroelectronics launched based on ARM Cortex-M3 kernel. Its working voltage is 2-3.6 V, the clock frequency is as high as 72 MHz, power consumption is only 36 mA, 12-bit A/D converter conversion time can reach 1 μ s, and serial ports (UART) output data rate is as high as 4 Mbps [9]. All the above fully meet the requirements of the sensor dynamic performance. Therefore the design of the sensor compensator with this microcomputer is reasonable.

2   Compensation Principle of Sensor

The dynamic characteristics of the sensor and its transfer function of pole position are close. On the principle of zero pole cancellation, its pole is cancelled   by combination of a compensation link, and the new pole is joined and adjusted to desired location, which improves the dynamic performance of the sensor [4~6]. When a sensor input and output data is given, it is needed to establish mathematical model for the sensor. After obtaining the transfer function, it is needed to find zero and pole to cancel and then design compensation link. Figure 1 shows the step response curve for a type of pressure sensor.

According to the methods of modeling in reference [1], the transfer function of the sensor model is

$$G(s) = \frac{0.01011S^2 - 7359S + 1.804 \times 10^9}{S^2 + 1.045 \times 10^4 S + 1.157 \times 10^9}.$$ (1)

Where,   the   damping   ratio   is $\xi = 0.154$,   the   natural   frequency   is $\omega_n = 3.4 \times 10^4 rad / s$.   This model fitting results is shown in Fig. 2.
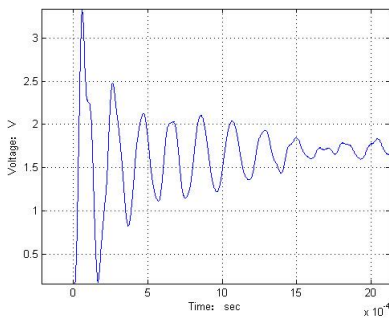

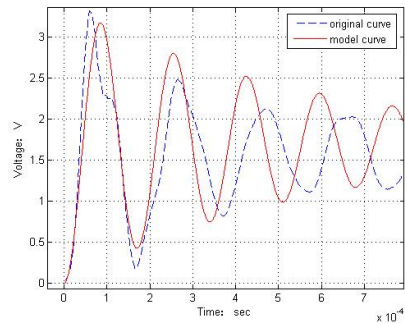**Fig. 1.** Step Response Curve of a Pressure Sensor   **Fig. 2.** Model Fitting Result

From Fig. 2, the nominal model well fits the original plant, therefore we can use (1) as the design model of the pressure sensor to compensate.

From Fig. 1 and the transfer function (1), sensor damping is small and oscillation is large. In order to get the oscillation smaller and the faster response speed, it needs combination of compensation link. Here the design of the zero-pole destructively compensator is shown as

$$H_1(s) = \frac{S^2 + 1.045 \times 10^4 S + 1.157 \times 10^9}{S^2 + 2 \times \xi' \times \omega_n' \times S + \omega_n'^2} \quad . \tag{2}$$

where, $\xi' = 0.6$, $\omega_n' = 2 \times \omega_n$。

The zeroes of (2) counteract the poles of (1), and the poles of (2) are set to new desired positions. Here establish the second order system as the result of the compensation, whose damping ratio is 0.6 and natural frequency is two times larger than original system, namely

$$G'(s) = G(s)H_1(s) = \frac{0.01011S^2 - 7359S + 1.804 \times 10^9}{S^2 + 2 \times 0.6 \times (2 \times \omega_n)S + (2 \times \omega_n)^2} \quad . \tag{3}$$

In order to realize the microcontroller compensation, it should discrete compensation links. Based on bilinear method, (3) is transformed into the form of difference equation

$$y(n) = 1.65 \times 10^7 x(n) + 3.3 \times 10^7 x(n-1) + 1.65 \times 10^7 x(n-2) \\ + 1.575y(n-1) - 0.6691y(n-2) \quad . \tag{4}$$

Write the difference equation shown in (4) into the program of microcomputer, then it can realize dynamic performance compensation effect for pressure sensor.

3   Hardware Design

System hardware mainly includes power part, smallest microcomputer system, A/D converter interface, data output interface and so on. The introduction scheme of its main circuit is shown as follows.

3.1   Power Part

Power part supplies energy for microcomputer compensator and its circuit is shown in Fig. 3.

The pressure sensor compensator based on microcomputer should be able to install in the sensor circuit, and request the volume smaller and the power consumption as low as possible. Therefore, it is required to use battery power supply, which is easy to carry and works for long hours. LT3973is a regulator that accepts a wide input voltage ranging up to 42V, and consumes only 2.5μA of quiescent current. Moreover, it maintains high efficiency at low output currents while keeping the output ripple below 10mV in a typical application, and has a low shutdown current of 0.75μA. Applied to the design of compensation as much as possible in reducing the interference of ripple effects, it also can try to minimize the loss of the battery power and extend use cycle. Reference voltage of microcomputer A/D uses external reference voltage, as shown in Fig. 4.
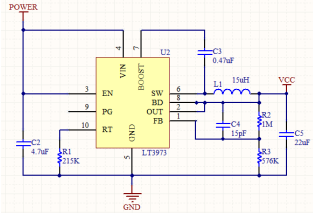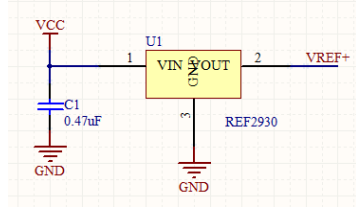
Fig. 3. Power Circuit

Fig. 4. Circuit of Reference Voltage

Stable voltage VCC from Fig. 3 circuit transfers into accurate power VRER+ through conversion of circuit for Fig. 4, which is as the reference voltage of A/D convertor.

3.2   Smallest Microcomputer System

Smallest microcomputer system includes power supply, crystals, system reset, JTAG interface, serial port and so on. STM32 microcomputer contains reset circuit, and needs no exterior. The clock includes three clock sources: the internal high speed of 8M clock signal, the external high speed clock signal and the external low speed of 32.768 kHz crystals. In order to track the dynamic characteristics of the sensor, the system requests A/D converter with very high sampling rate, so it puts forward very high demand for the clock. For meeting the requirement of high speed acquisition, the system uses external 8M crystals, and amplifies frequency to 72MHz through the phase locked loop(PLL)[7~9]. JTAG interface is used to download program and debug. It is often used early in the design and will not be used once finish debugging. Serial port is used to output compensation data, and its transmission rate will affect the dynamic performance of sensor tested. The system uses the most simple three wire serial connection mode, and outputs internal data to file. Hardware structure principle diagram of smallest microcomputer system is shown in Fig. 5. Real picture of smallest microcomputer system is shown in Fig. 6.
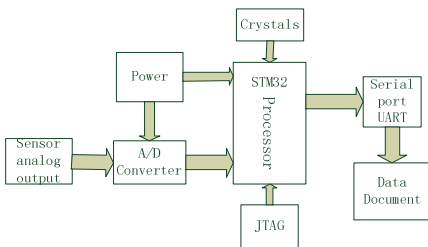
Fig. 5. Hardware Structure of System

Fig. 6. Real Picture of Minimum System

4   Software Design

The role of microcomputer in the whole system is: sampling analog output of sensor, transferring to discrete data, then doing pre-design compensation treatment for these

data, real-time outputting compensation data, at the same time, saving to document for following use. As close as possible to meet the requirement of real-time, two pieces of cache are opened in programming for saving data and transmitting data in turns. Software flow chart is shown in Fig. 7.
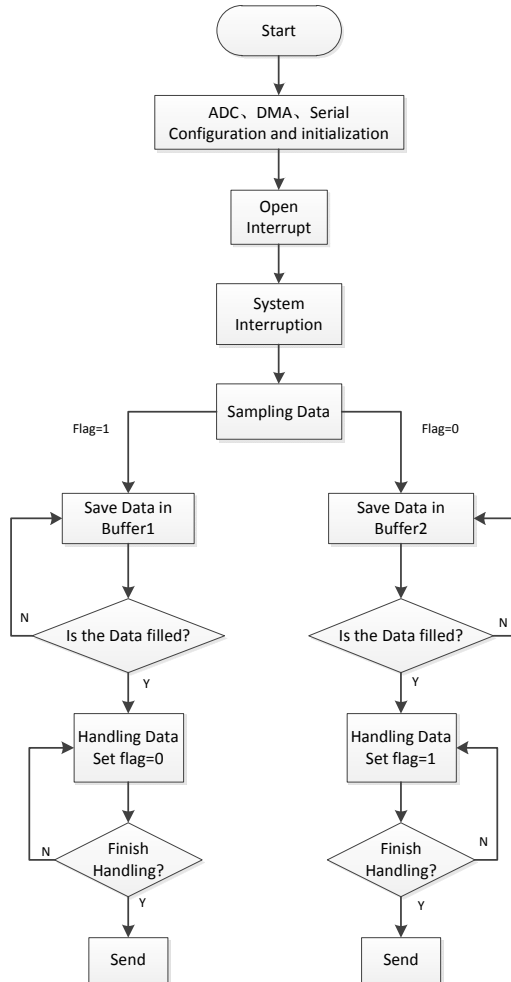
Fig. 7. Program Flow Chart

The data processing part in the program flow chart is sensor compensation link mentioned in the second section of the paper. Microcomputer continuously samples sensor output, handles sampling data by compensation processing link, then outputs data through serial, and finally realizes the sensor dynamic compensation. Part program is shown as follows.

```
if(Send_flag)
    {if(SaveData_Index == SaveData_0)
    { //flag=1, start to save data in buffer 1,handling data
 of buffer 0
      for(i=0;i<300;i++)
      { x3=SaveData_1[i];      //import new input variable
 Data[i]= 1.65e7*x3+3.3e7*x2+1.65e7*x1+1.575*y2-0.6691*y1;
                          //algorithm of compensation
          x2=x3;
          x1=x2;                      //input varible deliver
          y1=y2;                     //output varible deliver
          y2=Data[i];
          SaveData_1[i+300]= Data[i];
                       //handled data save in send buffer
      }
      SendData(SaveData_1, 600); //data in buffer handled
 after and send to serial port
    }
    else      //flag=0, start to save data in buffer 0,handling
 data of buffer 1
    {for(i=0;i<300;i++)
      {x3=SaveData_0[i];
Data[i]= 1.65e7*x3+3.3e7*x2+1.65e7*x1+1.575*y2-0.6691*y1;
                          //algorithm of compensation
          x2=x3;
          x1=x2;                      //input varible deliver
          y1=y2;                     //output varible deliver
          y2=Data[i];
          SaveData_0[i+300]= Data[i]; //handled data save in
 send buffer
      }
       SendData(SaveData_0, 600);//data in buffer handled
 after and send to serial port
    }
```

5   Results of Experiment

The sensor compensator based on STM32 is tested on the sensor in Fig. 1, and the result is shown in Fig. 8. The bode diagram of original plant and new plant added compensator is shown in Fig. 9.

From Fig. 8 and Fig. 9, we can conclude that through microcomputer compensation, pressure sensor response time becomes short, oscillation becomes weak, dynamic performance improves with a certain degree, and the system response is more stable in
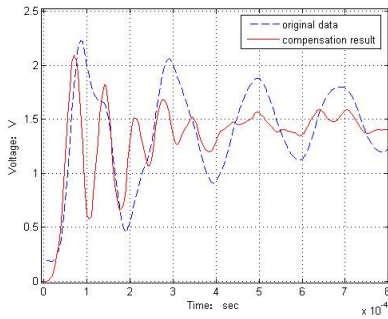
Fig. 8. Compensation Result in Time Domain
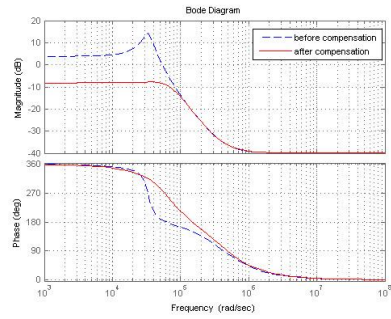
Fig. 9. Compensation Result in Frequency Domain

the pass-band. From Fig.9, it is obvious that the peak in amplitude-frequency curve gets well suppression after compensation, the inflection point in phase-frequency curve disappears, and then the curve becomes smooth. Thus the method is proven to be correct and practical, and it is feasible to act as sensor dynamic compensation with STM32 microcomputer.

6   Conclusions

This paper introduces the principle of pressure sensor dynamic compensation, and the method of using STM32 microcomputer in sensor dynamic process of compensation. It makes full use of characteristics of microcomputer including high performance, low power consumption, high speed, easy design and installment, and reliable work for long time. The system gives a good solution for pressure sensor in overcoming shortage of dynamic performance in some cases, which makes the pressure sensor measure well in real-time environment with fast changes and short response time. It provides the prerequisites for the controller well controlling system.

References

1. Huang, J.Q.: Test System Dynamics. National Defense Industry Press, Beijing (1996)
2. Zhang, A.P., Zhang, Z.J.: Correction and Improvement for Dynamic Performance of Pressure Sensor. Journal of North China Institute of Technology 19(1), 8–10 (1998)
3. Ye, X.B., Zhang, W.N.: Analysis and Improvement for Dynamic Performance of Pressure Sensor. Journal of Transducer Technology 21(2), 11–14 (2002)
4. Yu, D.C.: Analysis and Calibration for Dynamic Characteristic of the Pressure Sensor. China Instrumentation 7, 9–11 (2003)
5. Zhu, Z.H., Ding, J.N.: Research Progress of Dynamic Characteristics of Sensor. Machinery & Electronics 6, 75–77 (2008)
6. Wang, D.B.: Investigation of Dynamic Performance for Pressure Transducer. Measurement & Control Technology 27, 307–312 (2008)

7. Yao, Z.H.: Studying on Portable Detecting Instrument of Vibration Signal Based on STM32. Techniques of Automation & Applications 30(2), 45–46, 50 (2011)
8. Zhai, Y.J., Wu, Z.C., Shen, C.S.: Design of Standardized Module of Sensor Interface Based on STM32. Electronic Technology 3, 57–60 (2011)
9. Wang, Y.H., Xu, W., Hao, L.P.: STM32 series ARM Cortex-M3 microcontroller principle and practice. Beihang University Press, Beijing (2008)
10. Zhao, Y.: Research on Nonlinear Correction and Dynamic Compensation of Sensors. Shanghai Jiaotong University, Shanghai (2007)

The Mechanism of Abnormal Detection and Distributed Localization of Nodes Based on Trust Management in WSN

Kun Chang[*], Qingwei Liu, Mandan Liu, and Hailong Xiong

Key Laboratory of Advanced Control and Optimization for Chemical Processes
(East China University of Science and Technology), Ministry of Education, Shanghai 200237
liumandan@ecust.edu.cn

Abstract. The fault of Beacon nodes, as well as the interference in radio channel environment, will all affect the performance of localization. To eliminate these interference factors, a new distributed localization algorithm is proposed based on trust management. Firstly the algorithm splits WSN into several clusters with a certain cluster protocol, and then initiates the global trust rating (GTR) and local trust rating (LTR) of Beacon nodes. To get more precise location, it needs to localize frequently and update the GTR and LTR of beacon nodes. Theoretical analysis and simulation results show that the method can not only improve the precision of the positioning, but can also be used to examine abnormal beacon nodes and infer possible interference factors in wireless channel.

Keywords: WSN, Localization, GTR, LTR, Abnormal detection, Cluster.

1 Introduction

Wireless Sensor Network (WSN) is widely used in daily life and military equipment. Localization of the WSN nodes is one of the most important aspects of WSN technologies. As the WSN application fields spread, higher precision and anti-interference ability of localization is required. Apparently, classical traditional localization algorithms such as DV-hop, Central id algorithm, APIT algorithm can't afford actual requirement[1]. In paper [2], the authors proposed the BOA algorithm, which could get higher precision relatively in the condition of interference. In paper [3], they proposed LLV to measure the influence of localization error on the network and to analysis the possibility of localization redress. Then, they used LV-FEED to redress localization results and to improve the localization precision. Recently, Lazos[4] proposed an safety localization algorithm, which has no relation with distance. The algorithm can avoids the attack of malicious Beacon nodes, but it can do nothing to abnormal nodes nearby.

This paper proposed an Distributed Localization algorithm Based on Trust Rating (DLBTR) . This algorithm can not only promote the precision of localization, but also can detect abnormal nodes and avoid radio channel interference to localization.

2   Distributed Localization Based on Trust Rating

This paper proposed a DLBTR algorithm. This algorithm abstracts sensor node as social individual and build a trust evaluation mechanism to quantify the liability of sensor nodes. Next, we will give some explanation and demonstration on the trust evaluation mechanism and locating computation.

2.1   Network Structure

In order to achieve distributed localization, we need to cluster the network first. We designed an clustering method basing on virtual grid. The process of clustering network consists of two stages: cluster set-up stage and cluster steady stage. In the cluster set-up stage, the base station splits the surveillance area into several virtual grid after the sensor nodes deployed. Every virtual grid is the prototype of cluster. First, Beacon nodes will find the nearest virtual central point and join the cluster. Then, Node finds the nearest Beacon nodes by signal intensity and joins its cluster. In the steady stage, cluster member doesn't change. Node keeps collecting data and sending data to the cluster head frequently.

2.2   Trust Rating Evaluation

In the actual locating process, there always exist several abnormal Beacons. These Beacons can't provide accurate information about location coordinates, signal strength etc, leading to incorrect localization, or even unable to locate. There are   two aspects of Beacon node being abnormal: Between Beacon and Node exists radio channel interference; Because of lacking of electricity or hardware failure, a Beacon provides inaccurate information to all nodes.

In order to eliminate abnormal Beacon's influence to Node, we set trust rating for every Beacon. Node will verify the location information of the reference point in each round of localizing computation. If a Beacon is found suspicious, the Node will punish it, namely decrease its trust rating. Conversely, the Node will award the Beacon, namely increase its trust rating. For the first case, the Node will think the Beacon as abnormal node and decrease its trust rating. However, the Beacon is still a trusted reference node to all the other Nodes. Thus, Cluster Head set two trust ratings for each Beacon in case one Node's punishment got two much influence on the Beacon. As follows:

Local Trust Rating (LTR): One single Node's evaluation to Beacon. It's a standard for a Node to choose its localizing reference nodes. Every LTR has an appropriate Local Trust Threshold (LTT). If a Beacon's LTR is lower than its LTT at this round,

this Beacon couldn't take part in this Node's localization. The LTR of Beacon$i$ near Node in the nth round is:

$$f_{trust}[(n,Beacon_i)|Node]=\begin{cases}\max\{f_{trust}[(n-1,Beacon_i)|Node]-\dfrac{k_1}{LocationNum},0\} & flag=0\\[2mm]\min\{f_{trust}[(n-1,Beacon_i)|Node]+\dfrac{k_2}{2*LocationNum},1\} & flag=1\\[2mm]\min\{f_{trust}[(n-1,Beacon_i)|Node]+\dfrac{k_2}{LocationNum},1\} & flag=2\end{cases}$$

$$i=1,2,3,4 \tag{1}$$

In equation (1), n is localization round. $LocationNum$ is the total localization rounds. k_1 is punishing coefficient, and k_2 is awarding coefficient. LTR initialize with 1. There exist three situations in trust rating computation:

flag = 0: In the four Beacons, $Beacon_i$ is suspicious. We decrease its LTR.

flag = 1: There is no suspicious Beacon . We increase Beacon's LTR.

flag = 2: In this round localization, we find a suspicious Beacon, but it's not $Beacon_i$. We increase $Beacon_i$'s LTR.

Global Trust Rating (GTR): GTR is a base trust. It's a complicated trust evaluation to a Beacon. When a Beacon's GTR is lower than global trust threshold, the Cluster Head declears that the Beacon is invalid. The GTR of $Beacon_i$ in the nth round is:

$$F_{trust}(n,Beacon_i)=\frac{\sum_{j=1}^{NbrNodeNum}f_{trust}[(n,Beacon_i)|Node_j]}{NbrNodeNum} \quad {}_{i=1,...,BeaconNum} \tag{2}$$

In equation (2), *NbrNodeNum* is the number of Nodes near $Beacon_i$. *BeaconNum* is the total number of Beacons.

2.3   Calculation of Node's Location

In the cluster, Cluster Head calculate Nodes' location by localizing frequently, and every node's neighborhood list information will be used in this progress. In each round of localization, if the number of a Node's neighborhood Beacon less than 3, this Node wouldn't be localized in this round. If the number is 3, we calculate the node's location by 3-node localization algorithm. If the number is bigger than 3, we calculate the Node's location by 4-node localization algorithm. We choose four Beacons as reference nodes in the rule of highest trust rate and lowest distance. At the same time, the Beacon's LTR will be updated. At the end of per round, all Beacons' GTR will be also be updated. After many rounds localization, abnormal Beacons' trust rating will decrease. Finally, it will be eliminated. Therefore, the precision of localized node's location gets up rounds by rounds, while the number of unlocalized Nodes will decrease gradually.

The rule of choosing reference nodes: In the first round localization, we choose the nearest Beacons as reference nodes. After the first round, we choose the highest trust rating Beacons as the Node's reference nodes. If the trust ratings of two Beacon are equal, we will choose the nearer one. If all the chosen Beacons are suspicious in the late round, all the Beacons will be eliminated. Then we will choose four Beacons randomly to redress search direction and prevent the algorithm falling into local optimum combination.

According to the rules of choosing reference nodes, Node's localization precision will grow up quickly. At the same time, we can find out the abnormal Beacons by evaluating the Beacons' trust ratings.

Here are some assumptions. Node's coordinates are (x, y). Its neighborhood Beacons' coordinates are $Beacon(x_1, y_1)$, $Beacon(x_2, y_2)$, $Beacon(x_3, y_3)$, $Beacon(x_4, y_4)$. The distance between Node to Beacon nodes are d_1, d_2, d_3, d_4.

3-node localization : Choose three Beacons as reference nodes. We got :

$$\begin{cases} (x-x_1)^2 + (y-y_1)^2 = d_1^2 \\ (x-x_2)^2 + (y-y_2)^2 = d_2^2 \\ (x-x_3)^2 + (y-y_3)^2 = d_3^2 \end{cases} \quad (3)$$

The objective function:

$$\min f = \sum_{i=1}^{3} |(x-x_i)^2 + (y-y_i)^2 - d_i^2| \quad (4)$$

We use least square method to solve this optimal problem.

4-node localization algorithm: Choose four Beacons as reference nodes. We got:

$$\begin{cases} \begin{cases} (x-x_1)^2 + (y-y_1)^2 = d_1^2 \\ (x-x_2)^2 + (y-y_2)^2 = d_2^2 \\ (x-x_3)^2 + (y-y_3)^2 = d_3^2 \end{cases} \\ (x-x_4)^2 + (y-y_4)^2 = d_4^2 \end{cases} \quad (5)$$

Firstly, we use the former 3 equation in (5) to calculate Node's Coordinates $N_1(x_1^x, y_1^y)$ by 3-node localization algorithm. $f_{error}(1)$ is the sum of localization errors between N_1 and the former three Beacons.

$$f_{error}(1) = \sum_{j=1}^{3} |(x_1^x - x_j)^2 + (y_1^y - y_j)^2 - d_j^2| \qquad j = 1,2,3 \quad (6)$$

Secondly, the distance error between N_1 and Beacon (x_4, y_4) is:

$$\Delta_1 = \left| (x_1^x - x_4)^2 + (y_1^y - y_4)^2 - d_4^2 \right|^2 \quad (7)$$

And so on, we get :

$$\mathbf{f}_{error} = [f_{error}(1), f_{error}(2), f_{error}(3), f_{error}(4)]^T \tag{8}$$

$$\Delta = [\Delta_1, \Delta_2, \Delta_3, \Delta_4]^T \tag{9}$$

At last, we can calculate Node's coordinate Node $(x_5^{\;x}, y_5^{\;y})$ and these four Beacon node's distance errors F_{error}. In order to choose the most precise coordinate from Node $(x_i^{\;x}, y_i^{\;y})(i=1,2,3,4,5)$, we set a choosing function (10). The value of function (10) will be a standard to choose the optimal coordinate Node (x, y).

$$g(i) = \frac{\Delta_i}{f_{error}(i)+1} \qquad i = 1,2,3,4 \tag{10}$$

If $F_{error} \le \min(f_{error}(i)), (i=1,2,3,4)$, no Beacon nodes is abnormal or there exists at least 2 abnormal Beacon nodes. $Node(x,y) = Node(x_5, y_5)$.

If $F_{error} > \min(f_{error}(i)), (i=1,2,3,4)$, one Beacon node has an serious effect on Node's localization and it will be eliminated out and punished. $Node(x_i^{\;x}, y_i^{\;y})$ corresponds with the maximum of $g(i)(i=1,2,3,4)$ will be the best choice of this round.

2.4   Validation

We will simulate DLBTR algorithm and mechanism of abnormal detection on Matlab7.0. Table 1 are the parameters of simulation experiment.

Table 1. The parameter of simulation experiment

Parameter	Value
Size of network field	$100 \times 100 m^2$
Beacon number	20
Node number	30
Beacon's communication radius (R)	$30m$
Node's communication radius (r)	$15m$
Number of clusters	4

After nodes deployed, we cluster the network using the method that we said before. Sink allocates unique ID to every node. Then, Beacon's LTR and GTR initiate with 1. According to the size of the network, it is the best choice to set GTT and LTT as 0.78 and 0.5. In order to validate DLBTR algorithm and mechanism of abnormal detection only, we assume there is no measurement noise between normal nodes.

Abnormal Nodes Simulation: If a Beacon is a malfunction node, we add interference to measurement distances between Beacon and all its relation Nodes to simulating this malfunction. If the Beacon itself is normal, but there exist obstacles between this Beacon and a Node, we add interference to measurement distances only between them. Figure 1 is the network model of WSN on Matlab7.0. Each cluster is an fractional WSN in the figure.

Fig. 1. The network model of WSN
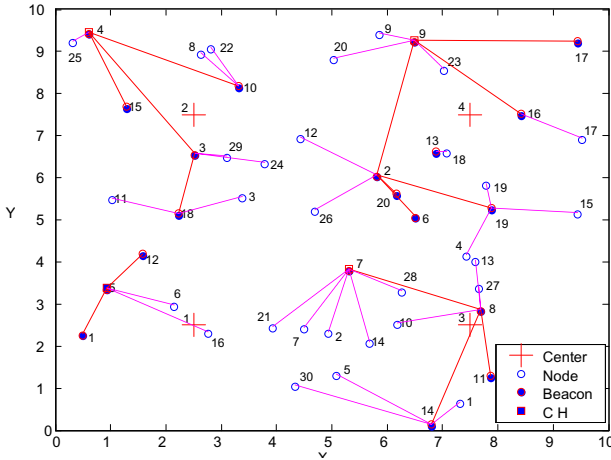
Experimental Model 1: Simulating Beacon's malfunction
In the 4th cluster, Beacon2, Beacon 13, Beacon 20 are malfunction nodes. There exists $20\% d_{ij}$ measurement distance error between these Beacons and their one jump neighbor Nodes. d_{ij} is the real distance. Figure 2 is the GTRs of Beacon 2, Beacon 20 and Beacon 13.
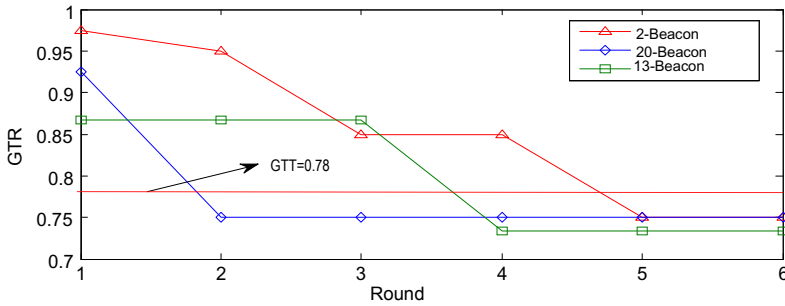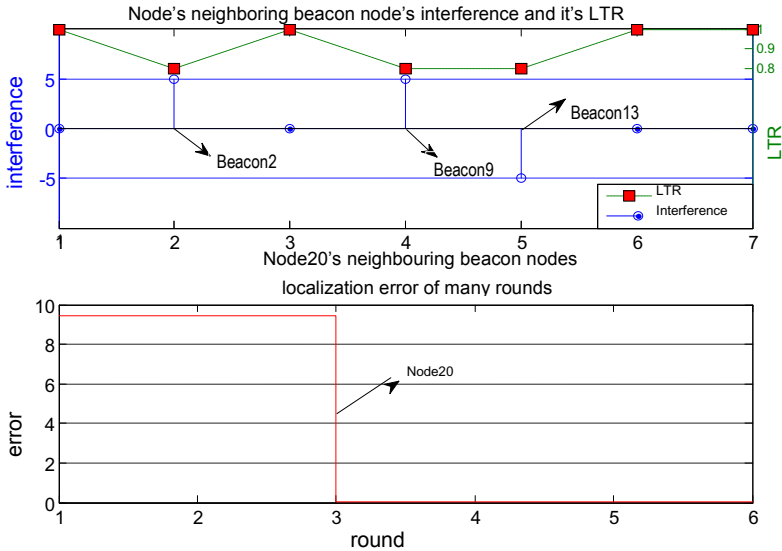

Fig. 2. The GTR of Beacon2, Beacon20 and Beacon13

From figure 2, we know most of Nodes' localizations weren't affected by the malfunction Beacons. Though Node 23 and Node 26 got big localization errors in the first two rounds, these malfunction Beacons' influence get smaller and smaller as the trust ratings of Beacon 2, Beacon 13 and Beacon20 decrease. When Node's trust rating lower than GTR, it will exit the network. Meanwhile, Nodes with low connectivity couldn't be localized because of lacking of localization reference points.

In this experiment , GTT is 0.78. Figure 2 shows that GTRs of Beacon2, Beacon13, Beacon 20 are lower than GTT in 5th ,4th and second round respectively, which means these Beacons are malfunction nodes. Therefore, the Cluster Head announces that these Beacons are invalid. So they won't affect other Nodes' localization.

Experimental Model 2: Simulating obstacles interference

In this experiment, we add interferences to radio channels between Node20 and Beacon2, Beacon9, Beacon 13 by increasing measurement distance error between Node and Beacons. Figure 3(the upper) shows, though radio channel affects Beacon's LTR, it doesn't affect Beacon's GTR. In figure 4, GTRs of Beacon2, Beacon 20 and Beacon 13 are all higher than GTT. Line1 and 2 show, because Beacon13 and Beacon 9 were trusted by other nodes, their trust rating turned to get up after being dropped by Node20. Therefore, this kind of radio channel interference only affects Node20, and it never affects other Node's localization. Figure 3(the lower) shows that abnormal Beacon's trust rating decreased after many rounds localization. In the third round, localization error disappeared because of abnormal Beacon's eliminated.

Fig. 3. The neighbor Beacons of Node20 and the localization error of rounds
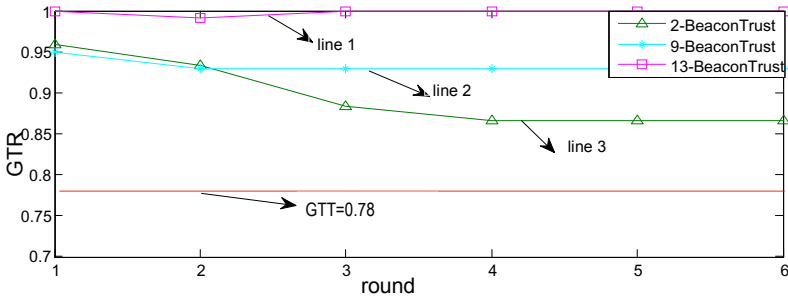
Fig. 4. The GTR of Beacon2, Beacon9 and Beacon13

3   Simulation Results and Analysis

MDS-MAP is an centralized localization algorithm based on MDS, and it's firstly proposed in paper [5]. In paper [6], Yi Shang, Wheeler Ruml etc, improved MDS-MAP and presented a distributed computing model. In this paper, we call the improved MDS-MAP algorithm as DMDS-MAP algorithm. Considering the similarity of network topology and computing steps between DLBTR and DMDS-MAP, we will compare this two algorithms combining with maximum likelihood method in localization precision, coverage, and other aspects to explore the advantages and disadvantages of the DLBTR algorithm.

In the next experiment, a wireless sensor network is uniformly deployed using parameters in table1except the node number. We set the node number as 100. The average measured distance error between Beacons and Nodes is $5\% d_{ij}$. d_{ij} is the distance between node i and node j. Set the computing time of maximum likelihood method as 1 unit. After 50 independent experiments, we got table 2:

Table 2. The comparison of positioning error and computation under 50 independent experiments

SCALE	MLM	DMDS-MAP	DLBTR	MLM	DMDS-MAP	DLBTR
	Localization error $\varepsilon(\%)$			Computing time		
100	31.4	15.5	17.2	1	156	82

As table 2 shows, at the same network environment, DMDS-MAP get the highest locating precision, while the next is DLBTR. However, the computing time of DMDS-MAP is the longest, which means its computational complexity is the biggest relatively. That's because the DMDS-MAP uses the MDS technology to build the relative plot. The use of the singular value composition in MDS involves a large number of matrix operations. The DMDS-MAP algorithm gets high locating precision, but high computational complexity demands high computing and storage capacity of the node. However, DLBTR algorithm improves locating precision in the use of certain optimization method on the base of maximum likelihood method. Therefore, DLBTR algorithm can save half of the computing times than DMDS-MAP algorithm when they

get the same locating precision. Thus, DLBTR algorithm is better than DMDS-MAP algorithm in terms of computational complexity and locating precision.

We have another simulation by changing the percent of beacon nodes into 40. After the simulation, we got figure 6. Figure 6 shows, maximum likelihood method and DMDS-MAP algorithm are very sensitive to abnormal Beacon nodes. As abnormal Beacon nodes grow, Node's positioning error gets up quickly. However, the DLBTR algorithm's localization error is very flat. This verifies DLBTR has the ability to detect abnormal Beacon nodes. The algorithm ensures the precision by eliminating abnormal nodes. Thus, DLBTR has strong fault tolerance and adaptability in unstable network.
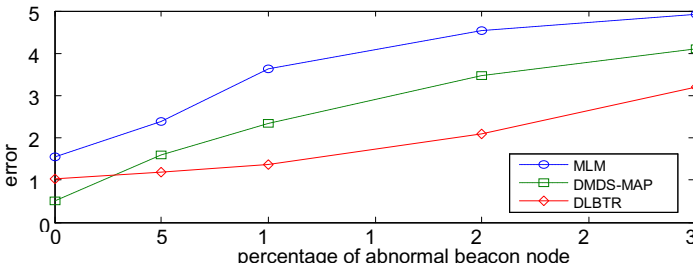
Fig. 5. The relationship of abnormal Beacon node's number and localization error

4   Conclusions

This paper proposed a Distributed Localization Based on Trust Rating (DLBTR) algorithm. DLBTR gets high locating precision by locating frequently and examining the liability of Beacon nodes and localization information. We evaluated the advantages and disadvantages of DLBTR algorithm by experimenting many times. Results show that DLBTR has the advantages of low computational complexity, high locating precision and strong fault tolerance and adaptability in unstable network. Validation of DLBTR algorithm shows it can detects abnormal Beacon nodes and radio interferences between nodes by analyzing Beacon nodes trust ratings.

References

1. Peng, B., Mautz, R., Kemp, A.H., Ochieng, W., Zeng, Q.: On the Effect of Localization Errors on Geographic Routing in Sensor Networks. In: Proceeding of IEEE Communications Society Subject Matter Experts for Publication (2008)
2. Moussa, A., El-Sheimy, N.: Localization of wireless sensor network using Bees Optimization Algorithm 30(5), 478–481 (2010)
3. Dawood, A.A., McNair, J.: On the interaction between localization and location verification for wireless sensor networks. Computer Networks 52, 2713–2727 (2008)
4. Abbasi, A.A., Younis, M.: A survey on clustering algorithms for wireless sensor networks. Computer Communications 30, 2826–2841 (2007)
5. Shang, Y., Ruml, W., Zhang, Y., et al.: Localization from mere connectivity. In: Proc. of ACM MobiHoc, Annapolis, MD, NY, pp. 201–212 (2003)
6. Shang, Y., Ruml, W., Zhang, Y., et al.: Improved MDS-based localization. In: Proc. of IEEE Information, Hong Kong, China, pp. 2640–2651 (2004)

Behavior Event Flow Analysis: A Method of Combat Simulation Creditability Evaluation

Ming Sun[1], Yalong Ma[1], and Huixian Tao[2]

[1] Department of Equipment Command and Administration,
Academy of Armored Force Engineering, 100072 Beijing, China
[2] Department of Computer Science and Engineering,
North China Institute of Aerospace Engineering, 065000 Langfang, China

Abstract. Credibility is an important part of combat simulation research, which restricts application qualities of simulation in the operational decision-making. Most of credibility evaluation methods measure the credibility by analyzing gap between simulation results and actual results, but the actual results of combat cannot be predicted. This paper brings forward a credibility evaluation method based on behavior event flow analysis, using the same form of event flow graph models to describe the battle course in simulation and theory. According to the merger thoughts, this paper also gives a strategy of contrast between the two event flow models, which can measure the flow rationality of simulation models by analyzing the event coverage degree and process matching degree between simulation logic models and theoretic logical models.

Keywords: event flow analysis, combat simulation, creditability, evaluation.

1 Introduction

With the development of simulation technology, combat simulation has become an important assistant to operational decision-making, and can give support to verify operational plan and analysis support requirement. However, combat simulation without enough credibility does not make any sense, or even misleads the decision-making. Therefore, combat simulation credibility evaluation is the basic premise to ensure the combat simulation applications.

Currently, simulation credibility research has been widely recognized and made great progress in the concepts, theories, methods, tools and standards. But in the combat simulation credibility evaluation research has encountered a bottleneck: the majority of the credibility evaluation methods analyze the gap between the simulation output and the actual object output based on statistical principles, which are applicable to the objects with abundant historical data or experimental samples. However, our army lacks of combat experience under modern conditions, and we can not experiment with the battle. Therefore, credibility can not be analyzed by simulation results only.

We note that the battle complexity is reflected in the complicated relationships of a large number of entities, and also reflected in the uncertain combat process. These behavior relationships and combat processes are showed by the entity's behavior

events and their logic relationship in the combat simulation models. If we can analysis the rationality of the events and their relationships in the combat simulation, the credibility evaluation of combat simulation models and process can be achieved. The basic idea as follows: According to the combat simulation results data and simulation process data, extract the dynamic behavior relationships and their action process between the entities and get the behavioral logic models using events as basic description unit; based on theoretical analysis and expertise, establish experiential event flow models in the same operational background as the reference standard of credibility evaluation; describe the two models with the same form and contrast between the both, analysis the event coverage degree and process matching degree between simulation logic models and theoretic logical models.

2   Elements of Event Flow Models

Behavior, which can be ascertained from exterior, is any change of system in its environment. Usually, behavior is expressed by event, activity, state, course and so on. Among these expression forms, event and state are the basic information of simulation. Event can be used to describe the qualitative change of state, including the key point of qualitative change. The behavior process analysis regards the qualitative change part of state, but not the middle state. State and its change can be described by event attribute or event flow. Event and activity can be transformed into each other, and two correlative events can compose an activity. And the course including events and activities sequence can be also transformed into event logic relationship. Therefore, the paper selects "event" as the basic description unit of combat behavior flow.

It is the basic premise of contrast analysis to describe the simulation process and experiential combat processes based on the unified event flow models. Such logical models should at least meet three requirements: First, the models can support the expression of the events and their logic relations; second, the model elements can be extracted and transformed from the combat simulation results easily and directly; third, the models can support the structured contrast.

Directed graph model is an intuitive description method of logic relationship, with good mathematical foundation and structural foundation, and can support a variety of graph-based intelligent algorithm. The common logic relational models are described with directed graphs [1], such as Bayesian networks, dynamic cause and effect graphs, fault propagation graphs, Euler net, Petri net, signed directed graph and so on. These models often contain the only two elements, node and edge, but lack explicit description of complex logic relationships such as concurrency, conflict and so on. For example, Bayesian reliability network uses the different paths from the same node to express the different events led by this node, but the logic relationship between the paths can not be clearly expressed.

Based on the directed graph theory, the paper proposes a kind of description model of operational behavior events process, called Event flow Graph (EFG). The model including event nodes, the logical sides and relation nodes uses event graph cell as

basic element and combines the sequence, concurrency, and conflict logic expression mechanism.

2.1   Event Node

The event node is the basic element of the EFG model, which is used to describe behavior events and expressed by circularity.

2.2   Logic Side

The logical side which is expressed by directed (arrow) line describes the logic relationship between the two nodes. The starting node of the logic side is called former node, and the end node is called the latter node.
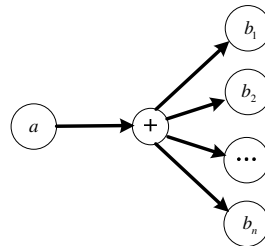
2.3   Logic Relation Node

The logic relation node describes concurrency and conflict relations of the related logical sides, including "AND" logic relation node and "OR" logic relation node [2].

1) "AND" logic relation node

It describes the concurrent relationship between the logical sides, including the "LEFT AND" logic form and "RIGHT AND" logic form and being expressed by "$\oplus$". "LEFT AND" logic describes that a number of former events ($a_1, a_2, \ldots, a_m$) co-occur and trigger the occurrence of a latter event b, whose graph cell expression is shown in Fig.1. "RIGHT AND" logic describes that a former event occurs and triggers the co-occurrence of a number of latter events ($b_1, b_2, \ldots, b_n$), whose graph cell expression is shown in Fig.2.
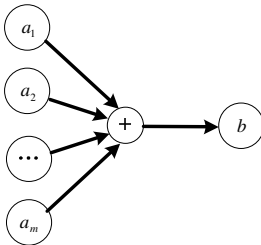
Fig. 1. "LEFT AND" logic   **Fig. 2.** "RIGHT AND" logic

2) "OR" logic relation node

It describes the conflict relationship between the logical sides, including the "LEFT OR" logic form and "RIGHT OR" logic form and being expressed by "$\odot$". "LEFT OR" logic describes that at least one of former events ($a_1, a_2, \ldots, a_m$) occurs and trigger the occurrence of a latter event b, whose graph cell expression is shown in Fig.3. "RIGHT OR" logic describes that a former event occurs and triggers the occurrence of at least one latter events ($b_1, b_2, \ldots, b_n$), whose graph cell expression is shown in Fig.4.
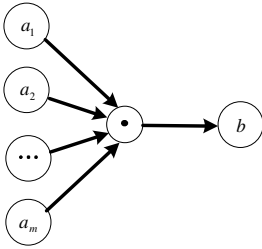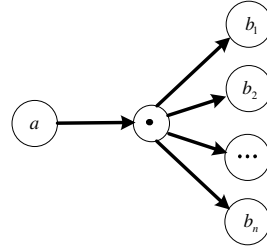
Fig. 3. "LEFT OR" logic

Fig. 4. "RIGHT OR" logic

3   Establishment Method of Event Flow Models

Event flow model can be not only used to support the establishment of experiential concept model of combat process, but also can describe the combat simulation process.

Conceptual combat process relies on operational theory, operational scenario or expert's experience, and is described by the way of manual work using the corresponding graph cell of EFG. Actually, the modeling process is the explicit expression of experiential knowledge.

It is more complex to establish the behavior event process model of combat simulation. And it is demanded that events should be extracted based on a large number of simulation data and transformed into abstract graphical description automatically. Hypothesis: the simulation data can support the establishment of event flow model, namely, the simulation process data uses event as basic record unit and each event record includes the description of the former events set which trigger the event occurrence directly. Modeling process can be described as:

1) Sort the event records in the combat simulation by the occurrence time and get the time sequence of simulation events.

2) Base on the reverse time sequence, each event record and the directly triggered events are mapped as the event nodes, the trigger relationships between the related event data records are transformed into logic sides, and the "AND" logic relation nodes are added to describe the concurrency relations. This is the directly graph cell mapping of the independent events process. In the process, the sequential triggered relations and the concurrency triggered relations are certain, therefore the event model only contains the basic event nodes, logic sides and " and" relation node.

3) The Similar events of the same operational entity are merged, while the similar "AND" logical nodes are aggregated, and the necessary "OR" logical nodes are added. If the former node and the latter node of an "AND" logic are one-to-one correspondence, namely, if any former node of an "AND" logic can be always found the corresponding node in the former nodes collection of another "AND" logical, and any latter node of an "AND" logic can be always found the corresponding node in the latter nodes collection of another "AND" logical, the two "AND" logical node could be aggregated. If the one-to-one correspondence in the former node and the latter node of logic side can not be found, "OR" logic nodes should be added and the nodes with conflict relations and the logic nodes should be connected.

4   Behavior Event Flow Analysis Principle

Completed the abstract description of combat process analysis in simulation and experience, the matching degree between combat process of simulation and theory can be analyzed. If there is obvious difference between the simulation event process and the theory event process, the credibility of simulation process and results should be doubted. This method can be regarded as order logic relation contrast between the combat process in simulation and experience. If the complexity is low, the gap between the two models can be obtained directly by manual judgment. Howard, due to the combat complexity, the actual combat event process model is often very complex, and must be auto handled.

The conceptual combat events logic models often only describe the main operational behavioral events and their order logic relations, and ignore some details. Whereas the event flow models established based on the combat simulation data is a detailed record of the behavioral events and their logical process. Therefore, there are differences between the two models in the description granularity, and it is difficult to contrast each other. The paper puts forward a treatment strategy based on the thought of "simplification - mergence – contrast". Its basic idea is as follows: First the EFG models of combat simulation processes are simplified, and the event nodes in the theoretical event process models are retained. And then according to the order logic relations between the event nodes, the simplified simulation event process models are merged into the theoretical event process model. The overlap section can be regarded as cognitive consonance, and the other section is hidden credibility peril.

Because of the structure complexity of event process models, it is difficult to merge and contrast the two models directly. If the complex event process models could be decomposed into a number of small simple logic sub-graph models, and these sub-graph models could be iteratively merged, the complexity of the merge process would be greatly simplified. Sub-graph models mergence is actually a kind of assemblage form with "OR" logic, namely, after mergence of the similar nodes, the different nodes constitute conflict event logic with "OR" logical nodes. The paper puts forward a redundancy method, that event process sub-graph models are merged by pre-adding "OR" logical nodes. The basic principle is: "OR" logical nodes are pre-added in all the logic sides of the sub-graphs. All the event nodes and logical nodes are not connected directly, but connects by "OR" logical nodes. Therefore, the mergence process of events nodes and logical nodes triggers the mergence of "OR" logical nodes and avoids the complexity in the "OR" logical node adding.
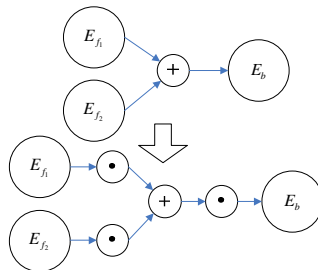

Fig. 5. Sketch map of "OR" logical node adding

The mergence algorithm is as follows [3]:

Step1: the event process model is decomposed into two basic sub-graphs. The first one is a simple logic sub-graph constituted by two event nodes and a logical side; and the second one is a logic sub-graph constituted by a logical node and its associated event node, logic sub-graph.

Step 2: All the sub-graphs are transformed into redundancy form including "OR" logic, namely, "OR" logical nodes are added between any two associated nodes (including the event node and logical node), and the logic connections are modified.

Step 3: The similar event nodes in the two sub-graphs are aggregated. The "AND" and "OR" logical nodes corresponding with the former nodes and latter nodes one-to-one are merged according to the order. It demands that the former nodes of the logical nodes should have been completed before the logic nodes merged.

Step4: Traverse all the sub-graphs to complete the mergence of the two event process models.

Step5: There are a lot of logic redundancy in the EFG model after mergence, especially the "OR" logical nodes. So the redundant nodes and the logical sides should be reduced.

5   Examples

The paper illustrates the principle of behavior event flow analysis with tank attack process. Fig.6 shows the simulation event flow model, and Fig.7 shows the experiential model by the theoretical analysis. Merge the experiential model into the simulation model, and it is found that the simulation model does not cover the process of "no hitting target event—>aiming event" in the experiential model. It shows that the simulation model did not consider the repetitive firing at the targets which were not hit. So the credibility of the simulation model may be dubitable.
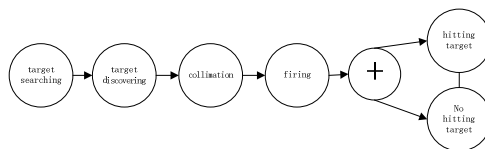

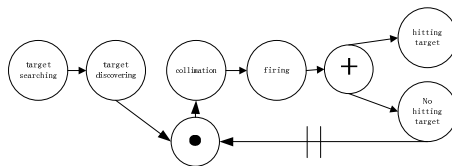**Fig. 6.** Example of operational event flow model by simulation

Fig. 7. Example of operational event flow model by theoretical analysis

6   Conclusion

The credibility evaluation based on EFG model is a kind of matching analysis of the simulation combat process and theoretical combat process. To a certain extent, it reflects the rationality of the model process. However, manual analysis of the operational event flow demands the assayer should have abundant theoretical cognition about combat process. It is especially difficult to analyze completely for the huge event flow path set, and the experiential event flow model can not fully cover all the possible combat process. In addition, the method of credibility evaluation can verify the direct logic relationship, but it can not validate indirect logic relationships. For example, the process of "event1 -> event2 ->Event 3" can be verified, but the process of "Event 1 -> Event3" can not be verified. Such indirect logic relationship validation should analysis by manual work.

References

1. Fan, X.-H.: Causality Diagram Theory Research and Fault Diagnosis Research Applying It to the Complexity System. Doctoral Dissertation of Chongqing University (2002)
2. Buss, A.H.: Modeling with Event Graphs. In: Proceedings of the 1996 Winter Simulation Conference (1996)
3. Shi, F.: Extended Event Graph Based Causal Tracing Analysis of Complex Simulation System. Doctoral Dissertation of National University of Defense Technology (2006)

A Simulation Model Validation Method Based on Design of Experiments[*]

Dezhi Dong, Jiangyun Wang, and Ping Zhang

School of Automation Science and Electrical Engineering, Beihang University,
100191 Beijing, China
gavintng@gmail.com

Abstract. When simulation factors are numerous while real-world observed data are sparse, the issue of validating the simulation models is problematic. An extreme case is focused that limited real-world observations are available cross the factor space, and only a single replicate is available on per simulation factor setting. A method based on design of experiments is proposed by which the validation experiments could be well arranged across the factor space through optimal design. The p-value test technology is employed to evaluate the statistical consistency of the static data, and for a set of validation experiments obtained by DoE, the combined analysis of all the p-values resulted from these experiments can be taken based on the inverse-CDF theorem, to make an overall characterization of degree of the simulation credibility on the entire factor space. An example of validation of a guided missile simulation is taken to demonstrate that the method is useful.

Keywords: simulation models, model validation, VV&A, design of experiments, p-values.

1 Introduction

With great development of the modeling and simulation (M&S), the M&S plays an ever important role in the design of systems and making decisions about them. The central issue of credible use of the M&S is simulation model validation, which is the process of determining if the M&S accurately represents the real-world from the perspectives of its intended use [1]. The degree to which M&S results are statistically consistent with observed data should be investigated in this process.

Through various combinations of the simulation factors, corresponding operation conditions are simulated upon which the real-world systems are conditioned. If the system is complex, the corresponding simulation factor space is really huge. To characterize the validity of the simulation, we need to validate the simulation model on every corresponding factor setting within the factor space, but this brings trivial and meaningless workload to us. Besides, it is extraordinarily expensive to obtain the observed data of the real-world systems, and generally no replicates are available on a

---

single operation condition. The classical method is often based on the assumption that under a typical operation condition, a real-world data sample is available to achieve a data comparison [2][3]. But the problems are, generally only a few of factor settings are considered, so the validity on whole factor space can't be characterized; the "real-data sample" assumption is too optimistic to be satisfied in practice.

In this paper we focus on an extreme case that limited real-world observations of the system output are available cross the factor space, and only a single replicate is available on per simulation factor setting. A novel validation method based on design of experiments (DoE) is proposed. The optimal DoE technology is introduced to screen the factor space and obtain representative validation experiments, according to the intended use of the simulation. To overcome the problem that the observation data are scarce, the data analysis based on p-value method is studied and a combined probability analysis is introduced to offer a credibility evaluation on the factor space with fewer real-world data needed. An example of validation of a guided missile simulation is taken to demonstrate that the method is useful.

2   Simulation Factors and the Simulation/Observation Pair

Simulation generally involves the issue of huge number of inputs and parameters. These inputs and parameters can be divided into two categories, and the one whose values are deterministic is called certainty factors or simulation factors, while the other whose values are stochastic is called uncertainty factors [4].

According to our intended use, we will explore the real-world system on a set of operation conditions which is called operation space of the system. Different combinations of the values of simulation factors correspond to the different operation conditions of the system, which are called simulation factor settings. What is called simulation factor space is a set of simulation factor settings upon which the model is conditioned to simulate the system conditioned upon its operation space.

In this analysis, a simulation /observation pair will involve a vector of realizations x_{sim}, of the simulation output random variable, X_{sim}, and a single realization x_{obs}, of the real-system output random variable, X_{obs}. For I factor settings interested, I simulation/observation pairs can be obtained.

$$X_{sim} \rightarrow x_{sim}^{i} = (x_{sim,1}^{i}, x_{sim,2}^{i}, ..., x_{sim,N}^{i}) \; for \; i = 1,...,I$$
$$X_{obs} \rightarrow x_{obs}^{i} \; for \; i = 1,...,I$$

(1)

Where I is the number of pairs in the series of experiments and N is the number of Monte Carlo trials during each validation experiment.

3   Design of Validation Experiments

During simulation validation process, if the number of the simulation factors is huge, the number of corresponding factor settings will be enormous. By classical methods,

validation on every factor setting may be a job impossible. We will introduce DoE technology to get a representative subset of the factor space to test. A countable subset of the simulation factor space, S_v^k, is a validation experiment set, if it satisfies: 1) the factor settings in S_v^k are filled throughout the factor space, to guarantee the coverage of the factor space; 2) the factors in S_v^k are not correlated with each other to avoid meaningless validation experiments. Validation experiment design is to find a validation experiment set S_v^k from the simulation factor space.

Latin hypercube design (LHD) is a representative method of the "space filling" design methods, and it can be used in our case. The LHD of k factors and n experiments can be denoted as a $n \times k$ dimensional matrix $D^{n \times k} = LHD(n,k)$. LHD tries to fill the design points spread in space, but this may weaken the orthogonality of the design, in addition, the LHD may be unstable due to the random generating of it. For these reasons, we will introduce the optimal LHD to address the problem of design of validation experiments.

3.1   Optimal Design Criteria of LHD

The minimum pairwise correlation between the factors and the maximum distance between the design points are both good criteria for finding optimal LHD. A criterion to minimize correlation between the factors was proposed by Owen and Tang [5], which is defined as follows.

$$\rho^2 = \frac{\sum_{i=2}^{k} \sum_{j}^{i-1} \rho_{ij}^2}{k(k-1)/2} \tag{2}$$

Where, ρ_{ij} is the pairwise correlation between the columns i and j of the design, and $\rho_{ij} \in [0,1]$. The smaller ρ^2 is, the weaker the pairwise correlation is.

Johnson, Moore and Ylvisaker proposed a "maximin" design criterion, according to this, the minimal distance between the design points will be maximized to ensure the design points are most "space filling". The "maximin" criterion [6] is:

$$\phi_p = \left(\sum_{i=1}^{\binom{n}{2}} \frac{1}{d_i^p} \right)^p \tag{3}$$

Where, p is a positive integer, and d_i is a pairwise distance measurement of the design points. Now the design that minimizes ϕ_p will be a "maximin" design. Note a fact that ϕ_p may be larger than 1. Based on the criterion to minimize the pairwise correlation and the criterion to maximize the distance between the design points given above, the criteria of optimal LHD can be given as a multi-objective criterion.

$$\psi_p = \omega\rho^2 + (1-\omega)\frac{\phi_p - \phi_{p,L}}{\phi_{p,U} - \phi_{p,L}} \tag{4}$$

Where $\phi_{p,L}$ and $\phi_{p,U}$ are the lower and upper bounds of ϕ_p employed to scale ψ_p to belong to the interval $[0,1]$, and they can be calculated according to the method in [6]. ω is some predefined positive weight. By this optimal design criterion, the design will be considered to be optimal when ψ_p is minimized.

3.2   Optimal LHD of Simulation Validation Experiments

The optimal design of the simulation validation experiments can be achieved by a process of iteration to find a LHD which makes ψ_p as small as possible. We introduce a version of the simulated annealing algorithm (MMA) [7] for optimizing ψ_p, it is presented as follows.

Step 1. Randomly chose a LHD $D = LHD(n,k)$.

Step 2. Randomly pick out two elements of a randomly chosen column of D, and then exchange these two elements to obtain a new LHD D_{new}.

Step 3. Calculate the values of $\psi_p(D)$ and $\psi_p(D_{new})$, and if $\psi_p(D_{new}) < \psi_p(D)$, $D = D_{new}$; else if $\eta < \pi = \exp\{-[\phi_p(D_{new}) - \phi_p(D)]/t\}$, $D = D_{new}$; else maintain D unchanged.

Step 4. If the optimal criteria are met, end the iteration; else go to Step 2.

In the algorithm above, t is a preset parameter known as "temperature" [7]. By design of the validation experiments using the optimal LHD, the validation experiments can be well arranged within the factor space.

4   Data Consistency of Simulation/Observation Pair

On each factor setting in validation experiment set, the simulation/observation pair can be obtained from experiment and real-world observation. The consistency of the pair needs to be analyzed to provide evidence for credibility of the simulation.

4.1   Validation Experiment Process and Experiment Sample Acquisition

For experiment on the i-th factor setting, the uncertainty factors will be randomly drawn from its distribution to get a simulation input sample. Using this input sample to execute the Monte Carlo trials of the simulation model repeatedly, ones can obtain an experiment sample x_{sim}^i of the simulation output. With an observation of the real-world system output conditioned upon the corresponding operation condition, a simulation/observation pair (x_{sim}^i, x_{obs}^i) will be collected. Generally speaking, the

distribution of uncertainty factors can be assumed according to some priori knowledge, historical data and real-world observations.

4.2   Analysis of Consistency of Simulation/Observation Pair

For each single factor setting in the experiment set, we can obtain a simulation/observation pair (x^i_{sim}, x^i_{obs}) from the validation experiments. The problem of consistency analysis can be characterized as a hypothetic test problem.

$$H_0 : x^i_{obs} \sim F^{(i)}_{sim}\Big|_{i=1,2,\cdots,I} \tag{5}$$

Where $F^{(i)}_{sim}$ is the distribution of the simulation output conditioned upon the i-th factor setting estimated from x^i_{sim}. The p-value analysis technology is introduced here to solve the problem of (5). To the simulation/observation pair of the i-th factor setting, consistency analysis is to decide whether the p-value the real-world observation x^i_{obs} relating to is extreme or not, assuming H_0 in (5) is true. The definition of p-values here is given below.

$$p_i = P\{|x - \bar{\mu}_{sim}| > |x^i_{obs} - \bar{\mu}_{sim}| \, \big| \, x^i_{obs} \sim F^{(i)}_{sim}\} \tag{6}$$

Where p_i is the p-value of x^i_{obs}, $P\{\bullet\}$ is the calculation of probability, and $\bar{\mu}_{sim}$ is the estimate of the mean value of X_{sim} in the i-th experiment. Given a pre-specified significance level α, If the p-value calculated by (6) is larger than α, we can consider that the observation of x^i_{obs} is optimistic for accepting the null hypothesis, and the statistical consistency of the simulation/observation pair is satisfactory. Else, we will be apt to reject H_0.

5   Combined Analysis of P-Values

The combined analysis here is to combine the consistency analysis results of all the simulation/observation pairs and offer an overall characterization of the simulation model upon the factor space.

Inverse-CDF theorem [8] implies that if the data in a sample is truly distributed following a same population, for large sample sizes, the cumulative probabilities of the data in this sample will be distributed uniformly between 0 and 1.

Assume for i-th validation experiment, the null hypothesis H_0 in (5) is true, and then by inverse-CDF theorem, the cumulative probabilities $F^{(i)}_{sim}(x^i_{obs})$ can be considered as a random draw from a uniform distribution on $[0,1]$. By analyzing for I experiments in validation, even if each experiment represents a different population pair, the cumulative probabilities can be analyzed in aggregate [9]. Note that the

uniform distributions are completely described by their bounds, so $F_{sim}^{(i)}(x_{obs}^i)$ from all experiments will constitute a random sample from a uniform distribution between 0 and 1, when I is large enough.

But, note that the analysis products of the validation experiments are p-values of x_{obs}^i, rather than $F_{sim}^{(i)}(x_{obs}^i)$, so we need to derive the relation between $F_{sim}^{(i)}(x_{obs}^i)$ and the p-values. By Equation (6), this relation can be derived as:

$$p_i = \begin{cases} 2(1 - F_{sim}^{(i)}(x_{obs}^i)), & x_{obs}^i > \bar{\mu}_{sim} \\ 2F_{sim}^{(i)}(x_{obs}^i), & x_{obs}^i \le \bar{\mu}_{sim} \end{cases} \tag{7}$$

Equation (7) implies the p-values are distributed as same as the cumulative probabilities, whatever x_{obs}^i is, so here it comes that the problem of characterization of simulation model validity on I factor settings in the validation experiment set has been transformed to the problem of evaluation of whether I p-values follow the uniform distribution bounded with 0 and 1. We will have enough confidence to declare that the simulation is credible on the experiment set if the p-values are uniformly distributed between 0 and 1. Further, because the validation experiment set is obtained by an optimal design of the simulation factor space, the conclusion that the simulation is credible upon whole factor space, can be made.

The test of the distribution uniformity of the p-values can be undertaken by the K-S test [10], etc., and we will not say more than is needed about this problem here.

6   Case Study

6.1   Example

A simulation model of a guidance missile was validated using the DoE based method. We pay close attention to the issue of characterization of the credibility of the missile miss distance simulation conditioned upon different missile/target relative motions. Assume the target is moving uniformly and rectilinearly, then we can take the initial angle of pitch (θ_{tgt}) and the initial translational speed (V_{tgt}) as the simulation factors.

Assume the simulation factor space can be given as a *2*-dimensional factor space:

$$D_M^2 = \{(\theta_{tgt}, V_{tgt}) | \theta_{tgt} \in [0.83\pi rad, 1.17\pi rad], V_{tgt} \in [280m/s, 350m/s]\} \tag{8}$$

We analyzed the effects of the drag error (ε_x) on miss distance here which was taken as an uncertainty factor to be analyzed. Generally the drag error can be characterized using normal distribution $N(0, (0.05q_x)^2)$ [11].

In design of the validation experiments, the size of the simulation validation experiment set was selected as $I=30$, and we employed the optimal LHD algorithm stated in Section 3.2, the factor settings obtained are presented in Fig. 1.
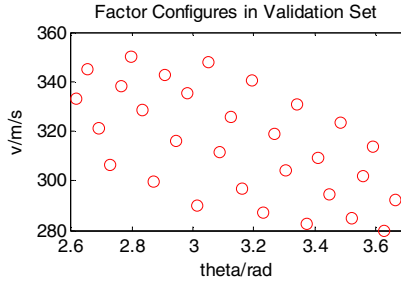
Fig. 1. The design of validation experiment is shown. 30 factor settings are presented as circles.

Assume to these 30 simulation factor settings obtained, we could obtain the observations of the real-world miss distance upon corresponding operation conditions, $x_{obs}^i, i = 1, \cdots, 30$. In each validation experiment, randomly draw 100 values of ε_x from its distribution, and run the Monte Carlo trials. After 30 validation experiments, totally 30 experiment samples could be obtained. We obtained an empirical CDF per sample, to analyze further the relative relation between the simulation/observation pair. We choose 4 of 30 samples and give their empirical CDFs here in Fig. 2.
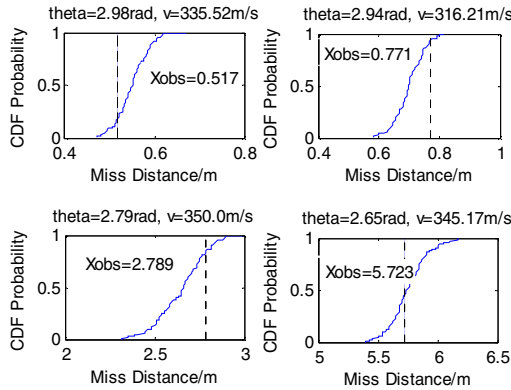
Fig. 2. The empirical distribution of 4 of 30 samples is presented. We use vertical lines to present the quantitative relation between the simulation/observation pair.

We calculated the p-values the 30 real-world observations relating to using the empirical CDF estimated from the experiment samples, and list them in Table 1.

Table 1. The p-values resulted from the experiments are listed, and each p-value denotes the degree of the real-world observation is deviated from the simulation output distribution

Index i	p-values
1 -10	0.705, 0.137, 0.311, 0.131, 0.822, 0.135, 0.177, 0.824, 0.775, 0.444
11-20	0.912, 0.651, 0.972, 0.794, 0.371, 0.513, 0.430, 0.203, 0.363, 0.731
21-30	0.734, 0.198, 0.470, 0.328, 0.960, 0.625, 0.429, 0.215, 0.648, 0.547

Given a significance level of 5%, almost all p-values are falling within the interval $[0.05,1]$, that is to say, there is no experiment in which the data consistency of the simulation/observation pair is too "bad" to make us rejecting the H_0 on the corresponding factor setting in our validation experiment set.

To achieve the combined analysis of p-values over the experiment set, we made the empirical CDF plot of these 30 p-values, and compared it to the CDF of the standard uniform distribution from 0 to 1, as the Fig. 3 (a) shows.

(a)   (b)
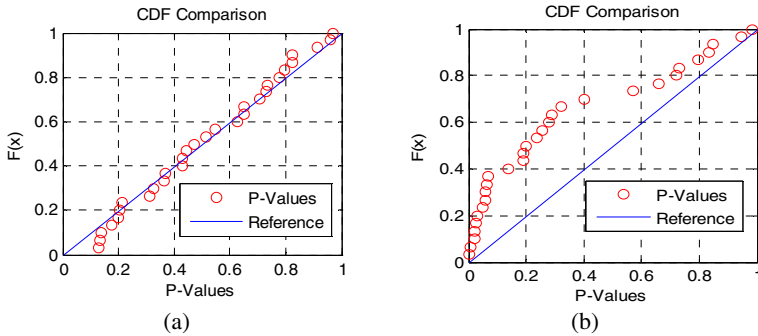
Fig. 3. A "probability paper test" is present, in which the closer the empirical CDF of the sample fit to the straight line, the closer the sample distributed follows to the reference distribution, and (a) is the test for the "real" case and (b) is the test for the "artificial" case.

From Fig. 3 (a) we can find that the empirical CDF of the p-values are close to the straight line, so we can consider that the sample of p-values uniformly distributed on $[0,1]$. For strict precaution, we employ the Kolmogorov-Smirnov method [10] to test whether the conclusion obtained from Fig. 3 (a) is right. The K-S statistic calculated is 0.1241, and the corresponding reject probability is 0.28, which is more lager than 0.05 (the pre-specified significance level), so we will have enough confidence to conclude that the p-values are uniformly distributed on $[0,1]$. Here, we can obtain a conclusion that the guidance missile simulation is credible on factor space D_M^2, or say, to the operation space of the missile, the simulation is credible enough.

6.2   Performance Analysis

To offer the reliability of the combined analysis and analyze the performance of the method, for each factor setting obtained through DoE, we set an artificial "real-world observation" which doesn't follow the distribution of the experiment sample. The p-values are calculated and tested of the uniformity and the Fig. 3 (b) is obtained, in which the empirical CDF of these p-values deviates from the one of the uniform distribution. We can conclude that if the simulation/observation pair is not consistent, the method we proposed can lead to a conclusion of reject of the credibility of the simulation model. This further proves that our method is reliable.

7   Conclusions

Considering the problems in validation of simulation that the simulation factor space is huge and the real-world observation data is hardly available, a novel method based on DoE is proposed. The case study in Section 6 demonstrates this method is useful to solve the problem of simulation model validation on a given factor space, and the better reliability and operability of the method is revealed. Some points for attention will be recommended as follows.

Firstly, a case may emerge in which some extreme p-values are obtained in validation, but the simulation model is judged to be credible after combined analysis. This means even the model is not "good" on a few factor settings, but it does not affect the validity of the model on whole factor space.

Secondly, the issue of statistical consistency of the dynamic data is not considered in this article, and will be further studied in future work.

Finally, our method is more reliable and operable, and is suited for validation of the huge, complex and stochastic system simulations.

References

 1. Law, A.M., Kelton, W.D.: Simulation Modeling and Analysis. Journal of Economic Entomology 82, 1161–1168 (1991)
 2. Liao, Y., Liang, J., Yang, X., Deng, F.: Research on Validation Methods of Six-Degree-of-Freedom Simulation Model of Missile Weapon System. Aerospace Shanghai 6, 34–38 (2007)
 3. Gafarian, A.V., Walsh, J.E.: Statistical Method for Validating Simulation Models by Comparison with Operational Systems. In: Proceeding of the 4th International Conference on Operations Research, New York (1969)
 4. Seila, A.F.: Output Analysis for Simulation. In: Proceedings of the 1990 Winter Simulation Conference, Detroit, Michigan (1990)
 5. Tang, B.: Selecting Latin Hypercubes Using Correlation Criteria. Statistica Sinica 8, 965–978 (1998)
 6. Jin, R., Chen, W., Sudjianto, A.: An Efficient Algorithm for Constructing Optimal Design of Computer Experiments. J. Statist. Plann. Inference 134, 268–287 (2005)
 7. Li, X., Zhang, H.: Improvement of Simulated Annealing Algorithm. Software Guide 7, 47–48 (2008)
 8. Leon-Garcia, A.: Probability and Random Processes for Electrical Engineering, 2nd edn. Addison Wesley Longman (1994)
 9. Smith, B.D.: Simulation Validation Using a Non-Parametric Statistical Method. In: 2006 ITEA Modeling and Simulation Conference, Las Cruces, NM (2006)
10. Massey, F.J.: The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association 46, 68–78 (1951)
11. Peng, J.: Research on the VV&A Theory and Method of Missile Guidance Simulation System. A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Engineering in Control Science and Engineering, Graduate School of National University of Defense Technology, Changsha, Hunan (2010)

A Consistency Test Method for Simulation Data Considering Shape and Distance of Series

Yuwei Hu, Ping Ma, Ming Yang[*], and Zicai Wang

Control and Simulation Center, Harbin Institute of Technology, Harbin, China
2006huyw@163.com, {pingma,myang,wzc}@hit.edu.cn

Abstract. By utilizing contrasting results between simulation data and reference data, dynamic consistency test of simulation results is an important content of model validation. In this paper, a novel consistency test method considering shape and distance of series is proposed. Based on the analysis of dynamic consistency test problem, an improved grey relational grade model is established to assess the dynamic consistency among series, which considered the shape and distance of series as two main attributes. The consistency is tested in view of similarity and nearness between series to accomplish overall test task. Meanwhile, the method is effective to test the consistency for the two sets of simulation series being symmetric of reference data. Finally, an example of consistency test for launch current of electromagnetic railgun is given to illustrate that the improved method is effective and practical for assessment of dynamic consistency of simulation data.

Keywords: Dynamic Consistency Test, Grey Relational Grade, Model Validation, Electromagnetic Railgun.

1 Introduction

Model validation, as an important part of modeling and simulation, is a necessary way to ensure the simulation credibility [1]. The credible results contribute to increasing confidence in using model to analyze and predict the actual system performance, which will reduce cost and improve security of experiment. Dynamic consistency test for simulation data is one of the most directive approaches in validation methodology [2]. Based on such idea the credibility of model can be assessed by consistency test data.

During the past several decades, many dynamic consistency test method were proposed for model validation problem, which are classified into two kinds method including time domain validation and frequency domain validation [1],[3]. In this paper, we consider with validation method in time domain where the sample data is directly used to validate model. Majority of the methods are effective under the condition that sample data obey some distribution characteristic. But, grey theory is a method for researching few data, poor information and uncertainty [4]. From this on,

---

[*] Corresponding author.

grey relation analysis was introduced into solving the problem of model validation [5], [6]. Moreover, Wu [7] extended grey relational analysis method by integrating the acceleration and velocity of samples into deviation. However, shape and distance of series, considered as two important attributes of series similarity, are introduced to the construction of grey relational grade which can assess the dynamic consistency of series. The proposed method taking into consideration of the effects of the two main features on the deviation between series gives comprehensive consistency test for simulation results. Meanwhile, sensible conclusion can be drawn for the two sets of simulation series being symmetric of reference data.

This paper is organized as follows: the dynamic consistency test problem is analyzed in section 2. Considering shape and distance of series, a new consistency test method for simulation data is proposed and proven in section 3. Section 4 gives an example to illustrate the proposed approach. Finally, section 5 presents our conclusion.

2   Analysis of Dynamic Consistency Test Problem

Comparison of simulation and test data for weapon system is a direct way to validate the simulation model of the actual system. According to the output characteristic, the performance of the system can be classified into static performance and dynamic performance. The static performance is considered as random vector at distribution. Dynamic performance refers to multi-dimensional random process and can fully reflect whole process of the system. So, consistency test for the dynamic performance is very necessary to validate the model of system. Mathematics statistics is primary method for dynamic consistency test, which makes statistics analysis to simulation results in every simulation experiment and determine the consistency between simulation and reference data under a certain significant level. So, statistics method is dependent on the number and distribution of dynamic performance. However, for series containing few data and poor information, grey relational analysis is applicable to test the consistency. Grey relational analysis is a basic principle of the system by judging the relational degree of the development situation during grey process based on the geometrical figure similar degree of the series curves. The grey relational grade is real number between 0 and 1. The number 1 indicates the high similarity with reference series. But, the method only emphasizes on the similarity of shape between simulation and reference series and it is not completely suitable for dynamic consistency test.

For example, Fig. 1 illustrates the comparison between simulation and reference series. Using grey relational analysis method, the grey relational grade of Fig. 1(a) is 1, which is larger than the calculation results of Fig. 1(b). The analysis results show that simulation series A has good consistency with reference series. But, by direct observation, simulation series B is much closer to reference series. So, distance between series should be considered for dynamic consistency test.
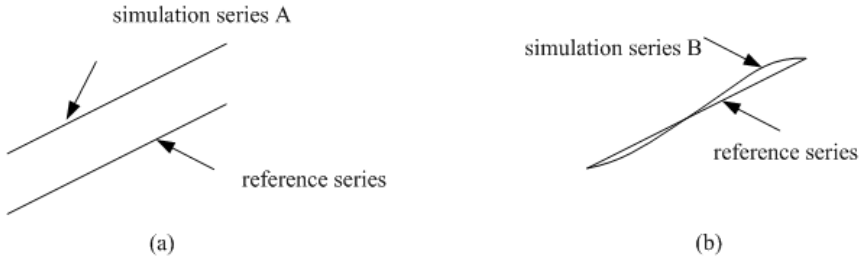
Fig. 1. Comparison between simulation and reference series

3　Consistency Test Method for Simulation Data

Through the analysis of the dynamic consistency test problem, the distance and shape of series are the two main attributes to determine the similarity degree of series. If the distance between two series is large, the degree of nearness can not compensate for the similarity degree of series. It indicates that the linear compensation of the two attributes is not obtained. The improved grey relational grade model for consistency test between simulation and reference series is established by multiplicative synthesis.

3.1　Constructing the Model for Dynamic Consistency Test

There is no requirement for sampling size and distribution, dynamic consistency test for simulation data can be performed by making direct use of the series. Suppose there are two series $X_0(t)$ and $X_i(t)$: $X_0 = \{X_0(1), X_0(2), \ldots, X_0(r)\}$ is reference series and $X_i = \{X_i(1), X_i(2), \ldots, X_i(r)\}$ is compared series. Absolute deviation of the ith interval numbers between two series is defined as $\Delta_{0i}(k) = |X_0(k) - X_i(k)|$. The traditional grey relational coefficient is expressed as

$$\zeta_i(t) = \frac{\min_{i,t} |X_0(t) - X_i(t)| + \rho \max_{i,t} |X_0(t) - X_i(t)|}{|X_0(t) - X_i(t)| + \rho \max_{i,t} |X_0(t) - X_i(t)|} \tag{1}$$

In equation (1), ρ is the distinguishing coefficient ranging from 0 to 0.5, which is set 0.5 in the following analysis.

Definition 1. The nearness between two series is relevant to relative error, which is defined as follows

$$M(k) = \frac{\Delta_{0i}(k)}{T(k)} + \frac{\Delta_{0i}(k)}{T(k) + \Delta_{0i}(k)} \tag{2}$$

Definition 2. $T(k)$ is relevant to the value of every data point of series. For two compared series being symmetric of reference series, the similarity relative to the reference series is identical. Then, $T(k)$ is determined using the following expression

$$T(k) = \begin{cases} \sqrt{|X_0(k)|}^{(1+\mathrm{sgn}(X_i(k)-X_0(k)))} \sqrt{|X_i(k)|}^{(1-\mathrm{sgn}(X_i(k)-X_0(k)))} & X_i(k) \geq X_0(k) \\ \sqrt{|X_0(k)|}^{(1+\mathrm{sgn}(X_0(k)-X_i(k)))} \sqrt{|X_i(k)|}^{(1-\mathrm{sgn}(X_0(k)-X_i(k)))} & X_i(k) < X_0(k) \end{cases} \tag{3}$$

Definition 3. Considering the distance and shape as two main attributes of series similarity, we can obtain the improved grey relational coefficient, which is defined as the following formula

$$\zeta_i'(k) = \left(\zeta_i(k) \times \exp(-M(k)) \right)^{1/2} \tag{4}$$

Get the grey relational trade between the compare sequence and reference sequence, defined as follows

$$\gamma_i(X_0, X_i) = \frac{1}{r} \sum_{i=1}^{r} \zeta_i(k) \tag{5}$$

The dynamic consistency for the compared series can be test by the value of the improved grey relational grade $\gamma_i(X_0, X_i)$. The bigger the grey relational grade is, and the more similar the compared series is.

3.2   Main Properties of the Improved Grey Relational Grade

The improved grey relational grade is satisfied with condition of distance measure axiom and it has four main properties including norm property, wholeness property, duality symmetric properties and approachability property.

Proof. (1) Norm property
Since $\Delta_{0i}(k) \geq 0$, we have that $0 < \rho'(X_0, X_i) \leq 1$ and $M(k) \geq 0$. Then, $0 < \rho(X_0, X_i) \leq 1$. When $X_0(k) = X_i(k)$ and $\Delta_{0i}(k) = 0$, then, $\rho(X_0, X_i) = 1$.
Therefore, norm property holds true.
(2) Wholeness property
If $X = \{X_i | i = 0,1,...,n; n \geq 2\}$, for any $X_{i1}, X_{i2} \in X$, generally speaking, $\max_{i1} \max_{r} |\Delta_{0i}(k)| \neq \max_{i2} \max_{r} |\Delta_{0i}(k)|$, then $\rho'(X_0, X_{i1}) = \rho'(X_0, X_{i2})$.

When $\Delta_{0i1}(k) \neq \Delta_{0i2}(k)$, then $\gamma_i(X_0, X_{i1}) \neq \gamma_i(X_0, X_{i2})$. So wholeness property is true.
(3) Duality symmetric property
For any $X_{i1}, X_{i2} \in X$, then, $\Delta_{i1i2}(k) = \Delta_{i2i1}(k)$ and $\rho'(X_{i1}, X_{i2}) = \rho'(X_{i2}, X_{i1})$.

If $X_{i2} > X_{i1}$, then

$$T_{i1i2}(k) = \sqrt{|X_{i1}(k)|}^{(1+\mathrm{sgn}(X_{i2}(k)-X_{i1}(k)))} \sqrt{|X_{i2}(k)|}^{(1-\mathrm{sgn}(X_{i2}(k)-X_{i1}(k)))} = |X_{i1}(k)|$$

$$T_{i2i1}(k) = \sqrt{|X_{i2}(k)|}^{(1+\mathrm{sgn}(X_{i1}(k)-X_{i2}(k)))} \sqrt{|X_{i1}(k)|}^{(1-\mathrm{sgn}(X_{i1}(k)-X_{i2}(k)))} = |X_{i1}(k)| = T_{i1i2}(k).$$

Obviously, for any $i1, i2$, $\gamma_i(X_{i1}, X_{i2}) = \gamma_i(X_{i2}, X_{i1})$, therefore, duality symmetric property is true.

(4) Approachability property

For $\left| X_0(k) - X_i(k) \right| \neq Const$, if $\Delta_{0i}(k) \to 0$, then $\exp(-M(k)) \to 1$ .And if $\Delta_{0i}(k) \to \infty$, $\exp(-M(k)) \to 0$.

For $\left| X_0(k) - X_i(k) \right| = Const$, then $M(k) = \dfrac{Const}{X_0(k)} + \dfrac{Const}{X_0(k) + Const}$.

If $Const \to 0$, then $\exp(-M(k)) \to 1$. And, if $Const \to \infty$, then $\exp(-M(k)) \to 0$. So, approachability property is realizable.

4   An Illustration Example

Consistency test method for simulation data has been applied to the validation of electromagnetic railgun simulation model. The electromagnetic railgun is a new weapon that utilizes electricity stored in power rather than chemical propellants to launch projectiles at long-range target. When operational, the railgun can provide high precision and time-critical fires in all-weather conditions by controlling the amplitude and shape of current through the rails. So, the launch current is a key index affecting the performance of electromagnetic railgun.
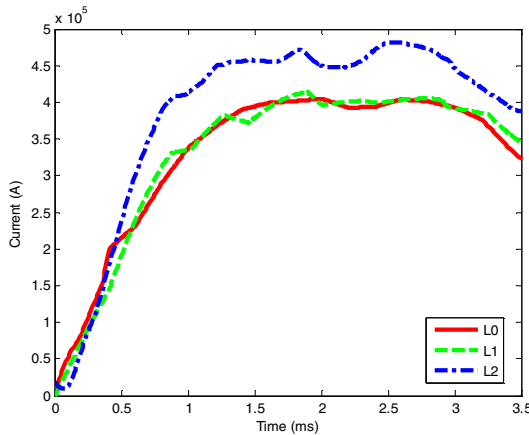

Fig. 2. Launch current profile in electromagnetic railgun system and two simulation models

Fig. 2 shows profile of current which is in a railgun system and two simulation models. So, there are three series curves to be analyzed. They are L_0 , L_1 , L_2 respectively. The series L_0 is output of the real system and the series L_1 , L_2 are the output of two different simulation model. To illustrate the effect of dynamic consistency test for simulation results, Table 1. list test results which are calculated based on three different methods including the proposed method, grey relational analysis and Tic method.

Table 1. Dynamic consistency test results for the whole launch current of railgun

Test method	Series L_1	Series L_2
Proposed method	0.825	0.572
Grey relational analysis	0.763	0.493
Tic method	0.019	0.084

Since the smaller the Tic coefficient is, and the more similar the compared series is, the results through proposed method are identical with the computation results based on grey relational analysis and Tic method from Table. 1. The series L_1 is closer to the reference series than the series L_2. The proposed method is effective to validate the simulation model.
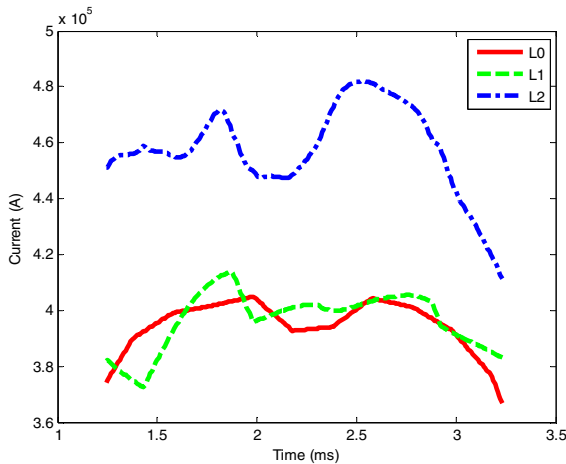
Fig. 3. Launch current profile in railgun system and two simulation models during flat-top stage

During launch, the current through the rails is expected to keep steady for the consideration of reducing the damage effects of the projectile on the bore. Dynamic consistency test for the current, shown in Fig.2, during flat-top stage is performed by the three methods and the test results are listed in Table 2.

Table 2. Dynamic consistency test results for the current during flat-top stage

Test method	Series L_1	Series L_2
Proposed method	0.806	0.782
Grey relational analysis	0.679	0.819
Tic method	0.009	0.074

By comparing the calculation results based on the three methods, we can discover that the results through the proposed method are identical with the Tic method results. It means that the series L_1 is closer to the reference series than the series L_2. The conclusion corresponds to the observation results. However, the conclusion from grey relational analysis method is not identical with the conclusion based on Tic method. So, the advantage of the proposed method is shown by testing the dynamic consistency for the launch current during flat stage.

5   Conclusions

Consistency test for simulation data is an important part of validation of simulation model. The results of consistency test have a major impact on credibility of the simulation model. In this paper, a novel consistency test method considering shape and distance of series curve is proposed to validate the model. Considered from similarity and nearness between reference series and compared reference, an improved grey relational grade model is constructed to fully assess the dynamic consistency among series curves. The effectiveness and feasibility of the proposed method is verified through the consistency test for the discharging current in a railgun launch process.

References

1. Sargent, R.G.: Verification Validation and Accreditation of Simulation models. In: Winter Simulation Conference 2000, pp. 50–59. IEEE Press, New York (2000)
2. Balci, O.: Principles and Techniques of Simulation Validation Verification and Testing. In: IEEE Simulation Conference Proceedings, pp. 147–154. IEEE Press, New York (1995)
3. Uselton, S., Dorn, G., Farhat, C.: Validation Verification and Evaluation. In: IEEE Conference on Visualization 1994 Proceeding, pp. 414–418. IEEE Press, New York (1994)
4. Wilsdon, T., Slay, J.: Digital Forensics: Exploring Validation Verification and Certification. In: First International Workshop on Systematic Approaches to Digital Forensic Engineering, pp. 48–55. IEEE Press, New York (2005)
5. Wei, H.L., Li, Z.W.: Grey Relational Analysis and Its Application to the Validation of Computer Simulation Models for Missile Systems. In: Systems Engineering and Electronics, Beijing, pp. 55–61 (1997)
6. Wang, S.Z., Liu, X.T., Tian, X.H.: Research on Simulation Model Evaluation Using Grey Correlation Degree. Journal of Air Force Engineering University, 73–76 (2007)
7. Wu, J., Wu, X.Y., Chen, Y.X., Teng, J.C.: Validation of Simulation Models Based on Improved Grey Relational Analysis. In: Systems Engineering and Electronics, Beijing, pp. 1677–1679 (2010)

Research on Credibility Evaluation Framework of Manned-Spaceflight Training Simulator

Jiangang Chao, Junjun Wang, and Pu Wang

China Astronaut Research and Training Center, Haidian District, Beiqing Road No. 26,
5132 Mailbox 22 Sub-mailbox, 100094, Beijing, China
{Xjtucjg,nudtcomputer}@163.com, wjunjun1979@sohu.com

Abstract. A Manned-Spaceflight Training Simulator (MSTS) was the main training platform for Chinese astronauts. To determine the effectiveness of the training, simulation credibility of the MSTS must be verified. Therefore, this paper aims to establish a simulation credibility evaluation framework for an MSTS. Through reviewing existing evaluation systems, the evaluation framework for an MSTS was fundamentally established according to the training purpose of the simulator and the astronaut's perspective. The framework contains two parts: static simulation and dynamic simulation, which can evaluate simulation credibility from visual, audio, and psychomotor sensory experience. The applicability and accuracy of the framework was basically proved by a questionnaire given to selected users of the MSTS and technical experts. Using this evaluation framework, the simulation credibility of an MSTS can be thoroughly scientifically tested.

Keywords: Simulation Credibility, Evaluation Framework, Training Simulator.

1 Introduction

Simulation credibility is the key performance factor for evaluating a simulation system, and determining how to evaluate simulation credibility is important. The first step is to establish a credibility evaluation framework. In previous research, different frameworks were established for different simulation systems and, therefore, definitions of "simulation credibility" vary. Sometimes, the definitions have no obvious differentiation from the definition of fidelity. But essentially, all definitions express these concepts:

- Simulation credibility is closely related to the simulation's purpose. In fact, research shows that this is the most important factor;
- What simulation credibility compares is not the "real" system, but the referent extracted from the real system according to the purpose of the simulation;
- Simulation credibility is used to relatively evaluate the degree that the simulation system matches the simulation's purpose, and the degree is usually expressed by the degree of resemblance to certain aspects of the simulation.

Schricher et al. (2001) established a generic model of fidelity measurement framework. By using this framework, all simulation task-critical objects can be identified as well as their associated behaviors and attributes for fidelity measurement. Prasad et al. (Rehman, 1995) listed the fidelity characteristics of a typical training simulation system. Lane and Alluisi (Rehman, 1995) identified four fidelity drivers that are used to determine simulation requirements: the mission to be simulated, objectives of the simulation, fidelity dimensions, and simulation components. Different evaluation frameworks of simulation credibility are mentioned in references 4~7.

In this paper, an MSTS credibility evaluation framework is established on the basis of the evaluation system's research results according to the characteristic of the MSTS and related standards of aviation simulators.

2   MSTS Profile

An MSTS is a spacecraft simulator built on the ground to simulate the inside environment and flight process of a spacecraft. After training in an MSTS, astronauts will be familiar with the functions and flight process of the craft and can accomplish aerospace flight tasks smoothly. An MSTS generally has such training functions as:

- Getting familiar with the cabin environment;
- Procedures training, such as standard procedures and incident response;
- Operation training on the instruments;
- Task training, such as Rendezvous and Dorking (RVD), Extravehicular Activity (EVA), etc.

During implementation, the simulation is divided into six parts according to the structure of the actual aircraft and the simulation's purpose, and all six parts are interconnected by interfaces such as networks, serial ports, CAN buses, etc. For example, the Fixed-Based Flight Training Simulator is made up of the cabin simulation, craft simulation, instrument simulation, visual and sound simulation, instructor console system, and assistant systems. The architecture is shown in Figure 1.

3   Analysis on Simulation Credibility Evaluation Framework

An MSTS is a typical man-in-loop system; furthermore, it is quite complex and multi-functional. For this type of system, evaluation framework should be built in levels, thus system credibility is integrative. But, exactly how to divide a system into different levels and how many levels a system should have are difficult problems in designing evaluation framework. Even different research papers describe different evaluation frameworks. In general, these frameworks are divided into two categories: frameworks built on simulation processes and frameworks built on simulation results.

Frameworks built on simulation processes are established according to the setup steps of the simulation, including physics, mathematics, and simulation models. These frameworks are also built according to different evaluation activities such as verification, validation, and accreditation. This kind of framework represents the integrated evaluation result of the whole simulation process.
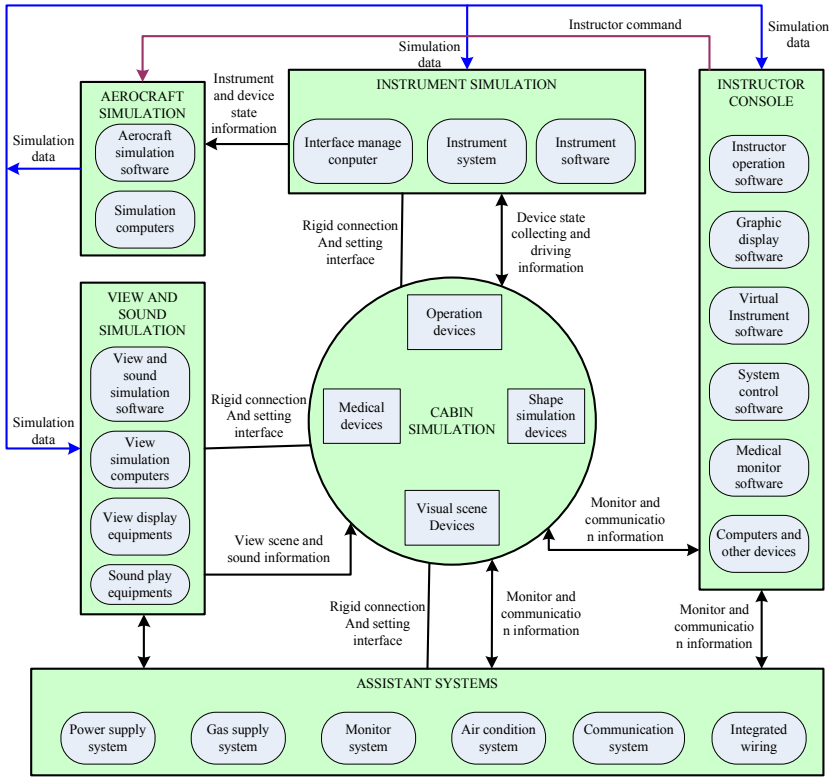

Fig. 1. Architecture of MSTS system

As for frameworks built on simulation results, the evaluation examines if the simulation system is applicable to its purpose, thus determining if the simulation's intention is accurately represented.

The underwater vehicle hardware-in-loop simulation system evaluation framework described in Jiao's paper (2005) is based on simulation processes, including mathematics model validation and accreditation, simulation model validation and accreditation, physical device validation and accreditation, and simulation results analysis. It is applicable when modeling steps and simulation process are well controlled and very clear, and the emphases of evaluation are models and devices. It is more applicable to data simulation systems than man-in-loop simulation systems.

For man-in-loop systems, an evaluation framework built on simulation results is more applicable. In general, the framework is built on two credibility aspects: object models and physical effect devices. In Figure 2, a typical man-in-loop system simulation credibility evaluation framework is shown.
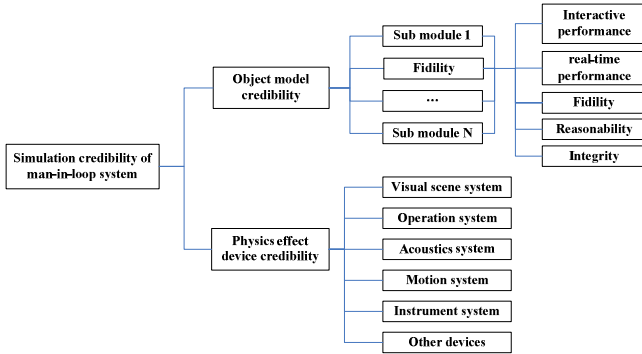

Fig. 2. A typical man-in-loop system simulation credibility evaluation framework

In this framework, the credibility of the models themselves is taken into account, together with the devices that provide physical sensations for "man." This framework can be adequately referenced for man-in-loop systems. But for an MSTS, there are some offsets, particularly during evaluation. Thus, this framework does not adapt to the simulation purpose of an MSTS.

The aim of an MSTS is astronaut training, in which the trainee is not only a participant, but also the service center. The credibility of these systems relies on how much the trainee (astronaut) trusts the system's capability to implement flight tasks. For the evaluation, the trainee's perception of the system is the most important factor. In other words, how closely the simulator makes the trainee feel as though he is really flying an aircraft. What the framework in Figure 2 evaluates is just the system itself, and the evaluation will determine the simulation's effect on the system, which was designed according to certain theories and forms. But as for the question of whether the system will feel real to a trainee, this framework could not answer. For example, in the simulated view system evaluation, credibility is evaluated on the software and hardware of view systems respectively, and the integrated result will be calculated from the results of the software and hardware. However, to answer the question of whether this system, made up of such software and hardware, will invoke a realistic visual feeling for the trainee, the framework in Figure 2 could not answer.

4   MSTS Credibility Evaluation Framework

4.1   Framework Method Analysis

Previous research describes that when analyzing credibility of a simulation system, the first target is to analyze the application purpose of the system. An MSTS aims to

provide astronaut training, and the result of the system is to fulfill the requirements of implementing a flight task for astronauts. If astronauts are trained in a simulator to implement some tasks, the simulator should have the ability to simulate a corresponding work environment and task process. As to whether the simulator does this is determined by the astronaut who implements the task. Thus, in analysis on credibility of a simulator, the evaluation framework should be built from the standpoint of the astronaut.

When a person perceives his surroundings, he undergoes a cognitive process that includes three phases: perception, thinking, and memory. Wherein apperception includes feelings and aesthesis, feelings are the basic process of apperception, which reflect the individual properties of objects felt by man's sensorium, for example, the light, color, and rigidity of objects felt by aural, olfactory, and haptic senses. Aesthesis are built on the basis of feelings and reflect the holistic properties and relationship of objects, such as the shape, size, and category of objects. Feelings and aesthesis make up apperception. Beyond apperception is the mental activity of thinking which, according to the information collected by apperception, and paired with intelligence, the relationship inside the objects and the natural properties of objects are recognized after analyzing, colligating, comparing, abstracting, and generalizing. Following the mental activity of thinking is memory, which gathers knowledge and experience from thinking and keeps these in the brain. The whole process is shown in Figure 3.
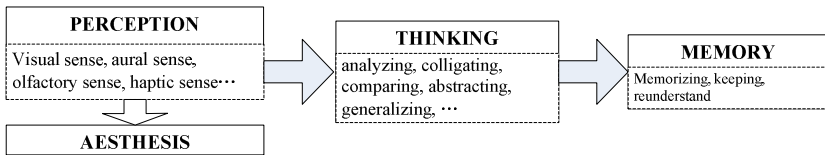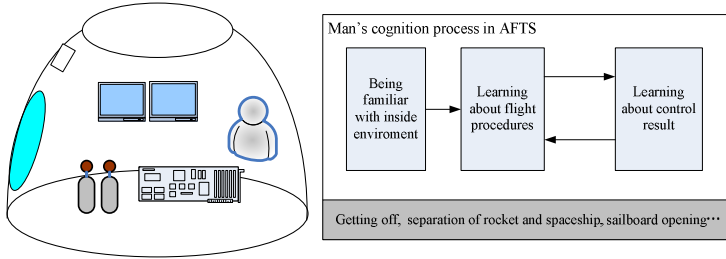

Fig. 3. Man's cognitive process

When an astronaut is training in an MSTS, first he would look around to observe the layout of the cabin, recognize the site and outlook of every device, thus building an inter-space profile in the brain. During the flight process, the current conditions of the aircraft are known through all the apperception results felt by visual and aural senses (MSTS does not relate to other perceptions such as olfactory, taste, and balance senses), thus the astronaut can fully experience the whole flight process. When the simulation begins, the trainee understands the result of his operations by his psychomotor sense and the state of the simulator after operation. The process is shown in Figure 4.

The environment and process that the MSTS creates stimulate the astronaut's apperception system in several ways, including visual, aural, and psychomotor senses; therefore, the astronauts would believe what happens in an MSTS and what he should do during training will be the same as for actual flight.

The areas that the MSTS reproduces can be divided into static simulation and dynamic simulation. Static simulation only relates to the properties of the simulated objects and will not change with the trainee's operation or flight process. Examples of

static simulation include inside environment simulation, operational power of a device, etc. Dynamic simulation refers to factors that will change during the flight process or astronaut's operation, such as the porthole view, RVD process, etc. In an MSTS, examples of static simulation and dynamic simulation are shown in Table 1.

Fig. 4. Man's cognitive process in MSTS

Table 1. Examples of static simulation and dynamic simulation in MSTS

Perception	Static simulation	Dynamic simulation
Visual	Inside environment	Simulated view
		Instrument data
Aural	Ambient background noise	Sound during flight such as engine retrofire, separation of spaceship and rocket
Psychomotor	Operational power and operation mode of device	Parameter changes, visual changes, and aural changes with man's operation, for example, attitude and position vary with operation of joy sticks

In evaluating dynamic simulation, besides visual, aural, and psychomotor evaluation contents, evaluation on dynamic simulation effect is more important. For the astronaut, simulation results are shown on the interfaces such as the porthole view, instrument data, and sound heard during flight. Therefore, the simulation purpose is to fulfill the requirements of instruments display, changing the view and sound, thus making the astronaut feel the same as he would on an actual flight by seeing changes on instrumentation and his view, and hearing changes in sound during flight and operation.

4.2   Credibility Evaluation Framework

According to previous analysis, MSTS credibility evaluation carries input from two angles: static and dynamic simulation, and visual, aural, and psychomotor senses. Of these, the psychomotor sense is special. Although it has properties in both static and dynamic simulation, in the actual simulation process, psychomotor sensations are felt when the trainee is operating the system. Thus, the psychomotor sense will be analyzed in dynamic simulation. Simulation credibility evaluation framework of an MSTS is shown in Figure 5.
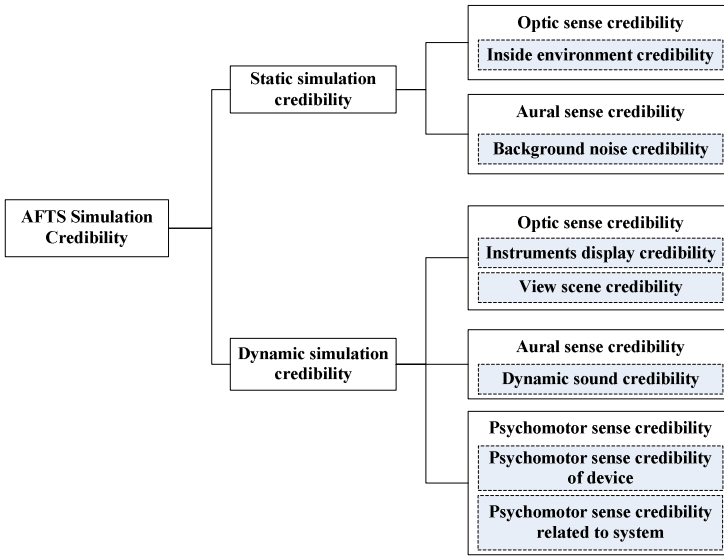
Fig. 5. Simulation credibility evaluation framework of MSTS

In Figure 5, the evaluation factors of static simulation credibility usually do not change, and we usually analyze these properties. But for dynamic simulation credibility, evaluation targets are related to flight processes. For different processes, the evaluated content is different. So, in evaluating dynamic simulation credibility, dynamic simulation processes should be distinguished first and credibility evaluation should be done for different processes.

When analyzing the simulation credibility of visual, aural, and psychomotor senses, it must be determined how to convert these properties into things that can be measured quantitatively. First, the process of forming a perception should be analyzed. The trainee receives significant amounts of information about the physical environment and about himself through sensorium. In sensorium, receptors accept information stimulated from the outside (such as light waves, sound waves, and skin distortion, etc), then collect and transact all the information and transfer it to the brain. The brain recognizes the information and explains it, and apperception and cognition about the information will be formed; the whole process is shown in Figure 6.
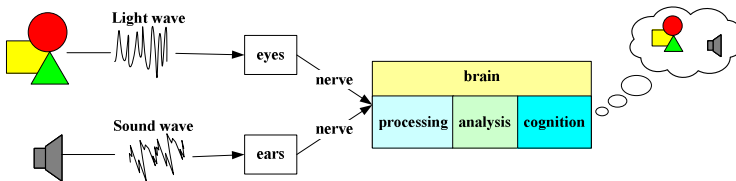

Fig. 6. Apperception process to outside stimuli

In researching different types of sense credibility, the most natural method is to analyze information itself. For example, when evaluating aural sense credibility, a sound wave is the best research object since a sound wave is easy to collect and measure. Thus, aural sense credibility could be evaluated based on amplitude/intensity, frequency, and orientation of a sound wave. However, for visual sense credibility, it is very hard to collect and measure light waves. Further, the psychomotor sense has a complex forming process, and it is almost impossible to collect or measure the related information. For these types of credibility, analysis should work around things that produce or influence information, listing all the related properties.

According to theory of human optics, eyes perceive light, and are sensitive to light and shade, shape, color, movement, and distance of objects. In the real world, this could mean lights in a room, and the shape, size, location, and velocity of objects all influence the visual sense. Thus visual sense credibility evaluation criteria should include all of the ideas above.

Psychomotor sense-forming processes include physical movement, muscles flexing, skin distortion, perhaps adding optic and aural stimulation — all of the things that contribute to nerve stimulation. When evaluating the psychomotor sense, objects that produce or influence the above should be taken into account. Since operation devices have many types, and different types of devices have different properties that influence the psychomotor sense, the evaluation objects of psychomotor sense credibility are quite different. In general, it could be evaluated on such aspects as device surface material, operation mood, operation power, and operation feedback. Therein, operation feedback is likely to include visual and aural feedback and should be further analyzed according to visual sense credibility and aural sense credibility evaluation objects.

Thus, for visual, aural, and psychomotor sense, the basic credibility evaluation objects are listed in Table 2.

Table 2. Visual sense, aural sense, and psychomotor sense basic credibility evaluation objects

Type of Perception	Credibility evaluation objects
Visual	Location, shape, size, color, material reflecting of objects
Aural	Intensity (amplitude), frequency, and orientation of sounds
Psychomotor	Surface material, operation mood, operation power, operation feedback of device

Dynamic simulation credibility evaluation objects are based on basic credibility evaluation objects, adding the properties which influence the dynamic parts. For example, dynamic aural sense evaluation should take the process of changing frequency and time of sound into account, and dynamic visual sense evaluation should concern visual changes. The applicability and accuracy of this framework was basically proved by a questionnaire given to users of the MSTS and experts on simulation credibility.

5   Conclusions and Suggestion for Future Works

In this paper, the simulation credibility evaluation framework of an MSTS that includes visual, aural, and psychomotor sense credibility is fundamentally established according to the purpose of the MSTS and the subject's perception forming process. As to whether the framework adapts to the actual evaluation, how to build different evaluation frameworks for different simulators with different purposes, and what kind of research should be carried out based on the framework are all considerations in MSTS credibility research work. Further detailed research should be discussed in depth and carried out in order to build an applicable simulation credibility evaluation framework and method for an MSTS.

References

1. Vincenzi, D.A., Wise, J.A.: Human Factors in Simulation and Training. CRC Press (2008)
2. Schricker, B., Framceschini, R., Johnson, T.: Fidelity Evaluation Framework. In: Proceedings of the IEEE 34th Annual Simulation Symposium, Seattle, WA (2001)
3. Rehmann, A.J.: A Handbook of Flight Simulation Fidelity Requirements for Human Factors Research. Rep. No. DOT/FAA/CT-TN95/46. National Technical Information Service, Springfield, VA (1995)
4. Zhang, W., Wang, X.: Simulation Credibility. Journal of System Simulation, 312–314 (2001)
5. Jiao, P., Zha, Y.: Application of AHP in Integrative Process for Assessing the Credibility of Guidance System Simulation. Journal of Computer Simulation, 68–72 (2005)
6. Zhang, W., Wang, J., Wang, X.: Study on Credibility of Physical Effects of Man-in-the-Loop Simulation System. Journal of System Simulation, 815–820 (2002)
7. Liu, S., Liu, X.: A New Method of Elevation of Confidence Level of Large-Scale Perplexing Simulation System. Journal of System Simulation, 666–669 (2001)
8. Hays, R.T., Singer, M.J.: Simulation Fidelity in Training System Design. Springer, New York (1989)
9. Roza, M.: Fidelity Considerations for Civil Aviation Distributed Simulation. In: Proceedings of the AIAA Modeling and Simulation Technologies Conference and Exhibit, Denver, CO (2000)
10. Roza, M., Voogd, J., Jense, H.: Defining, Specifying and Developing Fidelity Referents. In: Proceedings of the 2001 European Simulation Interoperability Workshop, London (2001)

Distributed Simulation Method for Homing Missiles Guidance, Navigation, and Control

Chen Dong, Tao Chao, Songyan Wang[*], and Ming Yang

Control and Simulation Center, Harbin Institute of Technology,
150080 Harbin, P.R. China
{chendong,chaotao,myang}@hit.edu.cn,
wangsyhit@yahoo.com.cn

Abstract. In view of efficient models and equipments integration, a distributed simulation method orienting homing missiles GNC demonstration and performance assessment is presented. GNC system of homing missile is introduced briefly as foundation. Distributed simulation method of GNC demonstration is presented. Operation principle, time promoting, and model operation sequence are discussed. Base on this method, a distributed simulation system is developed for GNC demonstration and performance assessment. Architecture of the simulation system, including hardware structure and software structure, is illustrated. Utilizing this system, GNC system of an infrared imaging homing missile is demonstrated with consideration of laser active jamming, disturbances and model uncertainties. Simulation results are obtained for GNC performance assessment. Demonstration and performance assessment of GNC are efficient and economical by the distributed simulation method.

Keywords: distributed simulation, homing missiles, guidance, navigation, and control, infrared imaging, laser active jamming.

1 Introduction

Distributed simulation is an efficient tool to deal with large-scaled system simulation and analysis. Based on distributed simulation, models from different sources can be integrated into one large model without having to recode the model logic of each of the individual models [1]. And the models are characterized by reusability and interoperability in distributed simulation [2]. With these characters, models developed by different departments can be assembled easily into a large-scale simulation system. Extension to this system becomes flexible. In addition, simulation execution time of a sequential simulation is decreased [3]. Economy and efficiency of simulation are achieved. With these advantages, distributed simulation has been applied in many fields, such as aerospace [4], manufacturing industry [5], power system [6], business [7], telecommunication [8], etc.

---

[*] Corresponding author.

In research and development of homing missile, design of Guidance, Navigation, and Control (GNC) system is an important part. Operational efficiency of a homing missile is strongly affected by performance of GNC system. Thus, demonstration to GNC system is necessary. As the development of homing missiles, GNC system becomes 'smart' and complicated to handle complex warfare situation, such as jamming, disturbances, uncertainties, maneuver targets, etc. These factors should be modeled and considered in GNC system demonstration and performance assessment. Based on distributed simulation method, models and equipments developed by different departments can be assembled and constitute a large simulation system. Extending the simulation system according to detailed simulation requirements is flexible.

This paper presents a distributed simulation method orienting homing missiles GNC demonstration and performance assessment. GNC system of homing missile is introduced briefly. Distributed simulation method of GNC demonstration is presented. Operation principle, time promoting, and model operation sequence are discussed. Base on this method, a distributed simulation system is developed for GNC demonstration and performance assessment. Hardware structure and software structure of the simulation system are introduced. GNC system of an infrared imaging homing missile is demonstrated vie the distributed simulation system. Jamming, disturbances, and uncertainties are taken into account. Simulation results are obtained for GNC performance assessment.

2   GNC System of Homing Missile

A general GNC system of homing missile is introduced briefly for understanding GNC principles. Main components of the GNC system are presented. Operation principle of GNC system is illustrated. And the correlative models are given. These are the foundations of distributed simulation for GNC system.

2.1   Main Components of GNC System

For a homing missile, the GNC system mainly includes seeker, airborne computer, and actuators. Moreover, there are various sensors measuring flight states. Based on these components, GNC algorithms can be implemented.

Seeker. In GNC system, seeker is a key equipment measuring data concerning missile-target engagement geometry. A servo system is functioned to adjust heading of the seeker, by which target is tracked and the data can be acquired continuously during flight. Based on these data, sightline angles and their rates are determined. Rates of sightline angles and other data are sent to airborne computer for solving guidance and control commands.

Airborne Computer. GNC algorithms are loaded in airborne computer. The data received by airborne computer is filtered to eliminate disturbances. According to these data, guidance algorithm solves guidance commands. Usually, the commands are desired aerodynamic loads. Control algorithm calculates commands of control surface deflections, which are sent to actuators.

Actuators. Commands of control surface deflections are followed with actuators. Through deflecting control surface, aerodynamic forces and moments suffered by the missile are changed. The desired aerodynamic loads are achieved. And missile can be guided to target.

2.2   Operation Principle of GNC System and Models

Operation principle of GNC system is illustrated as Fig. 1. In Fig. 1, a closed loop is composed of four main parts, i.e. GNC, missile dynamics, target dynamics, and missile-target engagement geometry. Moreover, effects of external disturbance, jamming, and model uncertainty are taken into account. Models, input, and output of the four parts are introduced briefly. Detailed models can be found in [9].
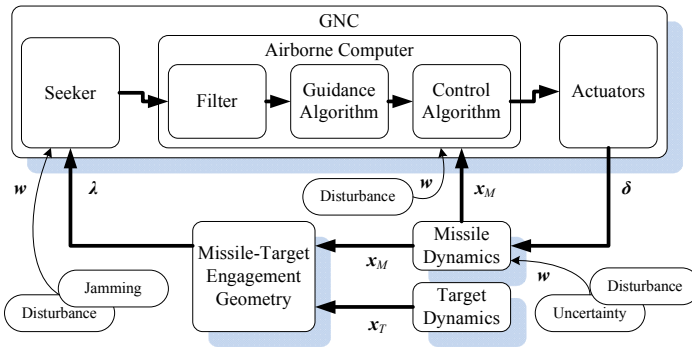
Fig. 1. Operation principle of GNC system

Model of GNC system is expressed as follows:

$$\boldsymbol{\delta} = f_{GNC}(t, \boldsymbol{\lambda}, \boldsymbol{x}_M, \boldsymbol{w}) . \tag{1}$$

In equation (1), f_{GNC} is a nonlinear function describing relationship between output and input of GNC system. $\boldsymbol{\delta}$ is vector of control surface deflections. t denotes time. $\boldsymbol{\lambda}$ is vector sightline angles. $\boldsymbol{x}_M$ is vector of flight states. Its elements include flight data measured by sensors. $\boldsymbol{w}$ is vector of disturbances. It describes external disturbances and jamming impacting on GNC system.

Missile dynamics is described by

$$\boldsymbol{x}_M' = f_{Missile}(t, \boldsymbol{\delta}, \boldsymbol{x}_M, \boldsymbol{w}) . \tag{2}$$

Equation (2) is a series of differential equations, which describe dynamics and kinematics of missile. $\boldsymbol{x}_M'$ is a vector, whose elements are derivatives of flight states with respect to time. $\boldsymbol{w}$ is a vector denoting external disturbances and model uncertainties. $\boldsymbol{x}_M$ can be obtained from initial values $\boldsymbol{x}_M(t_0)$ via numerical integration.

Target dynamics is described by

$$\boldsymbol{x}_T' = f_{Missile}(t, \boldsymbol{p}, \boldsymbol{x}_T) . \tag{3}$$

Where, x_T is state vector of target motion. x_T' is a vector describing rate of x_T. p is a vector containing parameters describing motion characters of the target. x_T can be integrated numerically from its initial values $x_T(t_0)$.

Missile-target engagement geometry is expressed by

$$\lambda = f_{MTEG}(x_M, x_T) \ . \tag{4}$$

Sightline angles can be computed by equation (4).

3   Distributed Simulation Method

Based on the operation principle of GNC system, distributed simulation method for GNC demonstration is presented in this section. Operation principle, time promoting , and model operation sequence of distributed simulation are discussed as follows:

3.1   Operation Principle of Distributed Simulation

Flowchart of homing missile GNC demonstration via distributed simulation is illustrated as Fig. 2.
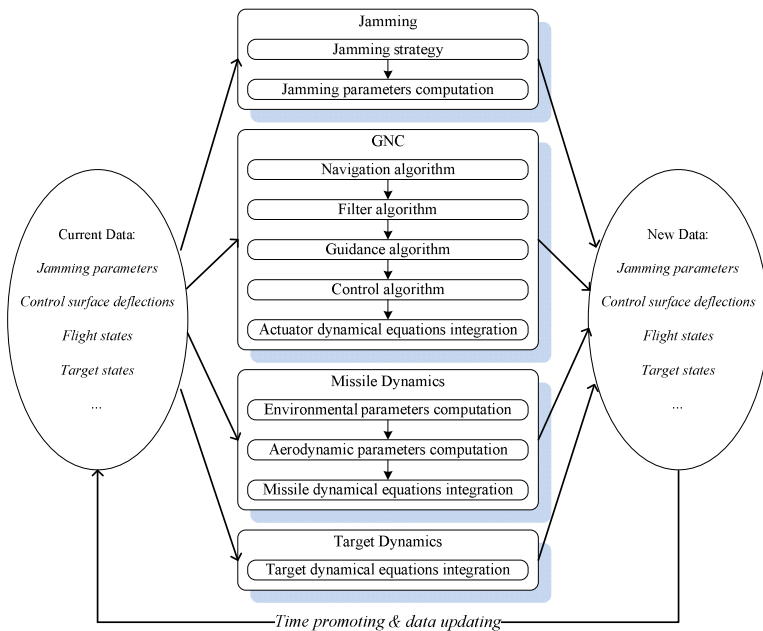
Fig. 2. Flowchart of distributed simulation

Four parts are involved in simulation, i.e. jamming, GNC, missile dynamics, and target dynamics. Simulation time is promoted with a specified increment, i.e. simulation period. In each simulation period, the above four parts are operated

independently, according to current simulation data including jamming parameters, control surface deflections, flight states, target states, etc. When operations of the four parts are completed in one simulation period, new simulation data are generated. The new data will replace current simulation data in the next simulation period.

For GNC part, navigation algorithm is functioned to generate sightline angle, according to relative motion of missile and target. Kalman filter is adopted in filter algorithm to eliminate disturbances in sightline angles and flight states. Guidance algorithm is conventional proportional navigation. Control algorithm is a PID law, which is widely used in engineering. Fourth-order Runge-Kutta method is used in integrating actuator dynamical equations.

For missile dynamics and target dynamics, fourth-order Runge-Kutta method is also applied in integrating the dynamical equations. Environmental parameters and aerodynamic parameters are computed according to flight states.

3.2   Time Promoting and Model Operation Sequence

Time control is important for distributed simulation. In distributed simulation for GNC demonstration, time promoting is regulated by main control of distributed simulation. Simulation time is promoted based on simulation period. All models should complete computation in a simulation period and export simulation data.
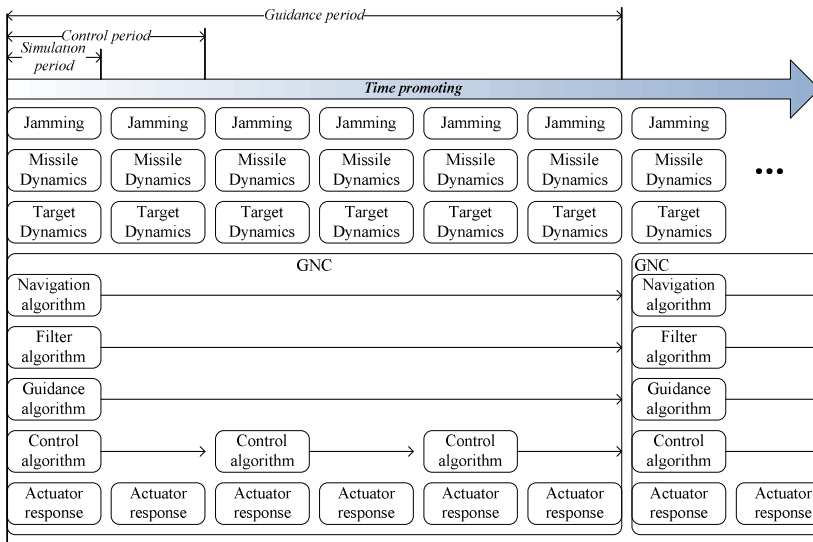
Fig. 3. Time promoting and sequence of model operation

However, some models are unnecessary to operate in every simulation period. For example, a sequence of model operation is displayed in Fig. 3. Jamming, missile dynamics, target dynamics, and actuator response are functioned in each simulation method. But, control algorithm is functioned every other simulation period. When control algorithm is not functioned, its output holds the history value, which drives

actuators. Thus, two simulation periods compose a control period. Similarly, navigation algorithm, filter algorithm, and guidance algorithm are functioned every seven simulation periods. Guidance period is composed of six simulation periods. Guidance commands hold their history values, if guidance algorithm is not active. Guidance period and control period are set according to actual GNC parameters. Simulation period are set with consideration of operation efficiency and computation accuracy.

4   Distributed Simulation System

Base on the aforementioned work, a distributed simulation system is designed for homing missiles GNC demonstration with infrared imaging seeker. For other types of homing missile, the system is also available by adjusting seeker model and correlative devices. Architecture of the distributed simulation system is described by the following hardware structure and software structure, respectively.

Hardware Structure. Hardware structure of the distributed simulation system is illustrated as Fig. 4. The system is composed of nine subsystems, a network card, and communication cables.

The nine subsystems are main control subsystem, GNC subsystem, missile dynamics subsystem, target dynamics subsystem, evaluation subsystem, data base subsystem, jamming subsystem with laser jamming equipment and optical coupling device, image collection subsystem with CCD camera, and visualization subsystem with projectors. Subsystems are connected to network card by communication cables.
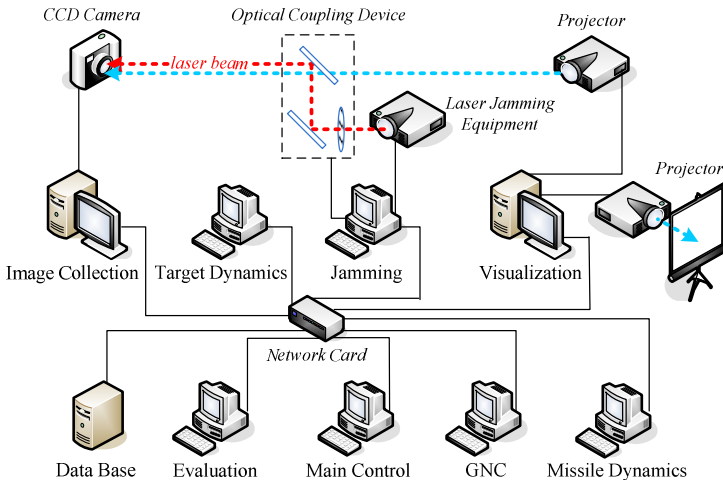
Fig. 4. Hardware structure of the system

Software Structure. Software structure of the distributed simulation system is presented in Table 1. In each subsystem, communication module is included. Data exchange between different subsystems is conducted via communication module following communication rules. Besides communication module, the subsystems contain other function modules.

Table 1. Function of the modules contained in nine subsystems

Subsystem	Module	Function
Main Control	Command	Sent commands to other subsystems.
	Clock	Provide simulation time for system.
	Initialization	Initialize simulation scenario and equipment parameters.
Target Dynamics	Dynamics	Solve dynamical equations of target.
	Kinematics	Solve kinematical equations of target.
Missile Dynamics	Dynamics	Solve dynamical equations of missile.
	Kinematics	Solve kinematical equations of missile.
	Aerodynamics	Provide aerodynamic parameters with disturbances and uncertainties.
	Environment	Provide environmental parameters with disturbances and uncertainties.
Data Base	Data base management	Store and manage experiment data.
Evaluation	Performance assessment	Conduct performance assessment.
GNC	Navigation	Provide navigation solution.
	Guidance	Provide guidance commands.
	Control	Provide control solution and follow guidance commands.
Image Collection	Image processing	Collect and process infrared images.
	Target identification	Determine target position in images.
Visualization	Image generation	Generate and project infrared images.
	Scene display	Display situation.
	Data display	Display experiment data.
Jamming	Optimal coupling device control	Regulate position and attitude of lens to simulate incidence angle change and attenuation of laser beam.
	Laser equipment control	Regulate frequency and power of laser beam.

5   Demonstration

Utilizing the distributed simulation system, simulations are performed to demonstrate GNC system of an infrared imaging homing missile.

For a specified simulation scenario, the scene generated by visualization subsystem is shown as Fig. 5 (a). For the same simulation scenario, the distributed simulation system is operated two hundred times independently to acquire experiment data for performance assessment. Laser active jamming, disturbances and uncertainties in missile parameters, atmospheric environment, and aerodynamic coefficients are taken into account. Distribution of miss distances deriving from two hundred experiments is displayed in Fig. 5 (b). Miss distances and other experiment data deriving from experiments can be used for GNC system performance assessment.

(a)   (b)
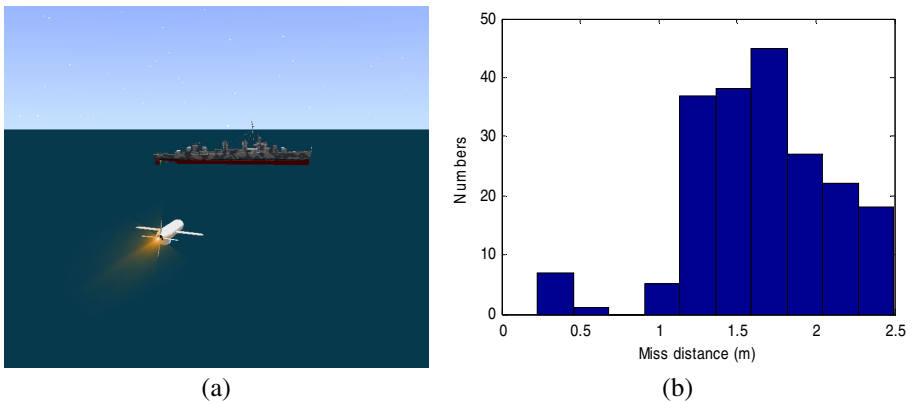
Fig. 5. Simulation results (a. Scene generated via visualization subsystem; b. Distribution of miss distances deriving from two hundred experiments)

6   Conclusions

A distributed simulation method orienting homing missiles GNC demonstration and performance assessment is presented. By this method, integrating simulation models and equipments into a large-scaled simulation system is flexible. According to different simulation requirements, extending or updating simulation system is flexible. Various GNC system models, missile dynamics models and jamming equipments can be involved in demonstration efficiently and economically, according to different simulation requirements. These characters are advantageous for GNC system demonstration and performance assessment. Cost of GNC system research and development can be decreased by the distributed simulation method presented in this paper.

References

1. Verbraeck, A.: Component-based Distributed Simulations: The Way Forward? In: 18th Workshop on Parallel and Distributed Simulation (PADS 2004), pp. 141–148. IEEE Press, New York (2004)
2. Boukerche, A., Zhang, M.: Towards Peer-to-Peer Based Distributed Simulations on a Grid Infrastructure. In: 41st Annual Simulation Symposium, pp. 212–219 (2008)
3. Bruschi, S.M., Santana, R.H.C., Santana, M.J., Aiza, T.S.: An Automatic Distributed Simulation Environment. In: 2004 Winter Simulation Conference, WSC 2004 (2004)
4. Miao, Y., Chen, C.-C., Sun, Z.W.: A Satellite System Distributed Simulation Design and Synchronous Control. In: 2009 IEEE International Conference on Mechatronics and Automation, pp. 3889–3893. IEEE Press, New York (2009)
5. Raab, M., Masik, S., Schulze, T.: Support System for Distributed HLA Simulations in Industrial Applications. In: Principles of Advanced and Distributed Simulation (PADS 2011), pp. 1–7. IEEE Press, New York (2011)
6. Huang, Q.H., Wu, J., Bastos, J.L., Schulz, N.N.: Distributed Simulation Applied to Shipboard Power Systems. In: Electric Ship Technologies Symposium (ESTS 2007), pp. 498–503 (2007)
7. Robinson, S.: Distributed Simulation and Simulation Practice. In: 7th International Symposium on Distributed Simulation and Real-Time Applications (DS-RT 2003). IEEE Press, New York (2003)
8. Iazeolla, G., Pieroni, A., D'Ambrogio, A., Gianni, D.: A Distributed Approach to Wireless System Simulation. In: 6th Advanced International Conference on Telecommunications (AITC), pp. 252–262. IEEE Press, New York (2010)
9. Pan, B.F., Tang, S., Wie, B.: Rapid Prototyping of a Guidance and Control System for Missiles. In: AIAA Guidance, Navigation and Control Conference and Exhibit. AIAA, Washington, DC (2008)

Numerical Simulation and Experimental Verification for Rice Using Distinct Element Method

Tomoyuki Miyamoto[1], So Noguchi[1], and Shinya Matsutomo[2]

[1] Hokkaido University, Graduate School of Information Science and Technology,
Kita 14 Nishi 9, Kita-ku, Sapporo 060-0814, Japan
satake_miyamoto@yahoo.co.jp
[2] Niihama National College of Technology, Department of Electronic Control Engineering,
7-1 Yagumo-cho, Niihama 792-8580, Japan
shin@ect.niihama-nct.ac.jp

Abstract. Distinct Element Method (DEM) has been developed for simulating behavior of discontinuous material by considering interaction between distinct elements, *i.e.* particles. After confirming the validity of DEM through some analysis, we can investigate behavior of material on a chute of a Sorting Machine with DEM. In the previous paper, we have reported that DEM is useful in simulating particles with various shapes, that is sorted by a Sorting Machine. The purpose of this paper is to discuss designing for an effective chute of a Sorting Machine from DEM results. At first, a repose angle of various types of rice in a stock box was simulated and the simulation results were compared with experimental ones. Secondly, the behavior of rice on a chute of a Sorting Machine was simulated and the simulated behavior was compared with actual one. Finally, we designed an effective chute of a Sorting Machine with DEM.

Keywords: DEM, Raw Rice, Parboiled Rice, angle of repose, Sorting Machine.

1 Introduction

A Sorting Machine that food safety problems have been solved is demanded for sorting grains, such as rice, barley, and wheat. Foreign material, such as a stone, has to be automatically removed in the Sorting Machine. In Japan, Sorting Machines have been developed for middle- and short-grain rice but they are unsuitable for other grain, such as long-grain rice. In recent years, due to the world food situation, it is desired to apply Sorting Machines to various kinds of grain. Therefore, in order to effectively develop such Sorting Machines, it is required to simulate the behavior of grain in them. We have developed a Distinct Element Method (DEM) simulation code to investigate the behavior of grain. By simulating a repose angle of grain in a stock box, the effectiveness of the developed DEM was confirmed in [1].

In this paper, fundamental simulation data of a rice processing machine is described for optimum planning, effective designing, and cost-reduction. Rice variously and complicatedly behaves in a processing machine because amount of

water, degree of adhesion, and grain size are various and depend on a breed in the rice, even if the shape of the rice grains is the same. The behavior of rice strongly affects the performance of a rice processing equipment. Numerical simulation was carried out using three-dimensional DEM in order to highly accurately reproduce the behavior of rice in a Sorting Machine. Furthermore, we have designed a new type of chute of a Sorting Machine from the simulation results.

2   Distinct Element Method

The DEM is a numerical method for computing stress and displacement in volume containing a large number of particles such as grains of sand [2] and [3]. Granular material is modeled as an assembly of rigid particles and interaction between particles is explicitly considered (shown in Fig. 1). Here, in normal and transaction forces, K_n and K_s are the elastic stiffness as springs, η_n and η_s are the viscous damping as dashpots, and μ is the friction slider. Usually, particle shapes and geometries have to be specified by a user. A sphere shape is commonly employed.

Normal Force   Transaction Force
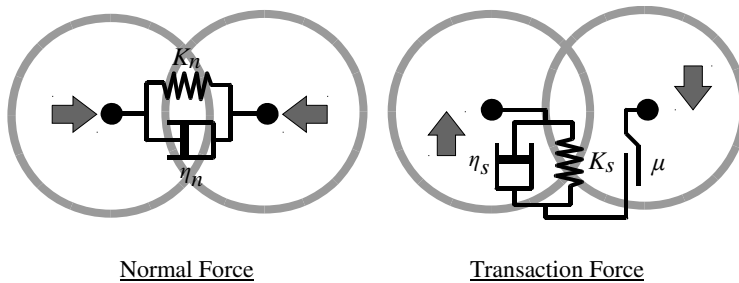
Fig. 1. Contact force model for DEM

3   Agenda Details

3.1   Characteristics of Rice

The varieties of rice are distinguished as Long Grain, Middle Grain, and Short Grain. Moreover, they are classified into Non-glutinous Rice and Glutinous Rice according to starchiness. The best harvested rice in the world is the Long Grain Non-glutinous Rice. The Long Grain is rice that the length is more than 6.6 mm and the aspect ratio (length/width) is less than 3.0. The breeds of the Long Grain are classified into Raw Rice and Parboiled Rice. The Raw Rice means simply milled Paddy and the Parboiled Rice means Paddy milled after steamed and dried. The characteristic of the Raw Rice is much different from that of the Parboiled Rice (shown in Table 1). Despite the same shape of grains, the Parboiled Rice has a higher angle of repose than that of the Raw Rice because the rice surface adhesion becomes high through a steaming process.

Table 1. Angles of repose in varieties of rice

Variety	Angle of Repose
Long Grain Raw Rice	40-42 degrees
Long Grain Parboiled Rice	45-47 degrees

3.2   Subject Outline in Sorting Machine

Commonly, a rice processing equipment has a Sorting Machine which sorts discolored rice and foreign material like a stone and a glass by recognizing them with a high-performance camera and blows away them with an ejector like an air-gun. Fig. 2 shows the schematic view of an Sorting Machine. There are three processes in the equipment. First, material is supplied with a chute. Second, bad particles are recognized with a high-performance camera. Third, the bad particles are blown away with an ejector. As a technical problem of the Sorting Machine, it is necessary to supply material with constant speed and suitable density to a sensing area. Thereby, influence of skin friction against chute surface is reduced. However, the surface structure of a common chute is very simple therefore it is impossible to regulate flow of material.
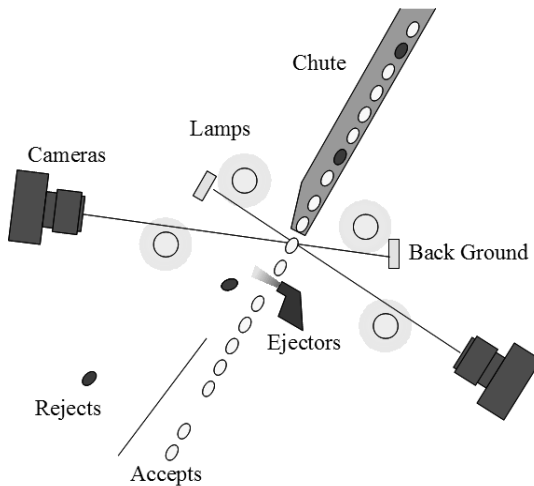
Fig. 2. Schematic view of Sorting Machine

Especially, the flow of the Parboiled Rice on the chute is uneven and agglomerated. The behavior of the grains causes to deteriorate sorting performance (quality, amount of throughput, and yield rate). The Raw Rice and the Parboiled Rice flowing on the chute are shown in Fig. 3. These images were captured by a camera of a Sorting Machine and binarized to make clear the density distribution of grains.

Raw Rice
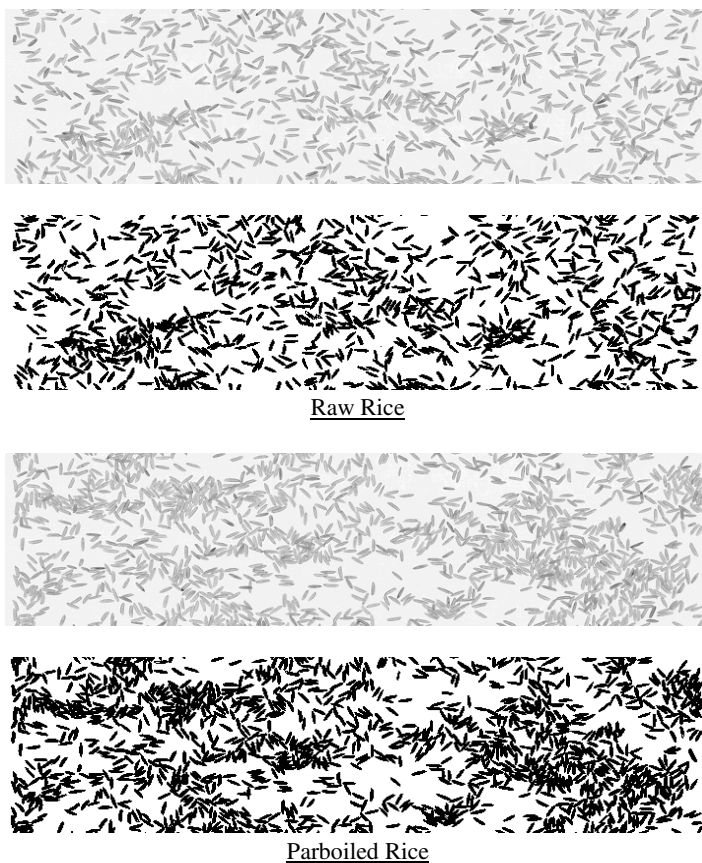
Parboiled Rice

Fig. 3. Captured images by Sorting Machine and binarized images

4 Suggestion for Improvement

4.1 Proposed Model

In the study, it is needed to improve the flow of the Parboiled Rice on the chute of Sorting Machine. The Parboiled Rice is adhesive, that is, it is easy to form a mass of rice. Additionally, the Parboiled Rice shows the feature of a high angle of repose, that is, it indicates less fluidity. It is difficult to change the state of the rice flowing on the chute, because the surface of the chute is less skin friction. The mass of the rice causes a detection error and a rejection of many good rice grains with defectiveness. It is needed to design a new structure that an external force separates a mass of rice to a grain on the chute. Therefore, we propose a new structure of the chute as shown in Fig. 4. The chute structure is designed with three stages at the top of the chute. As structural features, the slope of every stage is slightly larger than that of the ordinary chute and there are concave joints between the stages. Rice flows out as a parabolic orbit to a next stage and receives a small shock when landing on the stage.

Ordinary model   Proposed model
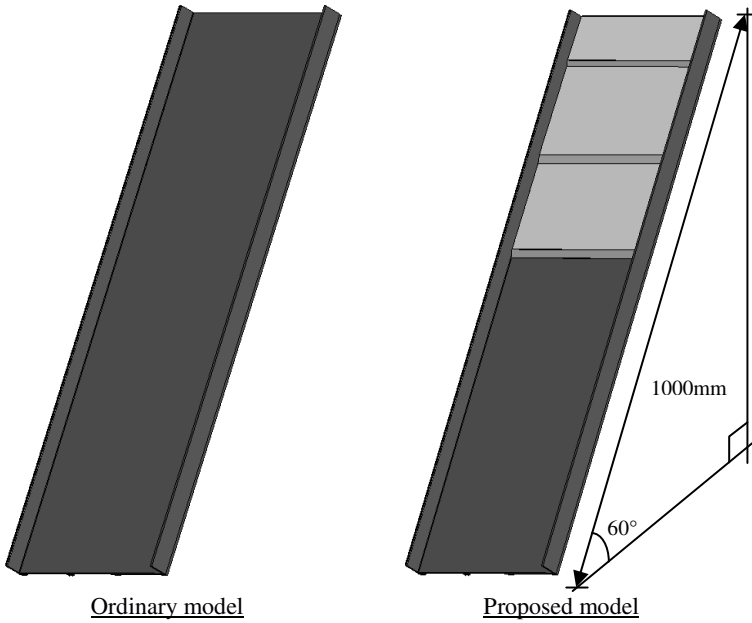
Fig. 4. New structure is newly designed and the proposed structure has three stages at the top of the chute

Experiment   Simulation

Raw Rice

Experiment   Simulation
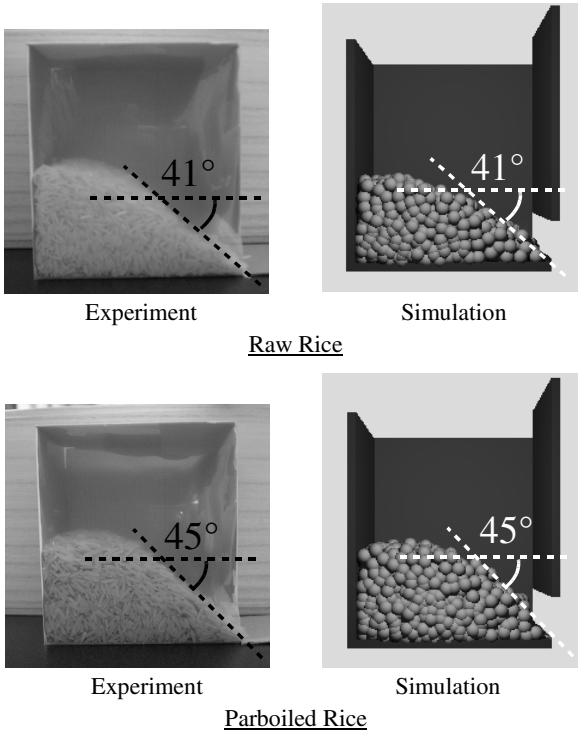
Parboiled Rice

Fig. 5. Simulation and experiment results for angle of repose

At first, an angle of repose for the Raw Rice and the Parboiled Rice in a stock box was simulated with the Ledge Method [4] and the simulation results were compared with experimental ones in order to identify some parameters for the DEM simulation (shown in Fig. 5). Here, rice is modeled as a sphere particle. Accordingly, the behavior of rice on a chute of a Sorting Machine was simulated using these parameters and the simulated behavior was compared with actual one (shown in Fig. 6). The effectiveness of the proposed structure was revealed by the DEM simulation. Then, the actual chute was investigated in experiments.

Raw Rice
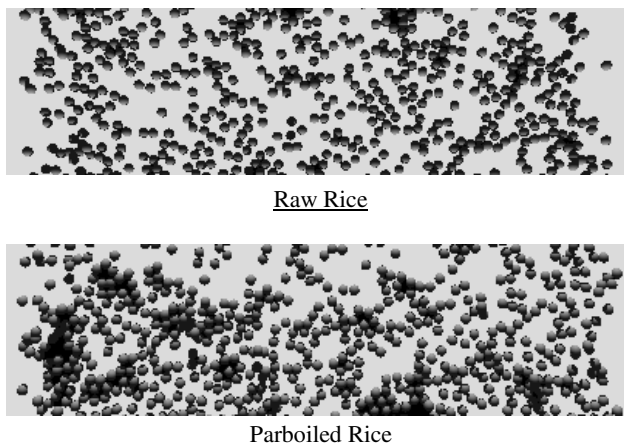
Parboiled Rice

Fig. 6. Simulation results with Raw Rice and Parboiled Rice on ordinary chute of Sorting Machine

4.2   Verification Results

The flow of the Parboiled Rice on the proposed chute of Sorting Machine was simulated with the DEM and it was figured out that it was effective in separating the mass of rice (shown in Fig. 7). Next, the proposed chute was actually made to do experiment. The simulation result agreed with the experimental one well (shown in Fig. 7). Moreover, the proposed structure which has three stages at the top of the chute was successful to separate the mass of rice to a grain.   Because of the concave joints between the stages, the rice flows out as a parabolic orbit to the next stage and receives a small shock to land on the stage in order to separate the mass of rice to a grain. However, the step between the stages is too high for the rice to bounce. Therefore, it is important that the mass of rice receives multiple small external forces. We will need to optimally design the step height between the stages and the number of stages. It is possible to use the DEM simulation for an effective design of the chute of Sorting Machine.

Simulation

Experiment

Ordinary chute model

Simulation

Experiment

Proposed chute model
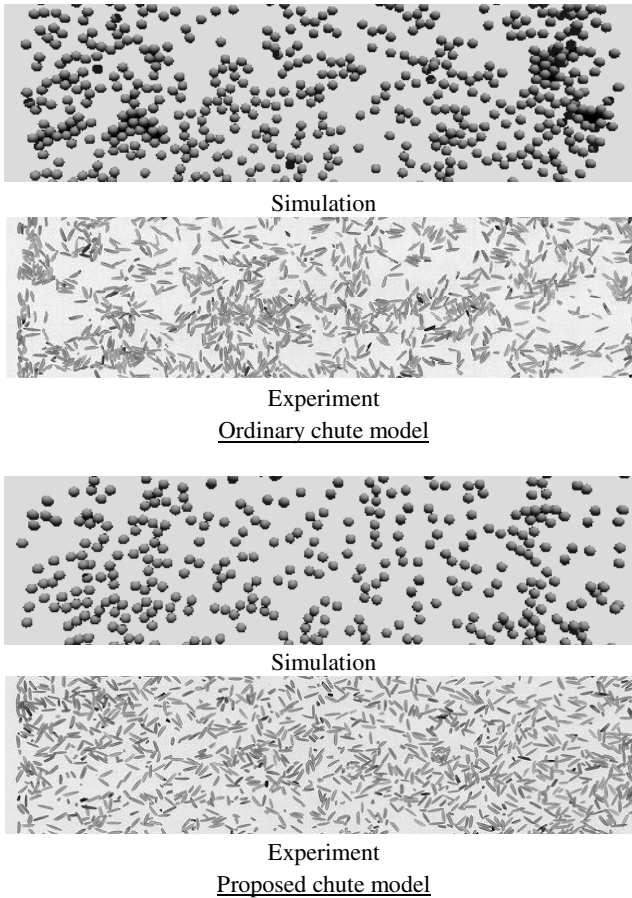
Fig. 7. Simulation results for ordinary and proposed chute model

5   Conclusion

In the study, the numerical simulation was carried out using the three-dimensional Distinct Element Method in order to highly accurately reproduce the behavior of rice in a Sorting Machine. It was figured out that the DEM is effective in the development of the Sorting Machine. Moreover, a new surface structure of chute was designed to improve the behavior of Parboiled Rice in the Sorting Machine with the DEM. The new chute is designed with multiple stages at the top. It was made of sheets of metal with bending work so it is less expensive and will be very easily installed. In addition, the step height between the stages and the number of stages are important for a good flow condition of rice in Sorting Machine. In the future prospects, we will consider to optimally design the structure of the chute in the Sorting Machine with an optimal design method.

References

1. Miyamoto, T., Noguchi, S., Matsutomo, S.: A numerical study for rice using distinct element method. Submitted to JSST 2012 Conference (2012)
2. Pradhan, T.B.S., Sawada, S.: Distinct element analysis of sand under cyclic loading using oval elements. In: Proc. of 10th WCEE, pp. 1163–1168 (1992)
3. Abdeli, M., Miedema, S.A., Schott, D.L., Alvarez Grima, M.: Discrete Element Method Modeling of Sand Cutting. In: WODCON XIX Beijing, Conference Paper (2010)
4. Neikov, O.D., Naboychenko, S.S., Dowson, G.: Handbook of Non-Ferrous Metal Powders: Technologies and Applications, p. 39. Elsevier Science (2009)

Safety Analysis of Computer-Controlled Real-Time Systems with Message Loss Using Communicating DEVS Models

Hae Sang Song[1] and Tag Gon Kim[2]

[1] Computer Engineering Dept., Seowon University,
337-3 Musimseoro, Heungduk-gu,
Cheongju 361-742, Korea
hssong@seowon.ac.kr

[2] Electrical Engineering Dept., Korea Advanced Institute of Science and Technology,
1291 Daehak-ro(373-1 Guseong-dong), Yuseong-gu,
Daejeon 305-701, Korea
tgkim@ee.kaist.ac.kr

Abstract. The Communication DEVS formalism is an analysis means for discrete event systems modeled by DEVS formalism which has been widely used as a system theoretical specification. This paper proposes a new method for analyzing safety of real-time discrete event systems using communicating DEVS formalism. It is a part of efforts toward a unified method for modeling, simulation, and logical analysis based on the DEVS formalism and associate theory. For safety analysis of such real-time discrete event systems we first define communicating DEVS and then propose a timed reachability analysis algorithm for the models. The algorithm visits all possible timed states of the model, which is not always possible by using a simulation based state traversal. The proposed method can be well used especially for DEVS-specified systems to check various logical properties such safety, liveness and so on. A case study of a safety analysis for a rail road crossing system illustrates the usefulness of the proposed method.

Keywords: Safety Analysis, DEVS Formalism, Communicating DEVS, Real-time Systems, Reachability Analysis.

1 Introduction

Computer-based controllers have been widely used as active or passive components of real-time systems such as traffic control, aircraft, nuclear power plants, hospital patient care systems and others [1]. In those systems a problem of safety would be regarded as one of the most important factors, for the consequences of failures of such systems cause serious and vital damages to human life and property. Thus safety analysis needs to be done from the early design stage in developing such computer-controlled real-time systems. In fact, a developer of such systems would implement the design after verification of safety is completed.

Safety of such a real-time system is defined that the whole behavior of the system does not go into any bad state at which the system would cause serious consequences. Thus, safety analysis should visit all states of a system to be analyzed. Safety analysis would not be completely done by simulation of a system model. This is because simulation would not visit all possible states of the model. States to be visited in simulation are controlled by an experimental design of input stimuli which is applied to a system model.

Safety analysis is usually conducted by reachability analysis in which all direct and/or indirect paths from an initial state are generated[1][2][3][4]. In this case the state explosion problem, however, is inevitable as the number of states of components becomes large and the time is an independent variable.

Generally, a generation of all possible states in reachability analysis of a real-time system with multiple components employs state transition rules of each component. Since the real-time system has time constrained such rules should include information of sojourn time between states. Moreover, since each component communicates with each other components the generation needs state transition rules for communicating between components. Furthermore, the generation should consider a message loss in communication between components. Thus, the generation would require a formal model with sound semantics to specify both timed state transition of each component and communication between components with message losses.

There have been several approaches for safety analysis based on different modeling formalisms. Petri-net has been used for modeling and safety analysis of real-time discrete event systems for various applications in recent decades[1][4]. However, it would be difficult to implement a Petri-net model in software in terms of a set of variables with associated operations, for a Petri-net model does not specify a system in terms of variables (or states). RT-DEVS formalism is an alternative for safety analysis of such systems[2]. But RT-DEVS formalism specifies the sojourn times between states in interval. Moreover the analysis did not consider a message loss in communication between components. The timed automaton has been used in this area but it uses integer time rather than continuous time in analysis, which makes the state explosion problem worse[3].

This paper employs the Communicating Discrete Event Specification (C-DEVS) formalism[5] for safety analysis, which has a sound semantics for both timed state transition and communication between components with a message loss. Having DEVS models as components C-DEVS formalism is an analysis means by generation of all possible timed state in a global state space.

The proposed method has been applied to safety analysis for the well-known rail-road gate problem. The proposed safety analysis method, though it is a yet another safety analysis method, is the first one in application of DEVS theory in safety analysis. The proposed method is different from a simulation-based approach using DEVS model in that simulation would not visit all possible states of a system model but visit only partial states set.

This paper is organized as follows. In the next section, we review the DEVS formalism first to define the Communicating Discrete Event Systems Specification (C-DEVS) formalism. Then we propose a C-DEVS model construction procedure from a set of atomic DEVS models. In section 3, we apply the C-DEVS formalism and the construction method to analyze the safeness of a railroad crossing controller problem. We conclude in section 4.

2   Background and Communicating DEVS

2.1   DEVS Formalism

As well known in the literatures [6], the DEVS formalism specifies a discrete event system in two aspects: one for the behavior of a basic component, and the other for the overall structure of a system. The former is modeled with atomic DEVS formalism, which describes the behavior of a unit component not further decomposable, which consists of three sets and four functions.

$$AM = < X, Y, S, \delta_{ext}, \delta_{int}, \lambda, ta >$$

, where

  X: input event set,
  Y: output event set,
  S: sequential state set,
        with total state set $Q = \{(s,e) \mid s \in S, \ 0 \le e \le ta(s) \}$,
  $\delta_{ext}: Q \to Q$: external transition function, for $\delta_{ext}(s, e, x) = (s', e')$, $e < ta(s), e' = 0$.
  $\delta_{int}: Q \times X \to Q$: internal transition function, for $\delta_{int}(s, e) = (s', e')$, $e = ta(s), e' = 0$.
  $\lambda: Q \to Y$: output function, for $(s, e) \in Q, e = ta(s)$
  $ta: S \to R_0^+$: time advance function,
  R_0^+ is the non-negative real number set.

There are two types of transitions of an atomic model: 1) external transitions entailed by external events; and 2) internal transitions in the case of no event occurrence until current state sojourn time has elapsed. In the latter case, just before the internal transition, an output event is produced at the state. In an analogy to the continuous systems, external transitions would correspond to the input driven state transition and internal ones the input-free state transition.

Every state is involved in a time advance, up to which the system can remain at the state. A total state $q = (s, e) \in Q$, often we call it a 'timed state', is a pair of a state and an elapsed time at the state. As an inversion, we would define a residual total state $[\![\tilde{q} = (s, ta(s) - e) \in \tilde{Q}]\!]$ of total state q, or shortly a 'residual state', to be a pair of a state and a residual time that is left to the time advance of the total state.

The coupled DEVS formalism, not recited here, specifies the structure of discrete event systems composed of components communicating with each other through event couplings. In this paper, to focus attention on C-DEVS, we assume that the structure of coupled models of a system is collapsed into a set of flattened atomic models and that any input and/or output events with the same name are all coupled together.

2.2   C-DEVS Formalism

Analysis of a discrete event system modeled by a set of DEVS models could be executed by traversal of the state space without simulation, which generates all

possible sequences of events and/or states that can reach from a given initial state. Communicating Discrete Event Systems Specification, or C-DEVS formalism, formally defines the timed interaction mechanism between N atomic DEVS models. Unlike RT-DEVS [2] that has two types of interaction – asynchronous and synchronous, C-DEVS has three types of interaction: asynchronous, synchronous and event loss.

Consider a system model of two atomic DEVS models, $M = \{AM_i, AM_j\}$, $AM_k = \; < X_k, Y_k, S_k, \delta_{ext,k}, \delta_{int,k}, \lambda_k, ta_k >$ for $k = i, j$. Let the atomic models are currently at their total states $q_i = (s_i, e_i)$ and $q_j = (s_j, e_j)$ respectively. In other words we could say they are at compatible residual states $\tilde{q}_i = (s_i, \sigma_i)$ and $\tilde{q}_j = (s_j, \sigma_j)$ respectively, where $\sigma_i = ta(s_i) - e_i$ and $\sigma_j = ta(s_j) - e_j$ are the residual times. Let's denote by $(s_i, s_i') \in \delta_{int,i}$ that $\delta_{int,i}(s_i, e_i) = (s_i', 0), e_i = ta_i(s_i)$ is defined. Then C-DEVS model of the system M can be defined as follows:

$$CDEVS(M) \equiv AM_i \;||AM_j =< E, S, T, ta, M >$$

where
  E: events set,
  S : composed discrete states set,
  T : transition relation of composed states,
  ta : time advance function,
  $M = \{AM_i, AM_j\}$: set of atomic DEVS models
with the following constraints:
  $E = (X_i \cup Y_i \; \cup \{\varepsilon\}) \times (X_j \cup Y_j \; \cup \{\varepsilon\})$, where ε is a null event,
  $S = S_i \times S_j$,
      where $Q = (S_i \times R) \times (S_j \times R) = S \times R^n$ composed total states set,
  $T \subseteq Q \times E \times Q$,
  $ta: Q \to R$.

The transition relation T and time advance ta are subject to the following rules:

  Rule 1 (asynchronous transition): Transit AM_i only for $\sigma_i = ta_i(s_i) - e_i \le \sigma_j = ta_j(s_j) - e_j$:
    (a) Transition relation

      If $(s_i, s_i') \in \delta_{int,i}$ and $(s_i, !a) \in \lambda_i$ and $(s_j, ?a, s_j') \notin \delta_{ext,j}$, then it implies
          $((q_i, q_j), (!a, \varepsilon), (q_i', \dot{q}_j)) \in T$, for all $j \ne i$, where $q_i' = (s_i', 0), \dot{q}_j = (s_j, ta_j(s_j) - (\sigma_j - \sigma_i))$

    (b) Time advance

      $$ta((q_i', \dot{q}_j)) = min \{ta_i(s_i'), ta_j(s_j) - (\sigma_j - \sigma_i)\}.$$

  Rule 2 (synchronous transition): Transit both AM_i and AM_j with event synchronization for $\sigma_i = ta_i(s_i) - e_i \le \sigma_j = ta_j(s_j) - e_j$:

(a) Transition relation

If $(s_i, s_i') \in \delta_{int,i}$ and $(s_i, !a) \in \lambda_i$ and $(s_j, ?a, s_j') \in \delta_{ext,j}$, then it implies

$((q_i, q_j,), (!a, ?a), (q_i', q_j')) \in T$, for all $j \neq i$, where $q_i' = (s_i', 0), q_j' = (s_j', 0)$

(b) Time advance

$ta((q_i', q_j')) = \min\{ta_i(s_i'), ta_j(s_j')\}$

Rule 3 (lossy transition): Transit only AM$_i$ with **event lost** for $\sigma_i = ta_i(s_i) - e_i \leq \sigma_j = ta_j(s_j) - e_j$:

(a) Transition relation

If $(s_i, s_i') \in \delta_{int,i}$ and $(s_i, !a) \in \lambda_i$ and $(s_j, ?a, s_j') \in \delta_{ext,j}$ with event a lost, then it implies

$((q_i, q_j,), (!a, \varepsilon), (q_i', \dot{q}_j)) \in T$, for all $j \neq i$, where $q_i' = (s_i', 0), \dot{q}_j = (s_j, ta(s_j) - (\sigma_j - \sigma_i))$

(b) Time advance

$ta((q_i', \dot{q}_j)) = \min\{ta_i(s_i'), ta_j(s_j) - (\sigma_j - \sigma_i)\}.$

Note that Rule 1 is asynchronous type of interaction when one model that has the smallest residual time is ready to send an output event while the other is at a state not ready to accept the event; therefore only the sender model will transit. The second rule explains a case where both the sender model and the receiver model with respect to a same event are ready to interact at right time at right state; then both of the model transit to start their new states. The third is intrinsic to C-DEVS, where both the sender model and the receiver one are ready to interact against an event, but the event is lost due to a communication mishap. Then only the sender model transits asynchronously in the same way with the first one.

We remark that the transition rules above of two atomic DEVS models can be easily extended to those of multiple atomic models by assuming model AM_i has the smallest resual time among models and applying the three rules for all the remaining models.

2.3   C-DEVS Model Construction Algorithm

A C-DEVS model for a set of atomic DEVS models can be constructed analytically by the C-DEVS formalism. We start with a given initial total state then we choose among the DEVS models a model with the least residual time. Holding on the model chosen, we pick one of the remaining models and apply the transition rules until all the models are examined. For each destination states determined by the transition rules we repeat the process described above. It ends when there is no destination state left to examine. Formally the procedure to construct a C-DEVS model from a set of models M is as follows.

1. Let $q_0 \in Q$ be the initial state, $Q_{done} \subseteq Q$ the finished states set and $Q_{exam} \subseteq Q$ the set of states to be visited. Initialize that $Q_{done} = T = \emptyset$, $Q_{exam} = \{q_0\}$. Then define $ta(q_0) = \min_{k=1..n}\{\sigma_k\}$, $\sigma_k = ta_k(s_k) - e_k$.
2. Pick a composed total state $q = (q_1, q_2, ..., q_n) \in Q_{exam}$.
3. Choose an element total state $q_i = (s_i, e_i)$ of the composed total state q, where $\sigma_i = \min_{k=1..n}\{\sigma_k\}$, $\sigma_k = ta_k(s_k) - e_k$.
4. Apply every transition rule for q_i and each element state q_j of q, $j \neq i$, which add up valid transitions and time advances into T and ta. If a destination state $q' = (q_1', q_2', ..., q_n')$ is obtained, then insert q' into Q_{exam} only if $q' \notin Q_{done}$.
5. If we have finished step 4, then move the composed state q from Q_{exam} to Q_{done}.
6. Repeat step 2-5 until $Q_{exam} = \emptyset$ or a predestined condition meets.
7. Then, we finally get $CDEVS = < E, S, T, ta, M >$.

The procedure describe above does not guarantee finiteness of transition relation even though the sets S and E are finite, for a composed total state is a pair of a composed discrete state and an elapsed time that might be infinite. However this state explosion problem is not of concern of the paper and in most cases the timed behavior converges if designed correctly in authors' experience.

3   Case Study: Safety Analysis of a Railroad Crossing System

3.1   The Problem Statement

Consider a railroad crossing system (RRC) in which trains and cars share the same crossing area. To prevent a disastrous accident we need safety measures such as a gate that blocks cars passing while a train approaches, and sensors that detect trains approaching and exiting the area. Thus the computer-based controller has two design objectives:

(1) (*safeness*) Prevent any accident by sensing the train's position and manipulating the gate in a right order as well as in a right time.
(2) (*liveness*) Cars eventually go through the gate in a finite time.

We also assume that messages between components of the railroad crossing system might be lost during the bad communication environment due to noise, software bugs or any human errors. To acquire a robust control system that satisfies the objectives, first we need to model the railroad crossing system and design a controller that *might* meet the objectives. Then we need to assure the safeness of the controller before implementation. DEVS formalism is an appropriate specification method for this type of problem since DEVS is very suitable specification formalism for modeling, simulation and analysis of real-time discrete event systems.

3.2   RRC DEVS Models

The railroad crossing system can be specified with three atomic models: TRAIN, GATE and CONTROLLER as in Fig. 1. Each circle denotes a total state (timed state) and a dotted arc an internal transition with an output event, and a solid arc an external transition with an input event. An output event is prefixed by '!' and an input event '?'. As assumed before, events with the same name are all coupled together and the system model is flattened into a set of atomic models.
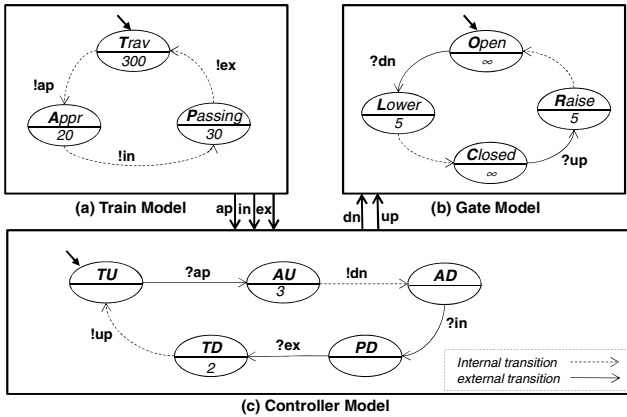

Fig. 1. A Railroad Crossing Model: Train, Gate and Controller

These atomic models in the figure can be mathematically specified in DEVS as follows.

TRAIN $=< X, Y, S, \delta_{ext}, \delta_{int}, \lambda, ta >$
  $X = \{\}, Y = \{! \, ap, ! \, ex, ! \, in\},$
  $S = \{Trav, Appr, Passing\},$
  $\delta_{int}(\text{Trav})=\text{Appr}, \quad \delta_{int}(\text{Appr})=\text{Passing}, \quad \delta_{int}(\text{Passing})=\text{Trav},$
  $\lambda(Trav) = ! \, ap, \qquad \lambda(Appr) = ! \, in, \quad \lambda(Passing) = ! \, ex,$
  $ta(Trav) = 300, \; ta(\text{Appr}) = 20, \; ta(Passing) = 30.$

GATE $=< X, Y, S, \delta_{ext}, \delta_{int}, \lambda, ta >$
  $X = \{? \, dn, ? \, up\}, Y = \{\},$
  $S = \{Open, Lower, Closed, Raised\},$
  $\delta_{int}(\text{Lower})=\text{Closed}, \quad \delta_{int}(\text{Raise})=\text{Open},$
  $\delta_{ext}(\text{Closed}, ?up)=\text{Raise},$
  $\lambda(Lower) = \varepsilon, \quad \lambda(Raise) = \varepsilon,$
  $ta(Lower) = 5, \; ta(\text{Raise}) = 5.$

CONTROLLER$=< X, Y, S, \delta_{ext}, \delta_{int}, \lambda, ta >$
  $X = \{? \, ap, ? \, in, ? \, ex\}, Y = \{! \, dn, ! \, up\},$
  $S = \{TU, AU, AD, PD, TD\}$

$$\delta_{int}(AU)=AD, \quad \delta_{int}(TD)=TU,$$
$$\delta_{ext}(TU, ?ap)=AU, \; \delta_{ext}(AD, ?in)=PD, \; \delta_{ext}(PD, ?ex)=TD,$$
$$\lambda(AU) = !\,dn, \; \lambda(TD) = !\,up,$$
$$ta(AU) = 3, ta(TD) = 2.$$

As we can see, the controller model receives the position events of the train and sends commands to the gate model. For reasons it is assumed that these events might be lost.

3.3 RRC C-DEVS Model Constructed without Event Loss

The safety analysis can be conducted either on the fly during C-DEVS model construction phase or be checked after we once get the C-DEVS model. For safety analysis we first need to define a set of bad states, $B \subseteq S$, any element of which should not be reached from the initial state. In the case study we define the bad set as $B = \{(P, O, -)\}$, that is, the gate is open while the train is passing the crossing area, which is very dangerous.

To be specific, Fig. 2 shown below depicts the transition relation and time advance function of a C-DEVS model for the RRC model, which is obtained by applying only Rule 1 and Rule 2, but not Rule 3. This analysis is conducted to see the composed behavior if no event is lost. The initial composed total state is ((T,O,TU),[0,0,0]) or equivalently ((T,O,TU),[300, ∞, ∞]) in residual form.
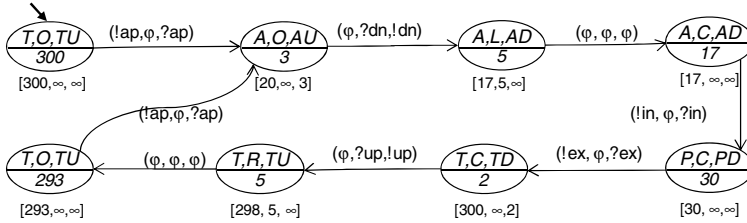
Fig. 2. Transition diagram of RRC model without message loss

We can easily see that each circle represents a composed state (above) with the time advance (below) of the state. A vector below a circle denotes the ordered residual times of each model at the states. For example, the circle upper-left represents that the initial state (T,O,TU) has its residual time vector [300, ∞, ∞], which means model TRAIN is at state Trav at which it may stay for 300; GATE is at state Open and CONTROLLER TU, and both wait forever. Consequently at the composed state the system will stay for 300. An arc represents a transition with an associated event when the time advance is fired.

We conclude in this analysis that there is no bad state (P,O,-) reached from the initial state (T,O,TU) and thus it is said to be safe if we have no message (event) lost. Note that ((T,O,TU),[300, ∞, ∞]) and ((T,O,TU),[293, ∞, ∞]) are not the same state

as the residual times are different. Furthermore we find that the system satisfies the liveness since the state sequence of the gate cycles through (-,O,-) → … → (-,C,-) → … →(-,O,-) in 355. The gate is available for 296 and unavailable for 59 in one cycle, which means cars eventually pass the crossing area.

3.4 RRC C-DEVS Model Constructed with Event Loss

The purpose of this analysis is to know how the system behaves if events or messages are lost. Fig. 3 is the resultant C-DEVS model for the railroad crossing system model in Fig. 1 using the analysis algorithm.   It is assumed that all events might be lost and thus we apply all the three transition rules. We can see that in normal case without event lost the behavior is the same as in Fig. 2. With any event lost, however, we can also see the system digresses to unexpected sequences of states.
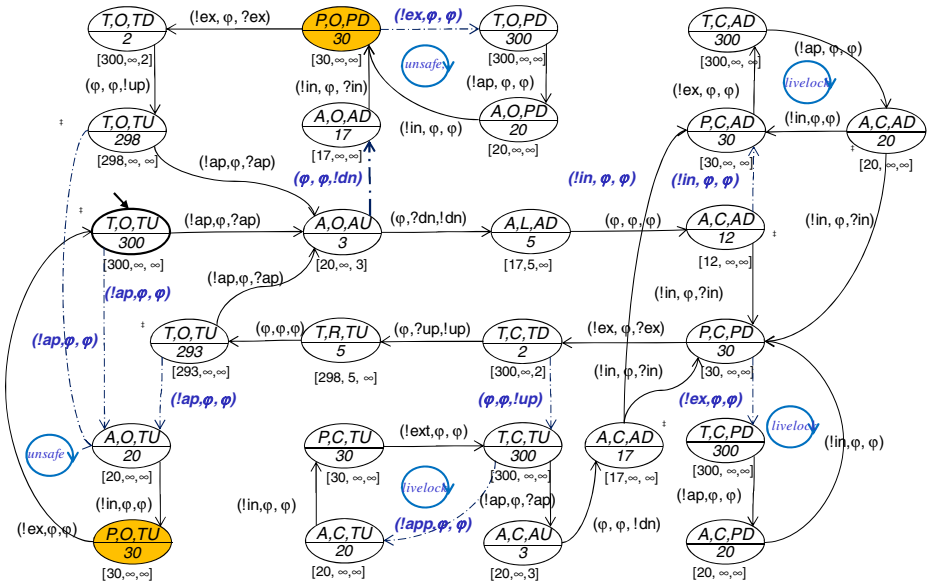
Fig. 3. C-DEVS model with message loss : lost message = {ap, in, ex, dn, up}

The RRC model starts with an initial composed timed state ((T,O,TU),[0,0,0]), which means model TRAIN is at timed state (T,0), and GATE (O,0) and CONTROLER (TU,0). Taking into account the time advances of the atomic models, the initial composed residual state is ((T,O,TU),[300,∞, ∞]) and the time advance of the composed state is 300. After the time elapsed, the train fires and generates an output event !ap, which can be accepted by the controller; then the two models transit synchronously by (!ap,ε,?ap) to a residual state ((A,O,AU),[20,∞,3]). However, it is possible the output event could be lost and the controller cannot receive the event. In this case the system will transit to a residual state ((A,O,TU),[20,∞,∞]) where the train is approaching while the gate is open and the controller never perceive the approching due to the event loss; thus after the time 20 has elapsed, the train is start to

enter the crossing area while the gate is still open, i.e., ((P,O,TU),[30,∞,∞]) that is a bad state. This means if event !ap is lost at (T,O,TU) then the whole system become unsafe. Moreover, we have another bad state ((P,O,PD),[30,∞,∞]) at top of Fig. 3, which comes from ((A,O,AU),[20, ∞, 3]) by the gate's losing event !dn sent from the contoller.

Thus we can conclude by this analysis that the system is unsafe in case of event losses, especially with event !ap and !dn lost. Fortunately any of the other event losses does not affect the safeness of the whole system. However, we can find that those violate the liveness as annotated by 'livelock' in the figure.

4   Conclusion

This paper proposed a safety analysis method for real-time discrete event systems based on the C-DEVS formalism that is mathematically renewed here. C-DEVS itself only defines the interaction mechanism between atomic DEVS models with a representation of the whole behavior as the result of the interaction. It is beneficial for analysis purposes to get the all possible behavior of DEVS models without simulation. We proposed a procedure to construct the C-DEVS model of a set of atomic DEVS models and illustrated the usefulness by a case study, the railroad crossing control system. For this, we first modeled the system in the DEVS formalism and then obtained the C-DEVS model using the construction procedure proposed here. From the resultant model we formulated and analyzed the safety of the system. The state explosion problem due to time, however, is inevitable and is not concern of the paper, though. As a future work, we will extend this research to the controller design problem, where the time advance is not fixed. In addition to this, we have a plan to develop a tool to automatically construct a C-DEVS model from an implemented DEVS simulator.

References

1. Leveson, N.G., Stolzy, J.L.: Safety Analysis Using Petri Nets. IEEE Trans. Software Engineering SE-13(3), 386–397 (1987)
2. Song, H.S., Kim, T.G.: Application of Real-Time DEVS to Analysis of Safety-Critical Embedded Control Systems: Railroad Crossing Control Example. Simulation 81, 119–136 (2005)
3. Bengtsson, J.E., Yi, W.: Timed Automata: Semantics, Algorithms and Tools. In: Desel, J., Reisig, W., Rozenberg, G. (eds.) ACPN 2003. LNCS, vol. 3098, pp. 87–124. Springer, Heidelberg (2004)
4. Boucheneb, H., Barkaoui, K.: Relevant Timed Schedules/Clock Vectors for Constructing Time Petri Net Reachability Graphs. Discrete Event Dynamic Systems 21(2), 171–204 (2011)
5. Kim, T.G.: DEVS Formalism for Modeling of Discrete-Event Systems. In: Handbook of Dynamic System Modeling, ch. 6. Chapman & Hall/CRC (2007)
6. Zeigler, B.P., Kim, T.G., Praehofer, H.: Theory of Modeling and Simulation. Academic (2000)

Study on Markov Chain-Based
System Readiness Assessment Method

Qisheng Guo, Guo Li, Liang Li, and Jinghua Song

Department of Equipment Command & Administration,
The Academy of Armored Forces Engineering,
Beijing, China
`liguo5252@sina.com`

Abstract. System readiness level (SRL) is used to measure the maturity of a system technical scenario so as to cover the shortage of Technical Readiness Level (TRL) by adopting Integration Readiness Level (IRL). When SRL is calculated based on TRL and IRL, the maturity of the Critical Technology Elements (CTEs) and their interplay would change with time going on. Thus, when SRL is calculated based on TRL/IRL, it may have time lags and the decision made according to the SRL will be also out of time. It may cost a large quantity of time and money. In order to solve the problem, markov chain is adopted. The concept of initial distribution, transition probability matrix and stationary distribution are used to describe and obtain the initial and stationary TRL/IRL/SRL. The markov chain-based System Readiness Assessment (SRA) method can reduce the time lag of the TRL/IRL/SRL. Thanks to the advantage, decision can be made more accurately and scientifically. An illustrative example is given to test and verify the method. The method is easy to operate and can be generalized to other field.

Keywords: system readiness assessment, markov chain, time lag, risk.

1   Introduction

In order to reduce the technical risks during the system designing, since the 1980's the National Aeronautics and Space Administration (NASA) has used technology readiness level (TRL) as a means to assess the maturity of a particular technology and a scale to compare technologies. In 1999, the Department of Defense (DoD) advanced a similar TRL concept in their programs. The TRL scale is a measure of maturity of an individual technology, with a view towards operational use in a system context. A more comprehensive set of concerns become relevant when this assessment is abstracted from an individual technology to a system context, which may involve interplay between multiple technologies. In 2006, Sauser proposed System Readiness Level (SRL) and System Readiness Assessment (SRA) which incorporates the TRL scale, and introduce the concept of an integration readiness level (IRL) to dynamically calculate a SRL index [1]-[4]. However, when the decision is made, the TRLs and IRLs may change. Thus, the SRL calculated based on TRL and IRL maybe out of

time or not so accurate because of the time lags of TRLs and IRLs. Of course, we can redo the assessment with time going on. But by the time we had redone the assessment, the results maybe out of time again. In order to solve the problem, we are proposing a new method based on markov chain. Thanks to the use of markov chain, the shortcoming of decision making with time lags can be reduced to some extent. Decision can be made more accurately according to both the current SRL and at the stationary SRL. Thus, the technical risks can be reduced.

2   Propaedeutics

The SRL index is an index of maturity applied at the system-level concept with the objective of correlating this indexing to appropriate systems engineering management principals. We contend that the SRL of a given system is a function of individual TRLs and the maturities of the links between them, which will be defined based on a scale of IRLs. The markov chain-based SRA Method mainly involves some basic notions and models such as: TRL, IRL, SRL [1], [5], [6] and markov chain etc. Some of these will be introduced briefly.

2.1   TRA & TRL

A Technical Readiness Assessment (TRA) is a formal, systematic, metrics-based process that assesses the maturity of technologies called Critical Technology Elements (CTEs) to be used in systems. The definition of a CTE is as follows: A technology element is "critical" if the system being acquired depends on this technology element to meet operational requirements (within acceptable cost and schedule limits) and if the technology element or its application is either new or novel or in an area that poses major technological risk during detailed design or demonstration.

2.2   IRA & IRL

Integration Readiness Assessment (IRA) is a systematic measurement of the interfacing of compatible interactions for various technologies and the consistent comparison of the maturity between integration points (TRLs). IRL is proposed using to describe the integration maturity of a developing technology with another technology, developing or mature. The addition of IRLs not only provides a check to where the technology is on an integration readiness scale, but also a direction for improving integration with other technologies. As TRL has been used to assess the risk associated with developing technologies, IRL is designed to assess the risk of integration. As system's complexity increases there must be a reliable method and ontology for integration that allows TRLs to collectively combine for develop these complex systems.

2.3   SRA & SRL

The SRL index is designed to be a function of the individual TRLs in a system and their subsequent integration points with other technologies, IRL. The resulting function of this interaction is then correlated to a five level SRL index.

2.4   Markov Chain

Markov chain model is adopted solve the problem. Some basic notions and solutions of markov chain are to be introduced as below [7]-[9].

1) Finite state homogeneous markov chain

Given a state space S equipped with a σ-field B we call a stochastic process $\{\xi_n, n=0,1,2,\ldots\}$ with each R.V. ξ_n taking values in S a markov chain if for every non-negative integer n and any set $T \in B$, almost surely

$$P\{\xi_{n+1} \in T|\xi_1,\xi_2,\ldots,\xi_n\}=P\{\xi_{n+1} \in T|\xi_n\} \tag{1}$$

When such a stochastic process has a modification of $P\{\xi_{n+1} \in T|\xi_n\}$ that does not depend on n (expected via the value of ξ_n) we say that it is a homogeneous markov chain. When S is finite we say such a markov chain is finite state homogeneous markov chain. Equation (1) is called markov property.

2) Stationary distribution

We call a probability distribution $\{\pi_i,\ i \in S\}$ a stationary distribution of markov chain ξ if

$$\pi_j = \sum_i \pi_i p_{ij}, \quad j \in S \tag{2}$$

Equation set (2) can be written in vector formation: $\boldsymbol{\pi}P = \boldsymbol{\pi}$

Thereinto, $P\{\xi_{k+1}=j|\xi_k=i\}=p_{ij}$, $\boldsymbol{\pi} = (\pi_0,\pi_1,\cdots,\pi_n)$, $P=(p_{ij})_{(n+1)\times(n+1)}$.

2.5   Time Lags

The time lag has two meanings in this paper. On the one hand, the readiness levels have time lags; on the other hand the decision has time lags. Time and space are changing from time to time. Due to the fact, for traditional evaluation method, when the assessment result is obtained, the maturity has changed. Thus, the evaluation result is out of time so it has time lags. The so-called decision with time lags is the decision which is made based upon readiness levels with time lags. The decision will be out of time because the readiness levels are out of time. Thus the decision has time lags 10.

3   Method and Procedure

3.1   Basic Assumption

1) Supposed that the TRLs & IRLs in the $(k+1)^{th}$ readiness assessment is only related to the kth readiness assessment. Thus, the TRL & IRL scale will have markov property according to the definition.

2) The development of a certain technique and the interplay between two or more techniques has inherent laws that can be found.

Under the assumptions above, the theory of finite markov chain can be used to describe and solve the problem of TRA & IRA and finally SRA.

3.2   Method in Detail

To use the markov chain related theory in readiness assessment, there're several woks that to be done.

1) Ascertain the state space of the levels for each CTE

For simple and convenient, the state space S of the readiness level is $\{m_0, m_1, \ldots, m_n\}$, i.e., the n-level scale method is used to describe the state space.

2) Obtain the initial distribution of the TRL/IRL

Do statistical analysis of the kth assessment, the initial distribution is

$P_0 = (p_0, p_1, \ldots p_n) = (P\{\xi(k)=m_0\}, P\{\xi(k)=m_1\}, \ldots, P\{\xi(k)=m_n\})$.

Thereinto, $P\{\xi(k)=m_l\} = n(m_l)/\sum n(m_l)$ $(l=0,1,\ldots,n)$, i.e., $P\{\xi(k)=m_l\}$ is the frequency with which the CTE get the readiness level m_l in the kth assessment.

3) Do statistical analysis and compute to get the transition probability matrix

Count the readiness level changing frequency from state i to state j, make it the transition probability p_{ij}, $P=(p_{ij})_{(n+1)\times(n+1)}$.

After the preparation, the initial and evolutional readiness level can be computed. The readiness level is actually the mathematical expectation of the readiness level.

Initial readiness level: $w_0 = P_0 \cdot (m_0, \ldots, m_n)'$

Evolutional readiness level after r step transitions: $w_r = P_0 \cdot P^r \cdot (m_0, \ldots, m_n)'$

In order to get the stationary readiness level, the stationary distribution should be obtained first.

It can be learned from the definition that the stationary distribution can be obtained by resolving the equation set (3).

$$\begin{cases} \pi P = \pi \\ \pi \cdot \mathbf{1}_{n+1} = 1 \end{cases} \tag{3}$$

$$\rightarrow \begin{cases} \pi(P-I) = \mathbf{0}' \\ \pi \cdot \mathbf{1}_{n+1} = 1 \end{cases} \rightarrow \pi(P-I \quad \mathbf{1}_{n+1}) = (\mathbf{0}' \quad 1) \rightarrow \pi = (\mathbf{0}' \quad 1)(P-I \quad \mathbf{1}_{n+1})^+$$

Thereinto π is stationary distribution.

Thereinto, $\mathbf{1}_{n+1}=(1, \ldots,1)'$; $\mathbf{0}=(0,\ldots,0)'$; $P=(p_{ij})_{(n+1)\times(n+1)}$; $(\ \cdot\ )^+$ means the Moore-Penrose generalized inverse matrix [10] (for its existence and uniqueness) of matrix "$\cdot$".

Stationary readiness level: $w = P_0 \cdot \lim_{r\to\infty} P^r \cdot (m_0 \quad \cdots \quad m_n)'$.

3.3   Detailed Procedure

The procedure of the method is shown in Fig.1.
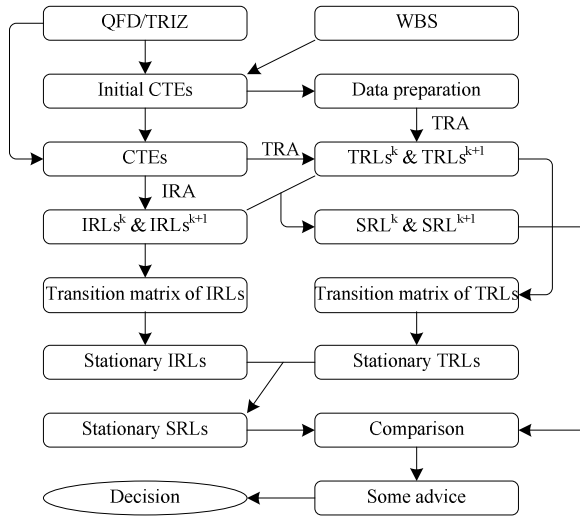
Fig. 1. Flow chart of the method

1) CTEs obtaining

In this step, CTEs should be obtained through Work Breakdown Structure (WBS), Quality Function Deployment (QFD) and Theory of Inventive Problem Solving (TRIZ, Russian initial abbreviation). Through WBS the initial CTEs can be obtained. With the aid of QFD/TRIZ, the weight of each CTE can be calculated. Take the initial CTEs and their weights into consideration, the CTEs are determined.

2) Initial TRL/IRL obtaining & Initial SRL calculation

Invite several assessment groups to do the readiness assessment. Record the data to calculate the initial TRL/IRL/SRL. In order to do the assessment, a symbolism (shown as Fig.2) is designed.
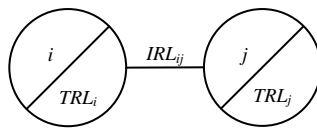
Fig. 2. TRL/IRL relationship representation symbolism

Fig.2 means that the technical readiness of CTE_i (CTE_j) is TRL_i (TRL_j) and the integration readiness level between CTE_i and CTE_j is IRL_{ij}.

If there were n CTEs, $TRL = (TRL_i)_{n \times 1}$, $IRL = (IRL_{ij})_{n \times n}$, $SRL = (SRL)_{1 \times 1}$

then the relationship between TRL, IRL and SRL is shown as formula (4) [2]-[4].

$$SRL = 1/n \cdot (1/l_1, \, 1/l_2, \, \cdots, \, 1/l_n) \cdot [(IRL/7) \cdot (TRL/9)] \tag{4}$$

According to formula (d), the initial TRL/IRL/SRL can be obtained as $\overline{TRL^0}$, $\overline{IRL^0}$ and $\overline{SRL^0}$.

3) The kth readiness assessment

Invite the same assessment groups as possible to do the kth readiness assessment. Mark the data involved as $TRL^k = (TRL_i^k)_{n \times 1}$ and $IRL^k = (IRL_i^k)_{n \times n}$.

4) The $(k+1)^{th}$ readiness assessment

Invite the same assessment groups as possible to do the $(k+1)^{th}$ readiness assessment. Mark the data involved as $TRL^{k+1} = (TRL_i^{k+1})_{n \times 1}$ and $IRL^{k+1} = (IRL_i^{k+1})_{n \times n}$.

5) Transition probability matrix obtaining

By counting the frequency of the transition from one level to another, the transition probability matrix can be obtained.

6) Stationary distribution obtaining

Through complex calculation, the Stationary distribution and the stationary of TRLs and IRLs can be obtained by solving equation set (3).

7) Stationary TRL/IRL & SRL calculation

According to formula (4), the stationary SRL can be obtained.

8) Comparison to support the decision

Compare the initial SRL and the stationary SRL; some advice can be obtained to support the decision making.

3.4   Comparison with the Classical Method

Compared with the classical method, the markov chain-based SRA mainly has advantage in two aspects.

1) Readiness level obtaining aspect

The markov chain-based method gets the readiness level computed with the initial distribution, transition probability matrix, and stationary distribution, while other methods with only simple statistical analysis. The markov chain-based method gets three formations (initial formation, evolutional formation, stationary formation) of the readiness level, while others only one.

2) Time lags solving aspect

The markov chain-based method solves the time-lag problem to some extent by using transition probability matrix and stationary distribution; while others often can do nothing with the time lags.

Comprehensively, thanks to the markov chain-based assessment method, decision would be made more accurately and the risks during system design would be reduced to some extent.

4   Illustrative Example

To test and verify the markov chain-based method advanced in this paper, an illustrative example is given as below.

1) Four CTEs are obtained through WBS/QFD/TRIZ. The results of the kth assessment (involving four assessment groups) is shown as Fig.3 and the $(k+1)^{th}$ assessment (involving four assessment groups) Fig.4.
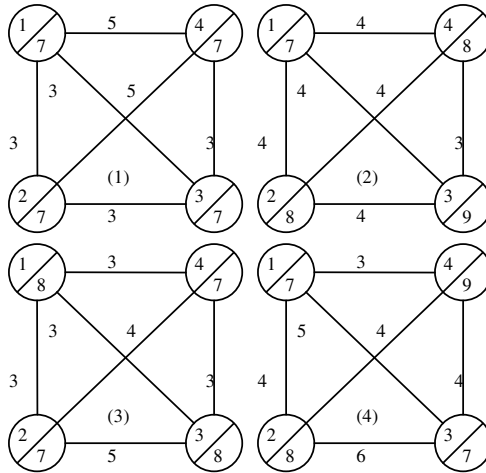
Fig. 3. The kth assessment
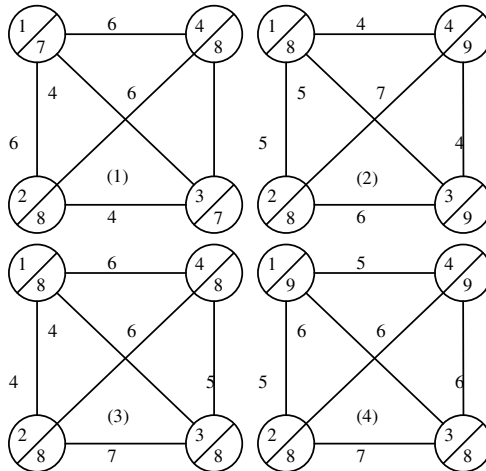
Fig. 4. The $(k+1)^{th}$ assessment

The information shown in Fig.3 can be summarized in Table 1 and Table 2. The information shown in Fig.4 can be summarized in Table 3 and Table 4.

Table 1. TRLs in the kth assessment

Group CTE	1	2	3	4
1	7	7	8	7
2	7	8	7	8
3	7	9	8	7
4	7	8	7	9

Table 2. IRLs in the kth assessment

Group	1				2				3				4			
CTE	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	7	3	3	5	7	4	4	4	7	3	3	3	7	4	5	3
2	3	7	3	5	4	7	4	4	3	7	5	4	4	7	6	4
3	3	3	7	3	4	4	7	3	3	5	7	3	5	6	7	4
4	5	5	3	7	4	4	3	7	3	4	3	7	3	4	4	7

Table 3. TRLs in the $(k+1)^{th}$ assessment

Group CTE	1	2	3	4
1	7	8	8	9
2	8	8	8	8
3	7	9	8	8
4	8	9	8	9

Table 4. IRLs in the $(k+1)^{th}$ assessment

Group	1				2				3				4			
CTE	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	7	6	4	6	7	5	5	4	7	4	4	6	7	5	6	5
2	6	7	4	6	5	7	6	7	4	7	7	6	5	7	7	6
3	4	4	7	5	5	6	7	4	4	7	7	5	6	7	7	6
4	6	6	5	7	4	7	4	7	6	6	5	7	5	6	6	7

2) Use markov chain-related theory to do the data analysis. Some important results are shown as below.

Initial State: $\overline{SRL^0} = 995/1792 = 0.5552$;

$$\overline{TRL^0} = \begin{pmatrix} 29/4 & 30/4 & 31/4 & 31/4 \end{pmatrix}';$$

$$\overline{IRL^0} = \begin{pmatrix} 7 & 14/4 & 15/4 & 18/4 \\ 14/4 & 7 & 15/4 & 21/4 \\ 15/4 & 15/4 & 7 & 13/4 \\ 18/4 & 21/4 & 13/4 & 7 \end{pmatrix};$$

Stationary State: $TRL^S = \begin{pmatrix} 9 & 9 & 9 & 9 \end{pmatrix}';$

$$IRL^S = \begin{pmatrix} 7 & 3969/640 & 1987/320 & 1987/320 \\ 3969/640 & 7 & 7889/1280 & 797/128 \\ 1987/320 & 7889/1280 & 7 & 7927/1280 \\ 1987/320 & 797/128 & 7927/1280 & 7 \end{pmatrix}.$$

$SRL^S = 1868/2043 = 0.9143$.

At the initial state, $TRL_1 < TRL_2 < TRL_3 = TRL_4$ and $SRL = 0.5557$; at the stationary state, $TRL_1 = TRL_2 = TRL_3 = TRL_4$ and $SRL = 0.9143$. Through the results, it is can be seen that if decision were made according to the initial readiness level, the technical scenario may be rejected because its SRL is too low (just SRL-2); but if decision were made according to the stationary readiness level, the technical scenario should be adopted because its SRL is very high (SRL-5). In other words, if a feasible program were rejected, what a pity it will be.

Through the example, we've known that the method put forward is reasonable and correct from theoretical aspect and meaningful from application aspect.

5   Conclusion and Prospect

The paper puts forward a novel method: markov chain-based SRA. The markov chain-based method can not only give the initial TRLs/IRLs, but also can give the evolutional TRLs/IRLs and stationary TRLs/IRLs. Thus, the SRLs calculated based on initial and evolutional TRLs/IRLs can keep pace with time. Thanks to the advantage, the new method can reduce the time lags and enhance the accuracy of the decision to some extent. It's easy to understand and can be generalized to other fields.

References

1. Sauser, B., Ramirez-Marquez, J., Verma, D., et al.: From TRL to SRL: The Concept of Systems Readiness Levels. In: Conference on Systems Engineering Research, Los Angeles, CA, April 7-8 (2006)
2. Sauser, B., Ramirez-Marquez, J., Verma, D., et al.: A System Maturity Index for Decision Support in Life Cycle Aquisition. In: AFRL Technology Maturity Conference, Virginia, VA, September 11-13 (2007)
3. Sauser, B., Ramirez-Marquez, J., Magnaye, R., et al.: A Systems Approach to Expanding the Technology Readiness Level within Defense Acquisition. International Journal of Defense Acquisition Management (01), 39–58 (2008)
4. Sauser, B., Ramirez-Marquez, J., Magnaye, R., et al.: System Maturity Indices for Decision Support in the Defense Aquisition Process. In: The 5th Annual Acquisition Research Symposium, Monterey, CA, May 14-15 (2008)
5. U.S. Department of Energy Office of Environmental Management. Technology Readiness Assessment (TRA)/Technology Maturation Plan (TMP) Process Guide (2008)
6. Department of Defense. Technology Readiness Assessment (TRA) Deskbook (2009)
7. Wang, Z.-K.: An Introduction to Probability Theory and Its Applications, 3rd edn., ch. 4. Beijing Normal University, Beijing (2007)
8. Loève, M.: Probability Theory II, pp. 263–281. Springer, New York (1978)
9. Feller, W.: An Introduction to Probability Theory and Its Applications, 3rd edn., pp. 392–399. John-Wiley & Sons, New York (1957)
10. Li, L., Guo, Q.-S., Yang, X.-Y.: Evaluation Method based on Markov Chain Model. In: 2008 Asia Simulation Conference - 7th Intl. Conf. on Sys. Simulation and Scientific Computing, Beijing (2008)
11. Wang, S.-G., Yang, Z.-H.: Generalized Inverse Matrix and Its Applications, ch. 4. Beijing Industry University, Beijing (1996)

Pre-motion Based VR Object Manipulation
Definitions and Preliminary Experiments

Shiori Mizuno[1], Asuka Komeda[1], Naoko Yoshii[1], Tomoko Yonezawa[2],
Masami Takata[1], and Kazuki Joe[1]

[1] Nara Women's University, Nara, Nara, Japan
[2] Kansai University, Takatsuki, Osaka, Japan

Abstract. In this paper, we propose a new method to manipulate objects by using user's pre-motion in a VR environment. Users are provided with an instinctually easy interface to VR with using their natural behaviors as their commands for manipulating VR objects. Hand pre-shaping is known as human's unconscious behavior to face a small object to be grasped. The definitions of pre-motions are determined by users' initial behaviors. Therefore, we present the definitions to classify characteristics of the kinds of objects and pre-motions.Using the above definitions, we develop a prototype system to validate the classifications. Consequently, manipulating VR objects by pre-motion is possible and promising.

Keywords: VR, object manipulation, pre-motion, Kinect.

1   Introduction

The VR is used in various fields such as architectures and preservation of cultural properties. In this way, the convenience of VR comes to the front while it is sometimes difficult for average users to manipulate objects in VR. Such users do not have enough knowledge of information technology, and some of them have a great deal of resistance to the manipulation by IT devices. Therefore, we need a method to manipulate VR objects easily and intuitively.

In general, keyboards, mouses, wands[1] and data gloves[2] are used as VR interface devices to manipulate objects. Using the devices, there are two problems for users to manipulate objects in VR. First, pre-learning for VR object manipulation is needed since commands to instruct the movement have been determined in each VR system in most cases. The imposition of available commands for VR object manipulation sometimes turns potential users off. The second problem is that the motion to issue a command is different from the natural motion to manipulate the object in VR. When a person wants to look at the details of a real object, he/she gets a close look. On the other hand, in the case of VR, he/she must issue a zooming command by clicking a button or drawing a mouse that is pre-defined for the VR system. In this way, there is big difference between human's natural behavior and operations for digital devices. The difference is too large to intuitively manipulate objects in VR. Data gloves are well known as a VR interface device, and used for capturing the motion of fingers to issue

a command in object manipulation. The motion is very close to human's actual behavior what we are interested in. However, as the VR interface device, the data gloves have some restriction. Because they are worn in hands, they just get the motion of figures and cannot capture larger parts of human's body. For intuitively manipulable VR systems, we need to solve the above mentioned problems: the pre-learning for manipulations, the difficult manipulation commands for some users, and limited parts of human's motion.

In this paper, we propose a new method to intuitively and easily manipulate objects in VR. The purpose of the method is to give common users an instinctually simple interface to a VR system with using their natural behaviors as their commands when they want to manipulate a VR object. To capture the natural behaviors as the interface to VR, we present the definitions to classify the kinds of objects in VR and the motions with which users may perform their behaviors to face the classified objects.

We focus the initial motions of users to face an object in VR. The initial motions includes "hand pre-shaping"[3]. Hand pre-shaping is known as human's unconscious behavior to face a small object to be grasped[4],[5],[6]. When a human looks at a small object and is going to grasp it, he/she makes the shape of his/her hand according to the shape of the object without thinking in advance of touching it. This unconscious behavior shows that we calculate the shape of our hands for the object to be grasped just from the visual information of the object. When we face another object to be grasped, our hand pre-shaping forms other shapes according to the new object. In this way, we unconsciously prepare the motion in advance of the actual behavior to the facing object according to the purpose of the behavior and the property of the object. In this paper, we call "preparation of motion" pre-motion. The pre-motion differs according to the purpose of behaviors and the property of objects. The hand pre-shaping is a typical example of the pre-motion in our definition. Given the definition of pre-motion, users' will can be guessed by the pre-motion.

In this paper, the characteristics of pre-motions to a facing object with various sizes, hardness and weights, are classified with the purpose and the property. To validate the pertinence of the classifications, preliminary experiments are performed. In the experiments, the pre-motions in these classifications to primitive objects are recognized by Kinect[7],[8], which captures human's body motions.

The rest of the paper is organized as follows. In Sect. 2, we present the definitions to classify the kinds of objects and pre-motions. In Sect. 3, we validate the effect of the definitions by experiments.

2   Definition of Classification

To manipulate VR objects by using natural motions, the characteristics of pre-motion should be carefully analyzed to detect where and how the feature of the pre-motion for a specific object appears. In addition, we need to classify the attribute of objects that provokes different pre-motions. In this paper, we classify the motions of grasping objects. In this section, we present the definition to classify objects and motions.

2.1   Definition of Object Classification by Feature

Aim at the classification of VR object attributions that trigger different pre-motions, we investigate the classification of VR object features. We focus on the following five features: shape, location, size, weight, and hardness.

I) Shape. We assume that human's pre-motions are similar for the target objects with a similar shape. Based on the assumption, we need to classify pre-motions by object shape. Shapes are classified into several primitives: cube, cuboid, sphere, cylinder, triangular pyramid, circular cone, etc. For example, a cardboard box, a microwave and a printer have a cuboid shape. A basketball and an orange have a sphere shape. A can and a tumbler have a cylinder shape. Figure 1 (a-1) and (a-2) shows an example of holding two kinds of objects with a cuboid shape. It illustrates that human's holding motions are similar when the target objects are classified in a shape with the same primitive.

II) Location. The direction of reaching out for an object varies according to the location where the object is placed as shown in Fig.1 (b-1) and (b-2). Figure 1 (b-1) shows that a cardboard box is located at shoulder height to be held by both hands while Fig.1 (b-2) shows that the cardboard box is located on the ground to be held up by both hands. From the figures, it turns out that the location of an object plays an important role in understanding user's will for the object in advance because different object locations cause the user to reach out for the objects in different directions.

III) Size. When we are going to hold an object, we unconsciously decide whether we reach out a hand or both hands for the object in advance. So we classify objects by size in two categories: hand size objects to be grasped by a hand and larger object to be held by both hands. Figure 1 (c-1) and (c-2) illustrates that human's motion changes regarding to the size of the target object. In Fig.1 (c-1) and (c-2), (c-1) and (c-2) present holding an object by both hands and grasping an object by a hand, respectively. From the figures, it turns out that a single hand is used for grasping an object with at most hand-size while both hands are used for holding an object with larger size than his/her hand. So the object size dominates human's motions of holding and grasping, and the object size should be classified.

IV) Weight. Lifting an object by hand, we judge whether we use one hand or both hands by the physical appearance weight of the object. In Fig.1 (d-1) and (d-2), since the target object is not so large, the examinee judges one-hand lifting (d-1), but it turns out that the object is too heavy to lift by one-hand. Then she lifts the object by both hands (d-2). From the figures, it turns out that the weight of the target object affects the motion of the examinee. So the weight of objects should be classified in two categories: heavy (requires both hands to hold) and light (requires one hand to grasp).

V) Hardness. When the size of the target objects are too large to grasp by one hand as described in Sec. 2.1-Ⅲ), some objects are graspable by one hand as shown in Fig.1 (e-1) and (e-2). In this case, the hardness of the target objects is different: soft (e-1) and hard (e-2). The motion of Fig.1 (e-1) is possible when the target is enough soft that a part of the object can be deformed to be grasped by one hand and the object is enough light for one-hand grasping. Although the object of Fig.1 (e-2) is as large as (e-1), it must be hold by both hands because it is very hard. From the figures, it turns out that some objects are soft and deformable so that they are grasped by one hand. So the hardness of objects should be classified in two categories: hard (requires both hands to hold) and soft (requires one hand to grasp).

(a-1)   (a-2)   (b-1)   (b-2)   (c-1)   (c-2)

(d-1)   (d-1)   (e-1)   (e-1)
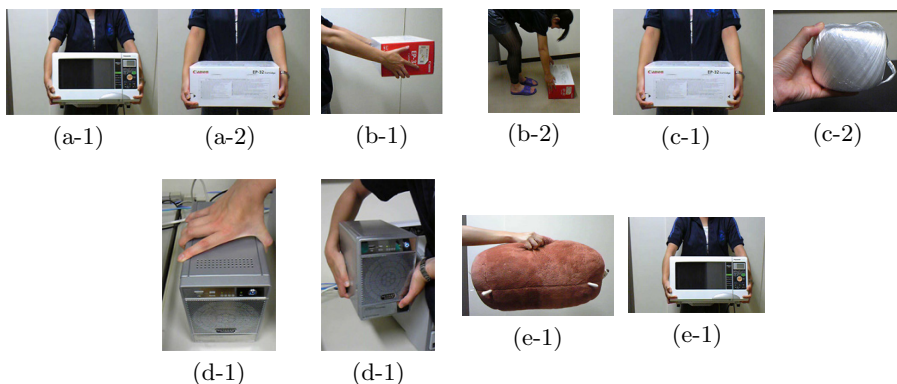
Fig. 1. Different motions by characteristics of object. (a-1)microwave (a-2)cardboard box (b-1)Midair (b-2)On the ground (c-1)Cardboad box (c-2)Tie down rope (d-1)one hand to grasp (d-2)requires both hands to hold (e-1)Soft (e-2)Hard.

2.2   Definition of Motion Classification

The motion of holding or grasping includes the pre-motion of reaching out a hand or both hands to the target object. When reaching out one's hand(s), there are two possible motions of touching and lifting. We believe that the difference between the two motions is predictable by analyzing the shape of the fingers, the number of the hands, and the way of opening one's hands regarding to the object features. In this subsection, we present a method to classify pre-motions when reaching out to the target object.

i) The number of hands. When reaching out to an object, we judge whether we use one hand or both hands regarding to the size, the weight, and the hardness of the object. In the case of one hand, when the object size is small and the object weight is light, we predict that the object is to be grasped or touched. Even when the object is large, it could be grasped, provided that it is light and very soft. Otherwise we predict it is to be touched. In the case of both hands, we reach out to the object with opening the hands when the object size is large. Even when the object is small, we use the both hands to hold a heavy object.

ii) The shape of fingers. When the target object is enough small and light to be grasped by a hand, the motion of reaching out may be also a touching. To predict whether the motion is for grasping or touching, we classify the shape of fingers. Figure 2 (a-1), (a-2), (a-3) and (a-4) shows that different purposes of reaching out a hand lead to different shapes of fingers. The shapes of fingers, exactly saying the shapes formed by the thumb and the other fingers, in Fig.2 (a-1) and (a-2) are sweeping while the shapes of fingers in (a-3) and (a-4) are with sharp and right angles. Since the target object is hand-size, the distance between the thumb and the other fingers in (a-1) and (a-3) is larger than the height of the object while the distance in (a-2) and (a-4) is shorter. To grasp the object, since it is taken between the thumb and other fingers with conforming the palm, we understand (a-1) represents a grasping motion. In short, the pre-motion for grasping includes a curve formed by the fingers and a longer distance between the fingertips than the object height (or width). In this paper, the hand shape for the grasping motion as shown in Fig.2 (a-1) is defined to be sweeping while the other hand shapes are defined to be rectilinear.

iii) Distance between two hands. In this paper, the horizontal, the vertical, and the depth directions to the VR screen are defined as x-, y- and z-axis, respectively. The object width along the x-axis and y-axis are defined as sideways and longitudinal width, respectively. Figure 2 (b-1) and 2 (b-2) show the object is hold by both hands. To hold an object by two hands, the distance between two hands is prepared to be longer than the sideways or the longitudinal width of the object in advance of actual holding. On the other hand, when the distance is shorter, the motion is predicted as touching. In this way, the distance between two hands should be classified so that the change of holding and touching motion is predictable.

(a-1)   (a-2)   (a-3)   (a-4)   (b-1)   (b-2)
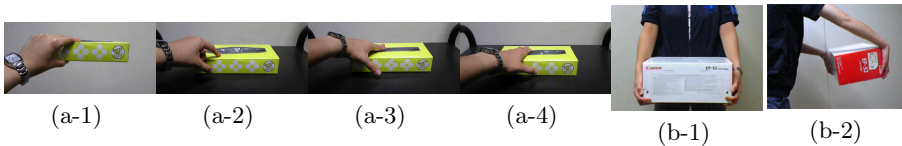
Fig. 2. Different motion by (a)finger shape and (b)distance between two hands. (a-1)Grasping (a-2)Pointing out (a-3)Touching (a-4)Putting on (b-1)x-axis (b-2)z-axis.

The classification of pre-motions for a target object we discuss is summarized.

When reaching out a hand, the motion to the target object is predicted based on the classification of finger shapes as shown in 2.2-ii): rectilinear for touching and sweeping for grasping. When reaching out both hands, there are possible three patterns as shown in 2.2-iii). When the distance between two hands is longer than the sideways or longitudinal width of the target object, the motion is predicted as lifting. Otherwise, it is predicted as touching. The classification of pre-motions for a target object is applied by object, and the motion is predicted by the purpose to the object.

3   Preliminary Experiment

For the pertinence evaluation of the classification we define in Sect. 2, we perform preliminary experiments using a prototype system provided that we focus a limited target with corresponding motions. The target object we use in this experiment is a cuboid. In this section, we explain the prototype system we developed for the preliminary experiment. We also show the result of the experiment and discuss the pertinence of our classification.

The preliminary experiment requires motion capture ability. So we use Kinect as a sensor device to capture user's motions.

3.1   Prototype System

The operation flow of the prototype is described. In Step 1, the prototype is activated. Kinect starts up and a pre-defined 3D object is displayed so that the information of user's motions detected by Kinect and the 3D object are related. In Step 2, user is detected. Note that the prototype does not support multiple users. In Step 3, the recognized user holds a pre-defined pose so that the prototype performs initial calibration. Namely individual difference of the pose must be corrected by calibration. In Step 4, user's motion is tracked with following captured information: joint positions. In Step 5, the coordinates of user's articulations are obtained based on the classification given in Sect. 2.2. In Step 6, using the coordinates and their tracking information, user's motion is recognized as a pre-motion based on the classification given in Sect. 2.1. In Step 7, the coordinates of a 3D object are calculated so that the 3D object is moved according to user's pre-motion. The resultant 3D object is displayed in Step 8. The movement of the 3D object is executed when user's pre-motion is for moving objects. On the other hand, the 3D object is not lifted when it is for touching. In Step 9, when detected user's pre-motion is recognized as the wave motion, which is pre-defined by Kinect, the prototype is terminated. Otherwise go back to Step 5 and continue to take user's pre-motion.

3.2   Experiment Result and Discussions

The purpose of human's motion provided that human's pre-motion is classified with our definition for the case of a cuboid target object. Since the cuboid is too large to be hold with one hand, both hands are reached out to hold the cuboid. When reaching out one hand, the pre-motion is recognized as a touching motion. In addition to the above two motions, pushing is another possible motion to cuboids. In this case, it is observed that user reaches out one hand or both hands toward a side of the cuboid so that he/she pushes the cuboid to the reaching out direction.

Figure 3 shows that a user reaches out a hand for some motion in a VR system when a cuboid is located on the ground. Figure 3 (a-1) illustrates user's pre-motion which is predicted as touching the cuboid. Figure 3 (a-2) shows user's motion is pushing, which is seen as an extension of the pre-motion touching, and the cuboid is moved by user's motion. At this time, the VR system knows the cuboid should be moved toward user's reaching out direction but not be lifted up.

(a-1)   (a-2)

Fig. 3. Reaching out a hand for (a-1)Touching and (a-2)Pushing.

In Fig. 4 (a-1) and (a-2), the user reaches out both hands toward the cuboid to lift it. Fig.4 (a-1) shows user's pre-motion of reaching out both hands, and it is not sure if the pre-motion is for touching or lifting the cuboid at this time. However, the VR system observes that the distance between user's two hands is longer than the sideways width of the cuboid, so the VR system predicts that user's pre-motion is for lifting the cuboid. The motion of the user is after all for lifting and the cuboid is lefted as shown in Fig.4 (a-2).

Figure 4 (b-1) and (b-2) shows the user reaches out her both hands with a different angle rather than in Fig.4 (a-1) and (a-2). This time, user's pre-motion shown in Fig.4 (b-1) is more understandable because her both hands are reached out with the hands angle of the longitudinal direction of the cuboid. It means that the user is going to lift the cuboid with holding its longitudinal as shown in Fig.4 (b-2).

(a-1)   (a-2)

(b-1)   (b-2)

Fig. 4. Reaching out both hands in (a)x-axis and (b)z-axis. (a-1)Touching (a-2)Lifting (b-1)Touching (b-2)Lifting.

Figure 5 (a-1), (a-2) and (a-3) also shows the user reaches out her both hands. In this case, user's two hands are set to the sideway, not longitudinal, direction of the cuboid as shown in Fig.5 (a-1). The distance between two hands is shorter than the sideway width of the cuboid, and the user reaches out her hands toward a side of the cuboid. For these reasons, the VR system predicts that the user pre-motion is for touching or pushing rather than lifting. Therefore, the cuboid moves by the user motion of pushing as shown in Fig.5 (a-2), and it is not lifted by the user motion of lifting as shown in Fig.5 (a-3).

Figure 5 (b-1), (b-2) and (b-3) shows the user reaches out her both hands with a different angle rather than in Fig.5 (a-1), (a-2) and (a-3). This motion is understood by the VR system as in Fig. 5 (a-1), (a-2) and (a-3). In this case, user's two hands are set to the longitudinal, not sideway, direction of the cuboid as shown in Fig.5 (b-1). The distance between two hands is shorter than the longitudinal width of the cuboid, and the user reaches out her hands toward a side of the cuboid. For these reasons, the VR system predicts that the user pre-motion is for touching or pushing rather than lifting. Therefore, the cuboid moves by the user motion of pushing as shown in Fig.5 (b-2), and it is not lifted by the user motion of lifting as shown in Fig.5 (b-3).

(a-1)	(a-2)	(a-3)
(b-1)	(b-2)	(b-3)
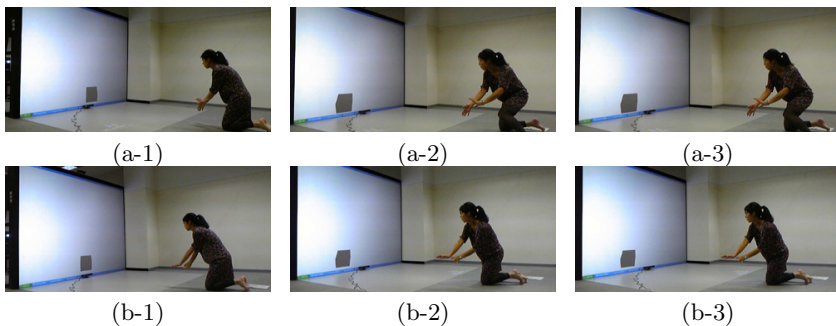
Fig. 5. Reaching out both hands in (a)x-axis and (b)z-axis. (a-1)Touching (a-2)Pushing (a-3)Lifting (b-1)Touching (b-2)Pushing (b-3)Lifting.

As the results of the experiment, we conclude that the prototype system based on the definition of definition works well. It is observed that user's pre-motion deeply dominates its following motion to the object in the VR. It means that user's request to VR objects is predictable and understandable based on the definition shown in Sect. 2.2. Therefore we show that it is possible to take in human's natural motion, as we here propose human's pre-motion, for manipulating VR objects.

4   Conclusions

VR is one of the most effective presentation methods for giving any people various information that is difficult to show such as world heritages or national

treasures. The most critical problem to use such VR is that typical VR systems require pre-defined control commands for presented 3D objects. In short, it is difficult for usual people, who are not familiar with VR operations, to manipulate 3D objects using such special commands in VR. In this paper, we proposed the definition to classify the relation between user's pre-motion and his/her target 3D object in VR. The purpose of the definition is to allow any user to intuitively and naturally manipulate 3D objects in VR. In such a VR system, user can manipulate 3D object without learning controlling methods in advance.

To get user's natural behavior as commands, definitions of classifications in user's motions are constructed based on user's initial, here we say "pre-", motions. So we classify objects and user's motions to the objects. To validate the pertinence of the classification, a prototype system with Kinect has been developed, and we performed some experiments with the prototype. In this experiment, a cuboid is displayed in the prototype VR system to be manipulated by an examinee. As the results, the examinee's pre-motions, reaching out one/both hand(s), are recognized by the prototype based on the classification definitions. We conclude that it is possible for ordinal people to intuitively and naturally manipulate 3D objects in VR.

Our future work includes smarter classifications for more complicated objects and various pre-motions.

References

[1] Deering, M.F.: The HoloSketch VR Sketching System. Communications of the ACM 39, 54–61 (1996)

[2] Greenleaf, W.J.: Developing the tools for practical VR applications. IEEE Engineering in Medicine and Biology Magazine 15, 23–30 (1996)

[3] Bard, C., Troccaz, J., Vercelli, G.: Shape Analysis and Hand Preshaping for Grasping. In: IEEE Workshop on IROS 1991, vol. 1, pp. 64–69 (1991)

[4] Jeannerod, M.: The timing of natural prehension movements. Journal of Motor Behavior 16, 235–254 (1984)

[5] Arbib, M.A., Iberall, T., Lyons, D.: Coordinated control programs for movements of the hand. Experimental Brain Research 10, 111–129 (1985)

[6] Miller, A.T., Knoop, S., Christensen, H.I., Allen, P.K.: Automatic grasp planning using shape primitives. In: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA 2003, vol. 2, pp. 1824–1829 (2003)

[7] XBOX (2012), `http://www.xbox.com/en-US/Kinect?xr=shellnav`

[8] XBOX (2012), `http://www.xbox.com/en-US`

Modeling and Verification of Warehouse Dynamic Scheduling Based on the IOQ Parameter of the Product

Wenqiang Yang and Minrui Fei

Shanghai Key Laboratory of Power Station Automation Technology,
School of Mechatronics Engineering and Automation,
Shanghai University,
Shanghai, China
yangwqjsj@163.com

Abstract. With the increasing competition of market economies, many companies are pursuing higher levels of production automation in manufacturing industry. For example, the automated warehouses are employed in the field of manufacturing and processing field, in the process of which automated warehouses play a more and more significant role. Therefore, it is meaningful to have a research on the automated warehouses scheduling issue. The warehouse scheduling algorithm is studied combining with the project on the automatic production line of an enterprise in this paper, and a warehouse scheduling optimization algorithm is proposed based on IOQ(Index of Quality) parameters. Then the process of getting the value of IOQ is also simplified by applying the idea of sparse matrix. In addition, the algorithm uses the maximum of the IOQs to schedule warehouse on line, and is compared with other warehouse scheduling algorithms. The simulation results show that the warehouse scheduling algorithm can not only improve the quality of the product effectively, but also improve the efficiency of the scheduling largely. The desired result is achieved in the end.

Keywords: the algorithm of IOQ, the optimization on line, warehouse scheduling, sparse matrix.

1   Introduction

As the storage period of the products of the enterprise is required strictly for a certain time in the process of production, it can be classified to the production model of the JIT(Just In Time). This requires a reasonable scheduling for the semi-finished products in the warehouse in order not to affect its quality. Thus, it can achieve the purpose of cutting down the cost.

According to the overview of the related literature, the type of scheduling can be summarized into two categories [1]. One is the stochastic scheduling, which has no model. That is to say it chooses an object random, and there is no optimization at all. Therefore there is little research value in it. The other scheduling is based on model, according to which finds the optimal solution by using the optimization algorithm.

The characteristics of the optimal solution generally show the shortest path, the least time, the highest utilization of equipments and so on. Combined with the research contents in this paper, there has been a large amount of literature on the scheduling of the workshop warehouse up to now [2,3,4,5,6]. However, some just use the genetic algorithms, the particle swarm algorithm, the niche algorithm, neural network, all of which belong to intelligent algorithms, to optimize the path. It is about getting out of and going into storage of the products. There is still little precedent research about the strict control of the storage time of product in the warehouse at present. At the same time, it may appear many task requests which are about getting out of warehouse at a time in view of the randomness of the production process [7,8,9]. Thus, how the semi-finished products are scheduled reasonably without affecting the quality gives us a great challenge. Because the traditional static scheduling methods have been unable to meet such needs, a warehouse scheduling system is modeled based on the weight timeout value of the products which marks the influence degree of the product quality.

2   The Description of Warehouse Scheduling Problem and Structure of Model

There exists some semi-finished products which have different processes on automation production line, thus they each have different storage time in the warehouse. At the same time, the random of the production is regarded. Therefore, it may cause that those semi-finished products which come out of warehouse at a time compete for the stacker, and then the storage time of the semi-finished products is longer than the normal time to some degree. This problem is even more serious in the case that there are a large number of semi-finished products. In order to the problem quantitatively, the definitions are done as follows.

Definition 1. The unit timeout value which affects the degree of the product quality is represented as weight, depending on the different storage time of product in the warehouse.

Namely,

$$w = \begin{cases} \dfrac{1}{2} & \text{storage time is two days} \\ \dfrac{1}{3} & \text{storage time is four days} \\ \dfrac{1}{5} & \text{storage time is five days} \\ \dfrac{1}{8} & \text{storage time is ten days} \end{cases} \tag{1}$$

Definition 2. According to the locations of the semi-finished products and their storage time in the warehouse , a buffer is opened up in the RAM, in which the storage time matrix and the weight matrix are generated dynamically corresponding to the semi-finished product in the warehouse.

Namely,

$$T_{m \times n} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix} \qquad (2)$$

$$W_{m \times n} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix} \qquad (3)$$

Definition 3. The past time matrix indicates the time pass by the semi-finished products since they have been stored in the warehouse.

Namely,

$$OT_{m \times n} = \begin{bmatrix} ot_{11} & ot_{12} & \cdots & ot_{1n} \\ ot_{21} & ot_{22} & \cdots & ot_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ ot_{m1} & ot_{m2} & \cdots & ot_{mn} \end{bmatrix} \qquad (4)$$

The scheduling model is found as follows [10, 11, 12] .

$$IOQ_i = \max((ot_i - t_i) * w_i)$$
$$i \in (1, 2, 3, \cdots, m*n) \qquad (5)$$

The scheduling will be carried out in descending order of IOQ_i, the scheduling sequence is showed as follows at this moment.

$$IOQ_{i1} IOQ_{j1} IOQ_{k1} \cdots IOQ_{z1} \qquad (6)$$

Where, $i1, j1, k1, \cdots z1 \in (1, 2, 3, \cdots, m*n)$

As it always takes the stacker a certain amount of time to carry the semi-finished product out of the warehouse, this scheduling sequence(6) shows a local optimum rather than a global optimum, which is proved as follows.

Supposing

$$IOQ_i > IOQ_j \qquad (7)$$

Namely,

$$(ot_i - t_i) * w_i > (ot_j - t_j) * w_j \tag{8}$$

Then the time passes Δt,

$$IOQ_i = ((ot_i - t_i) + \Delta t) * w_i \tag{9}$$

$$IOQ_j = ((ot_j - t_j) + \Delta t) * w_j \tag{10}$$

the size between IOQ_i and IOQ_j at this moment.

$$\begin{cases} IOQ_i > IOQ_j, & \text{when } w_i > w_j \\ IOQ_i > or < or = IOQ_j, & \text{when } w_i < w_j \end{cases} \tag{11}$$

Therefore, scheduling sequence (6) is not a global optimum.

It is proved above that this scheduling model is a time-varying one. Therefore, the scheduling sequence needs to be updated dynamically all the time, so as to achieve a global optimum.

3   The Utilization of Sparse Matrix in Warehouse Scheduling Algorithm

Supposing

$$TT_{m \times n} = OT_{m \times n} - T_{m \times n} \tag{12}$$

In this paper, it is required to find out the semi-finished product in the warehouse on the basis of the objective function. If the m*n elements which are gone through in a matrix are calculated and sorted, the number of comparisons is expressed as follows by the recursive method.

$$T(m*n) = \begin{cases} 0 & m*n = 1 \\ 2T(\dfrac{m*n}{2}) + ((m*n) - 1) & m*n \neq 1 \end{cases} \tag{13}$$

The time complexity can be expressed as $O((m*n)\log(m*n))$ from the above formula.

The above method has little effect on the search efficiency when the size of a warehouse is small. Otherwise, it will have a huge influence on the search efficiency. In view of this situation, the rules are defined as follows.

Given a situation that the number of the semi-finished products whose storage time expire is a small part of the whole product in the warehouse in a short period of time, based on which the matrix $TT_{m \times n}$ is transformed as follows.

If there are some positions in which there are no semi-finished products or the storage time of semi-finished products don't expire or semi-finished products has been placed the scheduling sequence, the elements of the matrix $TT_{m \times n}$ are assigned to zero corresponding to such positions.

The number of zero elements is much larger than the number of non-zero elements by means of such transformation in the matrix $TT_{m \times n}$, so it becomes a sparse matrix, in which only non-zero elements of the matrix are calculated. Then it will improve the calculation and scheduling efficiency greatly.

4   The Design of Scheduling Algorithm Based on IOQ Parameter

The principle of the scheduling algorithm based on the IOQ parameter is that the IOQs are computed for several semi-finished products whose storage time has expired and the one of whose IOQ is the largest is scheduling first. With an eye to the different weights and the changing timeout value for such semi-finished products, the IOQs of the rest semi-finished products of the warehouse are also vary. It is needed to calculate the IOQs to achieve the optimal scheduling on line all the time. The specific scheduling steps are as follows.

i The mapping table 1 is maintained in Ram according to the size of the warehouse in the actual production. The elements of the table1 are structure variables which are associated with the semi-finished product. Those variables include the following members.

① The warehouse position number of the semi-finished product.
② The storage time of the semi-finished product.
③ The weight of the semi-finished product corresponding to the current storage time.
④ The start time of the semi-finished product in the warehouse.
⑤ The flag which marks whether the semi-finished product has been in the scheduling sequence.

ii The $T_{m \times n}$, $W_{m \times n}$ and $OT_{m \times n}$ are generated from the mapping table1 and the formula (2), (3), (4) at the current time. If the elements of $W_{m \times n}$ are all zero, then goes to v.

iii The $TT_{m \times n}$ is worked out by the formula (12), meanwhile, the matrix $TT_{m \times n}$ is transformed to the sparse matrix by the rules in this paper, then the Table 2 is generated whose elements correspond to the non-zero in the sparse matrix $TT_{m \times n}$.

iv The IOQs of the semi-finished products are calculated in the Table2 according to formula (5) and the maximum value among IOQs is placed on the tail of the scheduling sequence, and the flag of the semi-finished product, which identify whether they have been in the scheduling sequence, is set as the number 1.

v Under field conditions, the time which the stacker spends to carry the semi-finished product out of the warehouse obeys the exponential distribution whose parameter is 1.5 minutes, 1.5 minutes passes, then go to i. The following semi-finished product in the scheduling sequence is obtained.

The flow diagram of the IOQ algorithm is shown in Figure 1.
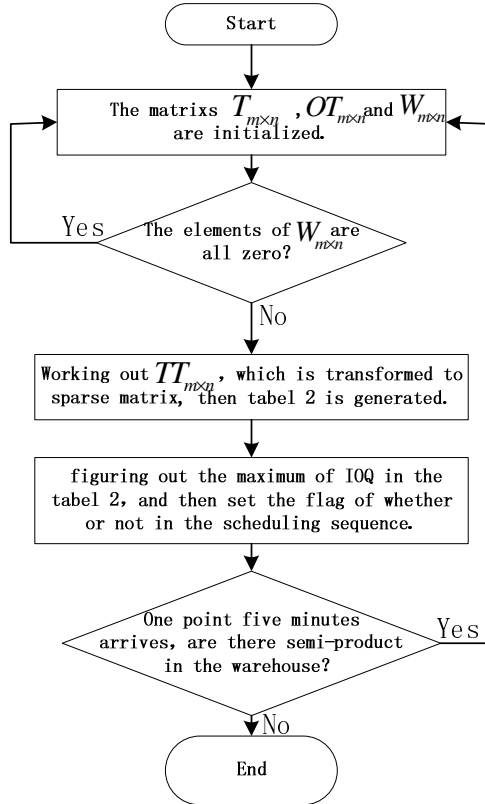
Fig. 1. The flow diagram of the IOQ algorithm

5 Experimental Verification and Analysis

The number of the semi-finished product and their timeout values in different sizes of the warehouse are assumed and the following agreement is done.

The height and the length of the warehouse are marked as H and L respectively. The sum of IOQ and the executive efficiency of a batch task are marked as SOI and T respectively. The semi-finished products whose timeout value is 10 minutes is marked as A, in the same way, the semi-finished products whose timeout value is 20 minutes

is marked as B, the semi-finished products whose timeout value is 30 minutes is marked as C, the semi-finished products whose timeout value is 40 minutes is marked as D, the semi-finished products whose timeout value s 50 minutes is marked as E, the semi-finished products whose timeout value is one hour is marked as F, in the end, if the number of timeout category which belongs to Y is x, then the number of timeout category which belongs to Y is marked as x_Y.

The proposed algorithm is verified and compared with other warehouse scheduling algorithms on the condition whose platform is Windows xp, frequency is 1.49GHz, memory size is 1.99GB , development environment is VC++ 6.0. The results are shown in Table 1 and Table 2.

Table 1. The tasks of the warehouse on different conditions

task	$L{\times}H$	subtask1	subtask2	subtask3	subtask4
task1	$80{\times}4$	$(3_A,5_C,3_D,2_F)$	$(6_C,3_D)$	$(4_B,4_D,1_F)$	$(7_B,3_C,3_D,1_F)$
task2	$92{\times}5$	$(1_C,3_D,1_F)$	$(1_C,3_B,2_E)$	$(2_B,3_C)$	$(2_A,2_C,2_E)$
task3	$100{\times}6$	$(2_A,1_C,1_D)$	$(2_B,3_D,3_E)$	$(3_A,5_C,1_D,1_F)$	$(2_B,1_E,2_F)$
task4	$80{\times}10$	$(3_A,1_C,2_E,6_F)$	$(3_A,3_C,4_D,2_F)$	$(2_C,3_D,2_E)$	$(2_A,3_B,1_F)$
task5	$75{\times}12$	$(6_B,2_D,1_E)$	$(6_A,3_B,1_F)$	$(4_C,3_D,4_E,2_F)$	$(5_A,4_B,6_C,1_D,1_E,1_F)$
task6	$95{\times}15$	$(4_A,3_C,3_D,1_E)$	$(4_B,3_C,5_E,1_F)$	$(2_C,1_D,1_F)$	$(6_A,3_C,3_F)$

Table 2. The comparison of the warehouse scheduling algorithm

Algo-rithm	task1		task2		task3		task4		task5		task6	
	SOI	T/ms	SOI	T/ms	SOI	T/ms	SOI	T/ms	SOI	T/ms	SOI	T/ms
IOQ	732.088	280	288.096	249	343.146	314	724.771	359	685.1	367	653.025	41425
FCFO	861.463	309	308.471	335	383.421	339	777.325	415	834.112	425	753.925	62825
Stoch-astic	857.65	323	318.946	339	368.583	337	801.696	405	806.963	431	753.825	62025

It can be inferred that the proposed IOQ algorithm compared with the other two algorithms has the smallest impact on the quality of the products on each warehouse instance in the Table 2. At the same time, the executive efficiency of the proposed algorithm also is superior to the others, which shows more advantages especially when the size of the warehouse increase. Therefore, it can be concluded that it is effective on improving the executive efficiency of the algorithm by adopting the idea of sparse matrix.

6 Conclusion

The IOQ mathematical model is founded on the basis of the idea that the extent of the unit timeout value whose influence on the product is associated with the storage time of the product, on which the IOQ algorithm is designed replying, the experiment is done by scheduling the semi-finished products with different storage time that should be carried out of the warehouse. And comparing with other warehouse scheduling algorithms, the results show that the proposed algorithm improves the quality of the product greatly. Meanwhile, the idea of the sparse matrix which is introduced to the IOQ algorithm, reduces the search space and collapses the time for solving the optimal solution defined by the target function, and then further improves the products quality and the efficiency of scheduling.

References

1. Hu, G., Kise, H., Xu, Y.: Optimization for input/output scheduling for automated warehouses. Transactions of the Society of Instrument and Control Engineers 18(4), 156–163 (2005)
2. Tian, G.H., Zhang, P., Yin, J.Q.: Research on the fixed shelf order picking optimization problem using a kind of hybrid genetic algorithm. Chinese Journal of Mechanical Engineering 40(2), 141–144 (2004)
3. Li, M.J., Chen, X.B., Liu, C.Q.: Pareto and Niche Genetic Algorithm for Storage Location Assignment Optimization Problem. In: The 3rd International Conference on Innovative Computing Information and Control, pp. 465–465 (2008)
4. Chang, F.L., Liu, Z.X.: Research on Order Picking Optimization Problem of Automated Warehouse. System Engineering-Theory & Practice (2), 139–143 (2007)
5. Wang, C., Zeng, M.M., Li, J.: Solving traveling salesman problems with time windows by genetic particle swarm optimization. In: IEEE World Congress on Computational Intelligence, pp. 1752–1755 (2008)
6. Cao, Q.H., Di, X.K.: A simulated annealing methodology to estate logistic warehouse location selection and distribution of customers' requirement. In: International Workshop on Intelligent Systems and Applications, pp. 1–4 (2009)
7. Shi, X.L., Bai, S.W., Tan, Y.: Research on task scheduling in AS/RS. Logistics Technology (3), 29–32 (2004)
8. Whitely, L.D., Howe, A.E.: Comparing Heuristic Search Methods and Genetic Algorithms for Warehouse Scheduling. Systems, Man, and Cybernetics (3), 2466–2471 (1998)
9. Ascheuer, N., Martin, G., Abdel-Hamid, A.-A.A.: Order picking in an automatic warehouse Solving online asymmetric TSPs. Mathematical Methods of Operations Research 49(3), 501–515 (1999)
10. Gademann, N., Van De Velde, S.: Order batching to minimize total travel time in a parallel-aisle warehouse. IIE Transactions 37(1), 63–75 (2005)
11. Lin, Q., Iyer, B.R.: Automated storage management with QoS guarantees. In: Proceedings of the 22nd International Conference on Data Engineering, p. 150 (2006)
12. Kim, B.-I., Heragu, S.S.: A Hybrid Scheduling and Control System Architecture for Warehouse Management. IEEE Transactions on Robotics and Automation 19(6), 991–1001 (2003)

A Simulation Model Validation Method Based on Functional Data Analysis[*]

Congmin Li, Jiangyun Wang, Liang Han, and Dezhi Dong

School of Automation Science and Electrical Engineering, Beihang University,
100191 Beijing, China
licongmin102@163.com

Abstract. Contrapose the disaccord between simulation data and real data in the process of simulation model validation, a simulation model validation method based on functional data analysis is proposed. This method is a kind of dynamic data analysis method, the simulation data and real data can be regarded as the sample data of two random processes, so the data also can be regarded as infinite dimension vector. Fit the data for function, so the data can be tackled in function form, the correlation analysis of the two random process of the simulation model and real system can be processed through the function data sample, the simulation model validation can be realized. This method proposes a new metrics to evaluate the credibility of the simulation model without the requirement of the time series consistency of simulation data and real data. An example is provided to prove the feasibility of the method.

Keywords: simulation model, model validation, functional data analysis, correlation analysis.

1   Introduction

Simulation model validation is an important way to ensure the credibility of simulation model. The basic method of simulation model validation analyses the output of the simulation model and the real system in the same input conditions [1], [2].

Dynamic data analysis method is important in simulation model validation, there are many traditional dynamic data analysis methods [3], such as error analysis method, inequality coefficient method (TIC), grey correlation analysis method and so on [4]. All these traditional dynamic data analysis method require the simulation data and real data meet the demand of time series consistency, however the data collected generally does not meet the requirement in practice, and the process of making the data meet the demand of time series consistency would cause a certain error.

Considering the issues mentioned above, a functional data analysis method is proposed, this method treats the simulation data and real data as the sample data of two random processes [5]. This method does not require the data meet the demand of time series consistency, and an example is raised to prove the feasibility of this method without restriction of time series consistency for the data.

---

[*] Supported by "the Fundamental Research Funds of Beihang University (No. YWF-12-LZGF-061)".

2   Functional Data Analysis Method

2.1   Characteristics of Functional Data Analysis Method

Functional data analysis method processes the data from the point of view of function [6], [7], [8]. In the process of simulation model validation, the dynamic data can be considered as functional data. When the method is used, there are many restrictions of the data: (1) the simulation data and real data must have the same starting point; (2) the simulation data and real data have the same sampling interval; (3) the simulation data and real data have the same number.

Compared with the traditional data analysis method, functional data analysis method has its own advantages: (1) functional data analysis method can process the infinite dimension data; (2) functional data analysis method requires less assumptions and structure restraint; (3) there is not any demand for number and sampling interval of the data.

2.2   Functional Data Correlation Analysis Method

Assume that two functions X, Y are observed N times, so N pairs of observed curve can be expressed as $(X_i, Y_i), i = 1, 2, \ldots\ldots N$. Describe the correlation of two functions, two comprehensive metrics constructed by the two vectors are needed. The correlation coefficient can describe the correlation of wo functions. Construct metrics $z =< X, \xi >= \int X(t)\xi(t)dt$ and $w =< Y, \eta >= \int Y(t)\eta(t)dt$, two weighting functions ξ, η are needed, the sample variance and covariance function can be expressed as:

$$v_{11}(s,t) = N^{-1} \sum (X_i(s) - \overline{X})(X_i(t) - \overline{X}) \tag{1}$$

$$v_{22}(s,t) = N^{-1} \sum (Y_i(s) - \overline{Y})(Y_i(t) - \overline{Y}) \tag{2}$$

$$v_{12}(s,t) = N^{-1} \sum (X_i(s) - \overline{X})(Y_i(t) - \overline{Y}) \tag{3}$$

The variance of the comprehensive metric z is:

$$Var(z) = \int \left[\xi(s) \int v_{11}(s,t)\xi(t)dt \right] ds \tag{4}$$

The variance of the comprehensive metric w is:

$$Var(w) = \int \left[\eta(s) \int v_{22}(s,t)\eta(t)dt \right] ds \tag{5}$$

The covariance of the two comprehensive metrics z, w are:

$$Cov(z,w) = \int \left[\xi(s) \int v_{12}(s,t)\eta(t)dt \right] ds \tag{6}$$

In order to facilitate the expression, define the operator V_{11}, V_{22}, V_{12}, then $V_{11}f$ stands for $V_{11}f(s) = \int_\tau v_{11}(s,t)f(t)dt$, the same as the V_{22}, V_{12}, so the square of the correlation coefficient of the two comprehensive metrics z, w can be expressed as:

$$\rho^2(z,w) = \frac{<\xi, V_{12}\eta>^2}{<\xi, V_{11}\xi><\eta, V_{22}\eta>} \tag{7}$$

To solve the formula (7), we can convert the formula:

$$\begin{aligned} &\max <\xi, V_{12}\eta> \\ &s.t. <\xi, V_{11}>=<\eta, V_{22}\eta>=1 \end{aligned} \tag{8}$$

The rough penalty term can be joined to the constraint conditions, λ_1, λ_2 are the coefficient of the rough penalty term, they can be valued subjective. The formula (8) can be converted to:

$$\begin{aligned} &\max <\xi, V_{12}\eta> \\ &s.t. <\xi, V_{11}>+\lambda_1\left\|D^2\xi\right\|=<\eta, V_{22}\eta>+\lambda_2\left\|D^2\eta\right\|=1 \end{aligned} \tag{9}$$

To derive the formula (9) further:

$$\begin{aligned} &\max <\xi, V_{12}\eta> \\ &s.t. <\xi, (V_{11}+\lambda_1 D^4)\xi>=<\eta, (V_{22}+\lambda_2 D^4)\eta>=1 \end{aligned} \tag{10}$$

To construct the Lagrange function:

$$G =<\xi, V_{12}\eta> -a(<\xi, (V_{11}+\lambda_1 D^4)\xi>-1)-b(<\eta, (V_{22}+\lambda_2 D^4)\eta>-1) \tag{11}$$

Take derivatives of the two weighting functions ξ, η, we get the next two formulas:

$$\begin{aligned} V_{12}\eta &= \rho(V_{11}+\lambda_1 D_4)\xi \\ V_{21}\xi &= \rho(V_{22}+\lambda_2 D_4)\eta \end{aligned} \tag{12}$$

Convert the equation set to matrix equation:

$$\begin{pmatrix} 0 & V_{12} \\ V_{21} & 0 \end{pmatrix}\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \rho\begin{pmatrix} V_{11}+\lambda_1 D^4 & 0 \\ 0 & V_{22}+\lambda_2 D_4 \end{pmatrix}\begin{pmatrix} \xi \\ \eta \end{pmatrix} \tag{13}$$

The correlation coefficient of the two comprehensive metrics z, w must reflect the correlation of two functions X, Y, so we choose the weighting functions that can maximize the correlation coefficient ρ, the range of the correlation coefficient ρ varies in $[-1,1]$. The two functions are positive correlative when the value range of

the correlation coefficient ρ varies in $[0,1]$, the bigger of the correlation coefficient ρ, the higher of the correlation, These two functions are negative correlative when the value range of the correlation coefficient ρ varies in $[-1,0]$.

In the process of simulation model validation, the dynamic data of the output of the real system and simulation model can be regarded as the sample data of two random processes, Make analysis of the data, the bigger of the correlation coefficient, the higher of the credibility of the simulation model.

3   Application of Functional Data Correlation Analysis Method in Simulation Model Validation

Assume that we get several sets of simulation data and real data, the data can be regarded as the sample data of two functions. The next specific calculation steps can be used to tackle the data:

Step 1. Fit the simulation data and the real data for function X_i, Y_i $i = 1......N$, X_i Y_i stand for the output of simulation model and real system respectively.

Step 2. Assume that the weighting functions are ξ, η, then we can expand the functions X_i, Y_i and ξ, η with the same basic function $\phi_1, \phi_2, ..., \phi_m$.

Step 3. Construct the matrix K and J, and $K(i, j) = < D^2\phi_i, D^2\phi_j >$, $J(i, j) = < \phi_i, \phi_j >$. Define the matrix C, D, and $X_i = \sum_{v=1}^{m} c_{iv}\phi_v, Y_i = \sum_{v=1}^{m} d_{iv}\phi_v$, a, b are coefficient matrix of the expansion of the weighting function ξ, η. Define the matrix V_{11}, V_{12}, V_{22}:

$$V_{11}(v, p) = N^{-1}\sum_i c_{iv}c_{ip} \text{、} V_{12} = N^{-1}\sum_i c_{iv}d_{ip} \text{、} V_{22} = N^{-1}\sum_i d_{iv}d_{ip} \tag{14}$$

Step 4. Construct the matrix equation:

$$\begin{bmatrix} 0 & JV_{12}J \\ JV_{21}J & 0 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \rho\begin{bmatrix} JV_{11}J + \lambda_1 K & 0 \\ 0 & JV_{22}J + \lambda_2 K \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} \tag{15}$$

Let $\lambda_1 = \lambda_2 = 1$. Solve the matrix equation, we can get the max correlation coefficient ρ, it represents the relativity of the correlation of the output of the simulation model and real system.

4   Case Study

Apply the functional data correlation analysis method to the simulation model validation, The sample interval and the number of simulation data and real data are

different, X_1, X_2, X_3 represent three groups of simulation data respectively, the sample interval of every group of simulation data is 0.1s, the number of every group of simulation data is 100. Fit the data for function with polynomial fitting method, the fitting result is shown in fig 1.
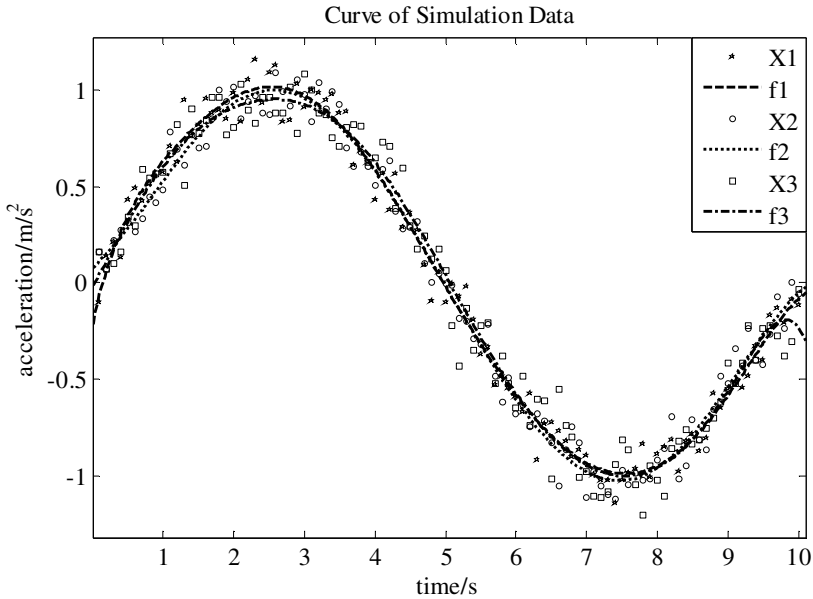
Fig. 1. Three curves of the three groups of simulation data are shown, they are the fitting result of Three groups of output data of simulation model

The fitting result can be expressed as function form, the results are:

$$f_1(x) = 1.878 \times 10^{-6} x^9 - 9.884 \times 10^{-5} x^8 + 0.002206 x^6 - 0.02721 x^5$$
$$+ 0.2021 x^4 - 0.9203 x^3 + 2.532 x^2 - 4.189 x - 1.147$$
$$f_2(x) = 5.077 \times 10^{-6} x^9 - 0.00025 x^8 + 0.00523 x^7 - 0.06063 x^6$$
$$+ 0.4253 x^5 - 1.846 x^4 + 4.884 x^3 - 7.679 x^2 + 6.918 x - 2.088$$
$$f_3(x) = 4.912 \times 10^{-6} x^9 - 0.000256 x^8 + 0.005674 x^7 - 0.06978 x^6$$
$$+ 0.5201 x^5 - 2.406 x^4 + 6.801 x^3 - 11.3 x^2 + 10.21 x - 3.119$$

Y_1, Y_2, Y_3 represent three groups of real data respectively, the sample interval of each set of data is 0.2s, the number of every group of data is 50.

Fit the data in for function with polynomial fitting method, the fitting result is shown in fig 2. g_1, g_2, g_3 are the fitting result of the three groups of real data. Y_1, Y_2, Y_3
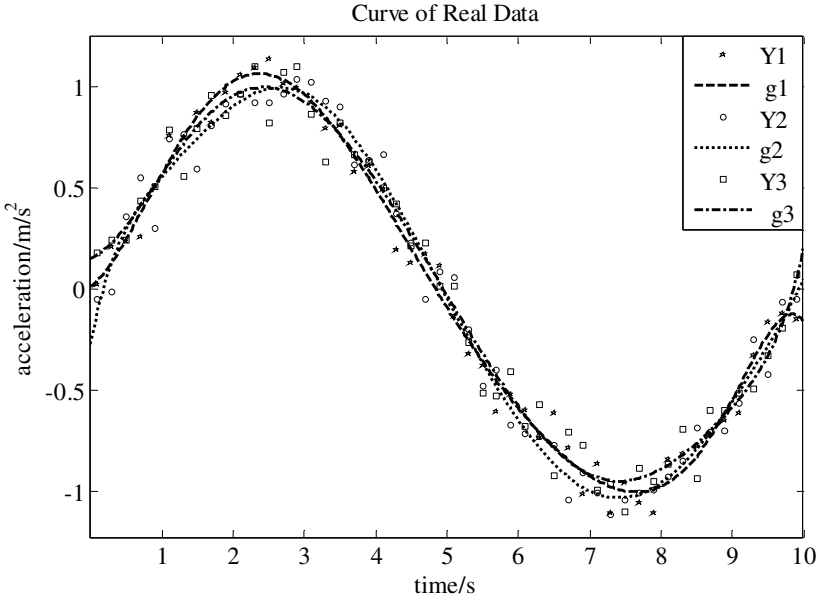
Fig. 2. Three curves of the three groups of real data are shown, they are the fitting result of Three groups of output data of real system

The fitting result can be expressed as function form, the results are:

$$g_1(x) = 3.145 \times 10^{-6} x^9 - 0.0001476x^8 + 0.002918x^7 - 0.03163x^6$$
$$+0.2046x^5 - 0.8045x^4 + 1.909x^3 - 2.845x^2 + 2.989x - 0.8644$$
$$g_2(x) = 3.805 \times 10^{-6} x^9 - 0.0001875x^8 + 0.003943x^7 - 0.04621x^6$$
$$+0.3294x^5 - 1.46x^4 + 3.968x^3 - 6.498x^2 + 6.275x - 2.043$$
$$g_3(x) = -2.62 \times 10^{-6} x^9 + 0.000145x^8 - 0.003394x^7 + 0.04402x^6$$
$$-0.3475x^5 + 1.729x^4 - 5.371x^3 + 9.736x^2 - 8.803x + 3.635$$

Expand the functions that fitted from the simulation data and real data with the same basic function, we choose the Fourier basic function:

$$\phi_1 = \cos(5/23\pi \times (x - 27/5))\phi_2 = \sin(5/23\pi \times (x - 27/5))$$
$$\phi_3 = \cos(10/23\pi \times (x - 27/5))\phi_4 = \sin(10/23\pi \times (x - 27/5))$$
$$\phi_5 = \cos(15/23\pi \times (x - 27/5))\phi_6 = \sin(15/23\pi \times (x - 27/5))$$
$$\phi_7 = \cos(20/23\pi \times (x - 27/5))\phi_8 = \sin(20/23\pi \times (x - 27/5))$$

Let $\Phi = [\phi_1 \ \phi_2 \ \phi_3 \ \phi_4 \ \phi_5 \ \phi_6 \ \phi_7 \ \phi_8]^T$ .The expansion of the simulation data function and real data function with Fourier basic function can be expressed as:

$$f_1(x) = -0.2598\phi_1 - 1.079\phi_2 + 0.001594\phi_3 + 0.1173\phi_4$$
$$+0.01714\phi_5 - 0.004596\phi_6 - 0.008957\phi_7 + 0.04237\phi_8$$
$$f_2(x) = -0.2431\phi_1 - 1.010\phi_2 - 0.005960\phi_3 + 0.06863\phi_4$$
$$+0.003421\phi_5 - 0.03917\phi_6 - 0.004178\phi_7 + 0.03976\phi_8$$
$$f_3(x) = -0.2412\phi_1 - 0.8988\phi_2 + 0.0001208\phi_3 + 0.1028\phi_4$$
$$+0.0008439\phi_5 - 0.06252\phi_6 + 0.01559\phi_7 + 0.03881\phi_8$$
$$g_1(x) = -0.2319\phi_1 - 0.9843\phi_2 + 0.003080\phi_3 + 0.1094\phi_4$$
$$-0.008781\phi_5 - 0.05501\phi_6 + 0.4868\times10^{-4}\phi_7 + 0.04684\phi_8$$
$$g_2(x) = -0.2422\phi_1 - 0.9782\phi_2 + 0.006256\phi_3 + 0.07485\phi_4$$
$$+0.00042\phi_5 - 0.03138\phi_6 - 0.003737\phi_7 + 0.04074\phi_8$$
$$g_3(x) = -0.2103\phi_1 - 1.027\phi_2 + 0.02230\phi_3 + 0.07588\phi_4$$
$$-0.01763\phi_5 - 0.04543\phi_6 - 0.003794\phi_7 + 0.05441\phi_8$$

Then construct the coefficient matrix of the functions. the element of C and D are the coefficient of the expansion of the simulation data functions and real data functions, and they meet the equation $f = C \times \Phi$ and $g = D \times \Phi$. Then we construct the variance matrix and covariance matrix V_{11}, V_{12}, V_{22} and K, J.

At last, we construct the matrix equation:

$$\begin{bmatrix} 0 & JV_{12}J \\ JV_{21}J & 0 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \rho\begin{bmatrix} JV_{11}J + \lambda_1 K & 0 \\ 0 & JV_{22}J + \lambda_2 K \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix}$$

Solve the equation. The eigenvalues are:

Table 1. The eigenvalue s olution of the matrix equation

Index	Eigenvalue value	Index	Eigenvalue value
1	-0.9314	9	-0.3303×10-5+0.1603×10-5i
2	0.9314	10	-0.3303×10-5-0.1603×10-5i
3	-0.00214	11	0.3303×10-5+0.1603×10-5i
4	0.00214	12	0.3303×10-5-0.1603×10-5i
5	-0.000125	13	0.1323×10-22+0.6056×10-6i
6	0.000125	14	0.1323×10-22-0.6056×10-6i
7	0.3282×10-2+0.6712×10-5	15	-0.1653×10-6
8	0.3282×10-2-0.6712×10-5	16	0.1653×10-6

We select the max value $\lambda_{max} = 0.9314$ as the correlation coefficient, the value is very close to 1, so we can conclude that the simulation data and the real data are very similar, and the simulation model is credible.

5   Conclusion

This paper introduces a simulation model validation method based on functional data correlation analysis. When we apply this method, there is no requirement for the data format, it is permitted that the simulation data and the real data have different sample interval and number. Meanwhile when the data is undulate violently, the error may cause the wrong result. So how to fit the data that undulate violently for function is worth considering in the further research.

References

1. Wu, J., Wu, X., Chen, Y., Teng, J.: Validation of Simulation Models Based on Improved Grey Relational Analysis. Systems Engineering and Electronics 32, 1677–1679 (2010)
2. Wu, X., Yang, L., Zhou, Y., Gao, Z.: Research of the Method for Simulation Model Validation Based on Wavelet Transform. Journal of Air Force Engineering University (Natural Science Edition) 9, 53–56 (2008)
3. Liu, H., Sun, Z., Liu, X.: Research and Implementation of Missile Model Validation System. Modern Electronics Technique 35, 1–4 (2012)
4. Li, Y.: Research on Validation Methodology and Tool for Models of Simulation System. Harbin Institute of Technology, Harbin (2008)
5. Jin, L.: The Study on the Methods of Functional Data Analysis and Their Application. Xiamen University, Xiamen (2008)
6. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Science Press, Beijing (2007)
7. Ramsay, J.O., Silverman, B.W.: Applied Functional Data Analysis. Methods and Case Studies. Springer, New York (2002)
8. Yan, M.: The Statistical Analysis of Functional Data: Thoughts, Methods and the Applications. Statistical Research 24, 88–93 (2007)

3D Gesture-Based View Manipulator
for Large Scale Entity Model Review

Hye-Jin Park[1], Jiyoung Park[2], and Myoung-Hee Kim[1,2,*]

[1] Dept. of Computer Science & Engineering,
Ewha Womans University,
11-1 Daehyundong, Seodaemungu, Seoul, Korea
`{clrara,lemie}@ewhain.net`
[2] Center for Computer Graphics and Virtual Reality,
Ewha Womans University,
11-1 Daehyundong, Seodaemungu, Seoul, Korea
`{mhkim}@ewha.ac.kr`

Abstract. Hand gesture-based Human Computer Interaction (HCI) is one of the most natural and intuitive methods of communication between humans and machines because it closely mimics how humans interact with each other. Its intuitiveness and naturalness are needed to explore extensive and complex data or virtual realities. We developed a 3D gesture interface to manipulate the display of a 3D entity model. For gesture recognition, we use the Kinect as a depth sensor to acquire depth image frames. We track the position of the user's skeleton in each frame and detect preset gestures. By simple gestures, the user can pan, zoom, rotate, and reset the view and freely navigate inside the 3D entity model in the virtual space. The proposed gesture interface is integrated with the stereoscopic 3D model viewer that we have previously developed for 3D model review.

Keywords: 3D gesture interface, view manipulator, battlefield visualization.

1   Introduction

Virtual reality techniques are important in battlefield visualizations because the ability to visualize a realistic 3D battle scene and impart a feeling of immersion helps commanders monitor the battle space better and evaluate the simulated operations in a more intuitive way. Stereoscopic 3D visualization is a common functionality in many battlefield visualization software products. In the current battlefield visualizations, however, the entity information is usually presented as simple text or images because the overview of the battlespace takes priority over the detailed view of each entity. In our previous work [1], we constructed a stereoscopic 3D model viewer to provide a separate view for reviewing 3D entity models and integrating them into a battlefield visualization tool that ran a battle scenario. However, the developed model viewer lacked an intuitive interface for manipulating the view and supported mouse and keyboard inputs only.

---

[*] Corresponding author.

In this paper, we describe our recent work on using a 3D gesture interface to manipulate the view of a stereoscopic 3D model viewer. For gesture detection, we use the Kinect sensor [2] to retrieve the user's skeleton. We track the 3D positions of the user's left and right arms and determine the manipulating actions needed to view the transformation.

This work contributes to enhanced user interactions by providing a 3D gesture interface for manipulating the view. Therefore, the user can review the 3D entity model intuitively using simple gestures. When using a large display, which occurs often in design review, a 3D gesture interface is more desirable than a mouse or keyboard because it allows the user to move freely and avoids the need to carry devices.

The rest of the paper is organized as follows. Section 2 offers a brief review of a few examples of military simulation systems, focusing on visualizing the simulated scenes and the interface for user inputs. Section 3 presents an overview of the proposed system, including our stereoscopic 3D model viewer. Section 4 presents the suggested 3D gesture interface using a depth sensor and the integration with the model viewer. Section 5 discusses the implementation of the system and the results obtained. Finally, in Section 6, we conclude the paper by presenting the potential benefits and pitfalls of the proposed system and by outlining imminent research areas. We also provide a tentative direction for future work.

2   Related Work

In the most of the present battlefield visualization systems, mice and keyboards remain the most used interaction devices. For systems that run an immersive display (such as virtual workbenches), 3D input devices (such as joysticks and datagloves) have been employed.

The Dragon system [3] developed by the Naval Research Laboratory is a battlefield visualization system implemented on a virtual reality responsive workbench. They use speech recognition, a pinchglove, and a hand-held joystick as input devices for the Dragon system. They discuss the problems of using speech interfaces and pinchgloves and choose to use a joystick. They place a six degrees-of-freedom position sensor inside a three-button game joystick to track the position and orientation. An interaction metaphor called a "virtual laser pointer" is implemented using a laser beam that comes out of the front of the joystick and enters the virtual space. Using the virtual laser pointer, the user points and selects an entity in the scene. For virtual environment navigation, various button combinations are provided for setting navigation modes such as pan, zoom and rotate. Jones et al. [4] developed a virtual reality environment to support the early-stage review process for complex ship designs. The user interacts with the system using a dataglove with vibrotactile stimulators that can sense the angle and relative position of the fingers and thumb. By receiving vibration and audio feedback, the user can pick up and operate the model freely in the space. Charissis et al. [5] developed a virtual reality system to visualize a submarine rescue vehicle and simulate a rescue mission. They construct a virtual reality display environment based on stereoscopic projection and a high resolution

wide screen. A dataglove with tactile feedback and 3D tracking is used as the interaction device. They also use a SpacePilot device for free navigation and a range of preset viewpoints. Although all of the above-mentioned systems have better performance than conventional PC-based visualization systems, the interaction devices cannot avoid the need for the user to wear or carry the device.

An intuitive interface is an important factor governing the effectiveness and usability of an interactive application, and several controller-free gesture-based techniques have been presented to manipulate views and 3D objects.

Controller-free gesture interfaces allow users to manipulate the graphical contents of the display with their bare hands. All of the interfaces employ tracking system to recognize mid-air hand or arm gestures as user input. Sato et al. [8] estimated the position of the user's hand in three dimensions and recognize hand shape patterns in real-time using multiple cameras. Van den Bergh et al. [10] introduced a real-time hand gesture interaction system based on adaptive skin color segmentation and Haarlets. The system employed only a RGB camera to detect hand gestures and hand movements. However, the system is often disturbed by skin-colored objects in the background, and the hands are not allowed to overlap with the user's face or with people in the background. To overcome issues with overlap and people in the background, Van den Bergh et al. [11] developed a hand detection algorithm based on depth and color. The depth data used in the system is acquired using a Time-of-Flight (ToF) camera. This approach both improves detection rates and allows the hand to overlap with the user's face or with the hands of other persons in the background. The proposed gesture interactions consist of selecting, panning, zooming and rotating. ToF depth cameras are quite expensive and have some disadvantages, however, including low resolution and sensitivity to ambient light.

The Kinect sensor, a real time depth sensing device, has recently been developed to support gaming with the Microsoft Xbox 360. Due to its low cost and wide availability, it has been employed in several research projects in various disciplines. The Kinect can track a user's body and movements without the assistance of attached devices or tracking markers. Gallo et al. [12] provide a controller-free interactive exploration system for medical imaging data using the Kinect. By using both kinetographic and metaphoric hand and arm gestures, the system allows users to execute basic tasks (such as image selection, zooming, translating, rotating and pointing) and complex tasks (such as manually selecting and extracting a region-of-interest (ROI) and interactively modifying the transfer function used to visualize medical images). Song et al. [9] use a handle bar metaphor to support controller-free object manipulation in a 3D virtual environment. It mimics the familiar situation of handling objects that are skewered with a bimanual handle bar. The users can manipulate the virtual handle bar by changing the relative position and orientation of the invisible line that joins their two hands in the physical space. The proposed bimanual interaction supports various visual manipulation tasks that involve translation and rotation, constrained rotation, and multiple object alignment using Kinect sensor.

In this paper, we describe a gesture-based interface that uses the Kinect for view manipulation. Our proposed gesture interface can support efficient and effective hand manipulation of 3D objects and the view without the need to wear or carry devices.
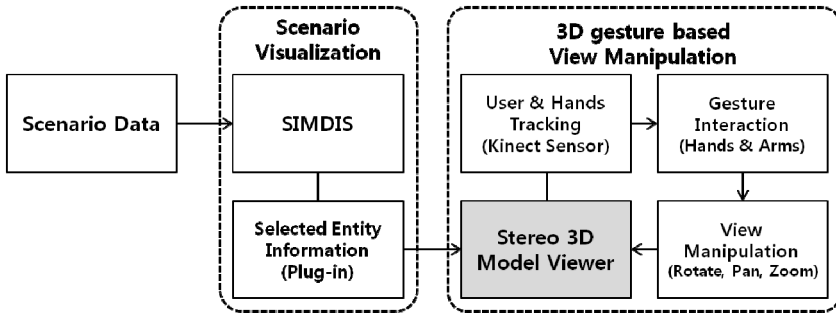
Fig. 1. System overview

3 An Overview of the Stereoscopic 3D Model Viewer

Our system overview is shown in Figure 1. The system consists of the SIMDIS [6], the stereoscopic 3D model viewer, a plug-in to connect the viewer to the SIMDIS, and the Microsoft Kinect sensor. The model viewer is implemented using the OpenSceneGraph (OSG) toolkit [7] on which the SIMDIS is based.

The operation of the system starts with loading the SIMDIS with the scenario data that are to be visualized. Using the SIMDIS GUI, the user can play and stop the scenario, and the user is allowed to click on an entity to load a more detailed view into the stereoscopic 3D model viewer. The underlying plug-in captures the click event, generates a message, and sends it to the model viewer. We use 3D models with more details than the models in the SIMDIS. The model viewer allows the user to view the selected entity in either monographic or stereographic mode. We support four different stereoscopic 3D outputs— anaglyph, side-by-side, top-bottom, and quad-buffered.

We provide two options for viewing the manipulation interface. The user can elect to use the mouse or gesture mode. When the gesture mode is activated, the Kinect sensor starts to acquire depth image frames, and the skeletal tracking module computes the user's skeleton in each real-time frame.

4 View Manipulation by Gesture Interface

We developed a 3D model viewer for large-scale entity model review using large displays in [1]. The proposed model viewer supports only a standard keyboard and mouse as the user interaction devices. However, these devices are not suitable for large displays because they were designed to be used in a PC environment. To use the mouse, the user needs to stay within arm's reach of a table, which reduces the mobility of the user. The user also needs to spend a small amount of time considering how their mouse movements will be mapped onto the large display [13]. The system needs to allow for more intuitive interactions between the user and the object shown on the large display. Therefore, we proposed manipulating the view with a gesture interface that uses information on the user's hand movements detected by the Kinect sensor.

In this section, we describe our system's gesture interface. All of the predefined gestures are mapped to the interaction commands. The user can execute the interaction by gestures at a distance from the display without using any devices. The user's body is represented as a skeleton that consists of line segments linked by joints. The motion of the joints provides the key to motion estimation and recognizing the full skeleton.

We use dynamic gestures characterized by the spatial-temporal motion structures in frame sequences to discriminate between actions. To provide an unambiguous method for specifying the start and end points of gestures, both in time and space, the gesture interface adopts the concept of an activation area. The system continuously checks whether at least one of the user's hands is inside the activation area. The proposed gesture interface is integrated with the stereoscopic 3D model viewer.

4.1   The OSG Manipulator

To observe a 3D model in detail, a user has to be able to move through the virtual scene and freely control his vantage viewpoint. The OSG library controls the virtual scene using a manipulator. A manipulator consists of a camera for the user viewpoint and commands (e.g., translation, rotation and zoom) for camera movement that are controlled by the user through input devices, such as a mouse, a keyboard, a joystick, touch, etc. Every interaction event between the user and the system is implemented by a manipulator.

The OSG toolkit offers trackball, flight, drive, terrain, orbit and spherical manipulators. These manipulators offer different capabilities, and a user can switch between them according to the characteristic of the virtual scene. In addition, the toolkit allows a developer to customize an appropriate manipulator for their system by overriding the existing manipulator functions. We defined a 3D gesture manipulator based on a trackball manipulator to accept the user's hand movement as a mouse substitute.

4.2   User Tracking Using the Kinect

To detect the user's hand gesture, we track the user and the user's hand positions in real time. As mentioned above, we use the Kinect Sensor and its Software Development Kit (SDK) for hand tracking. The Kinect Sensor has a built-in depth camera that acquires depth data; these data indicate the distance of the object from the Kinect sensor. The image frames of the depth map are 640 × 480, 320 × 240, or 80 × 60 pixels in size. Each pixel represents the Cartesian distance (in millimeters) from the camera plane to the nearest object at that particular Cartesian coordinate (x, y). The Kinect SDK allows the acquisition of depth data and skeletal tracking of the images from one or two persons who are within the Kinect sensor's field of view. The data are provided to the application code as a set of points that constitute a skeleton. The coordinate system for the skeleton space is a right-handed coordinate system measured in meters. The x, y, and z axes are the body axes of the Kinect sensor. The origin point places the center of the Kinect sensor, and the positive z axis extends in the direction in which the Kinect sensor is pointed. The positive y axis extends upward, and the positive x axis extends to the left.

The skeleton tracking engine assigns each skeleton a unique identifier. The identifier is an integer that incrementally grows with each new skeleton. If the skeleton engine loses the ability to track a user (e.g., if the user walks out of the view), the tracking

identifier for that skeleton is retired. Each time the event handler executes, it finds the first valid skeleton. If there are two visible users, the user closest to Kinect becomes the primary user. The gesture manipulator is controlled by the primary user. The system checks whether the torso and left and right hand joints are within the activation area. The activation area is not a fixed location in front of the Kinect. It can be defined as a relative distance between the torso and the hand position. If the user's arms are stretched forward more than 30 cm, the activation area is active.

4.3   Gesture Interaction for View Manipulation

Hand gesture interaction in our system is composed of four gestures for rotating, panning, zooming and reset the view. Each gesture's features are identified by the type of hand activation: RightHand, LeftHand, and TwoHands. The RightHand can perform the rotation command, and the LeftHand can perform the panning command. "TwoHands" means that both hands are simultaneously activated. There are two events in the TwoHands mode: the zooming command and the reset command. Figure 2 shows an overview of the gesture interaction for view manipulation.
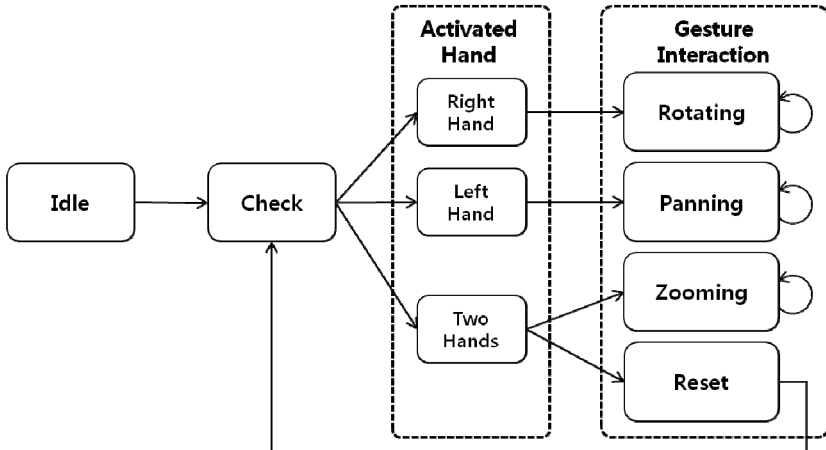
Fig. 2. The state transitions diagram for the gesture interface

The system updates the distance between the user's hands and torso in every frame. Before starting the gesture interaction, the system uses two steps to discriminate the gesture-to-action mapping. First, the system checks whether the hand position inside the activation area remains motionless on the z axis. If the hand position on this axis is not changed, the system stores the current position. In second step, the system checks whether the hands have moved more than a minimum distance from the stored hand position. If the motion meets the requirements for activating a gesture, the state enters the interaction state: rotating, panning, zooming and reset. An initiated interaction terminates when both hands leave the activation area.

To enter the rotating gesture state (Figure 3 (a)), the user's right hand is moved inside the activation area while the left hand remains outside of it. The system calculates the rotating axis and angle using the start and end positions of the hand

movements because our customized manipulator is based on a trackball manipulator. The manipulator finds the axis and angle of rotation using the cross product and dot product, respectively.

In contrast to the rotating gesture, the panning gesture (Figure 3 (b)) is activated by moving the left hand inside the activation area without the right hand. Once the system enters this state, the viewpoint follows the user's left-hand movements. The panning direction can only move in the xy plane. The magnitude of the movement vector is computed using the current and previous user hand positions.

The zooming gesture (Figure 3 (c)) is similar to the widespread gesture used in touch devices such as smartphones. If both hands remain inside the activation area while moving in opposite directions, the system calculates a zoom factor by comparing the distance between the hands with the distance in the first recognized frame. The distance between the two hands is calculated using only the x axis.

(a) rotating   (b) panning   (c) zooming
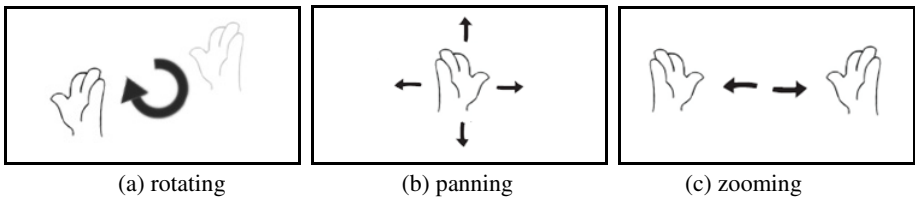
Fig. 3. The hand gestures for view manipulation

It is common to get lost in a 3D environment or to set a strange viewpoint. Therefore, a system should provide a method for resetting the home position. To reset the current viewpoint to its initial position, the user raises both hands over his/her head at the same time. When the positions of both hands are higher than the user head's, the viewer resets the viewpoint to the home position.

5   Results

A user can interact with the SIMDIS and the model viewer by predefined gestures. Our system allows combining our hand interaction system with standard mouse and keyboard interaction. The mode is switched through keyboard inputs. Pressing the '2' key activates the hand interaction mode, after which the user can use interactions to manipulate the view. Pressing the '1' key deactivates the hand interaction mode and returns to the standard mouse and keyboard interaction mode. Figure 4 is an example of the user rotating, panning, zooming, and resetting the view through the 3D model viewer. The left and right columns show views of the system operating at the start and the end of a gesture. View-manipulation interaction in the 3D model viewer is determined only by the user's movements and without image processing. We used a 3D display for the stereoscopic 3D model viewer. The user can see the stereoscopic images by wearing shutter glasses. The SIMDIS and the model viewer run on a standard workstation with dual 3.2 GHz Intel CPUs and 4 GB of RAM. The workstation runs on the Windows 7 operating system and is equipped with an NVIDIA graphics board. The resolution of the acquired depth image was 320×240 pixels. The skeletal tracking module updates at approximately 30 frames per second, and the entire system operates in real time.

(a) Rotating the model
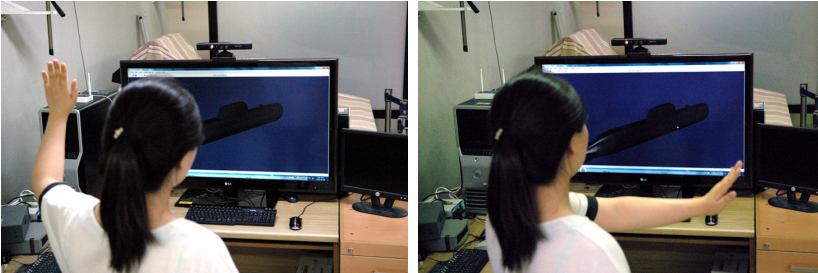
(b) Panning the model
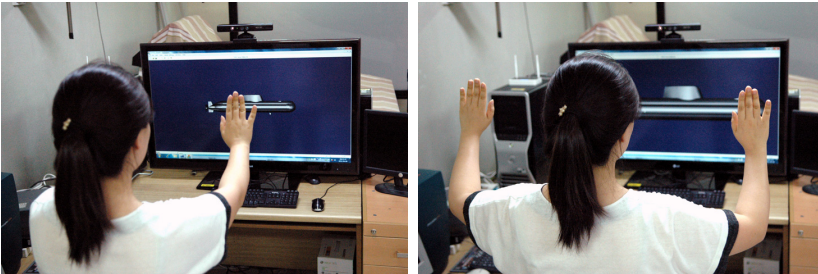
(c) Zooming the model
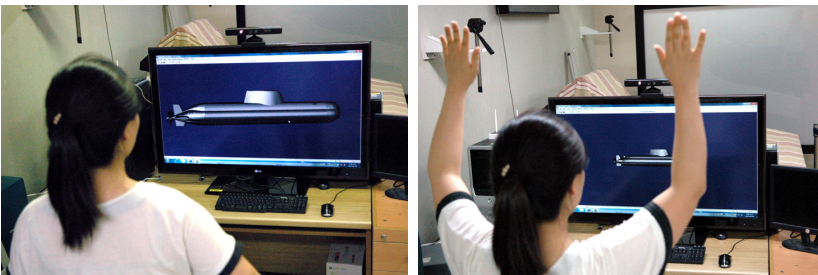
(d) Reset the viewpoint

Fig. 4. An example of the view manipulation by gesture interaction: (a) Rotating, (b) Pannning, (c) Zooming the model and (d) Reset the view

6   Conclusion

We developed a 3D gesture interface to manipulate the view in a 3D model and integrated it with a 3D model viewer for selected battlefield entities. The 3D gesture interface allows the user to control the view of the 3D model with hand gestures in real-time and without using special input devices. It recognizes 3D gestures using the depth positions of the hands and provides the user mobility for interacting with a large display. However, the user is required to be within the depth range (1.2 m ~ 3.5 m) of the Kinect sensor, and the issue of rapid onset of fatigue remains a problem when using gesture interactions for precise control.

Our system provides only four hand gesture interactions (rotating, panning, zooming and reset) and allows only a single user to manipulate the view. In future work, we will extend the set of viewing features and allow multiple users.

References

1. Park, J., Park, H.-J., Kim, M.-H.: Stereoscopic 3D Model Viewer with Improved Depth Perception for Battlefield Visualization. In: Proceedings of the Asia Simulation Conference 2011 (2011)
2. Microsoft Kinect, http://www.xbox.com/en-US/kinect
3. Durbin, J., Swan II, J.E., Colbert, B., Crowe, J., King, R., King, T., Scannell, C., Wartell, Z., Welsh, T.: Battlefield Visualization on the Responsive Workbench. In: Proceedings of the Conference on Visualization 1998, pp. 463–466 (1998)
4. Jones, B.S., Naef, M., McLundie, M.: Interactive 3D Environments for Ship Design Review and Simulation. In: Proceedings of the 5th International Conference on Computer Applications and Information Technology in the Maritime Industries 2006 (2006)
5. Charissis, V., Naef, M., Jones, B.S., Ramsay, J., Sharples, B.: 3D Visualisation of Submarine Rescue Systems and Rescue Mission Simulation. In: Proceedings of the RINA Warships 2008: Naval Submarines (2008)
6. The SIMDIS, https://simdis.nrl.navy.mil
7. OpenSceneGraph, http://www.openscenegraph.org
8. Sato, Y., Saito, M., Koike, H.: Real-time Input of 3D Pose and Gestures of a User's Hand and Its Applications for HCI. In: IEEE VR, pp. 79–86 (2001)
9. Song, P., Goh, W.-B., Hutama, W., Fu, C.-W., Liu, X.: A Handle Bar Metaphor for Virtual Object Manipulation with Mid-Air Interaction. In: CHI 2012, pp. 1297–1306 (2012)
10. Van den Bergh, M., Bosché, F., Koller-Meier, E., Van Gool, L.: Haarlet-based Hand Gesture Recognition for 3D Interaction. In: Workshop on Applications of Computer Vision (WACV), pp. 1–8 (2009)

11. Van den Bergh, M., Van Gool, L.: Combining RGB and ToF Cameras for Real-time 3D Hand Gesture Interaction. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), pp. 66–72 (2011)
12. Gallo, L., Placitelli, A.P., Ciampi, M.: Controller-free Exploration of Medical Image Data: Experiencing the Kinect. In: 2011 24th International Symposium on Computer-Based Medical Systems (CBMS), pp. 1–6 (2011)
13. Cheng, K., Pul, K.: Direct Interaction with Large-Scale Display Systems using Infrared Laser Tracking Devices. In: APVis 2003 Proceedings of the Asia-Pacific Symposium on Information Visualisation, vol. 24, pp. 67–74 (2003)

The Research Review on VV&A Working System of Complex Simulation System

Shuli Zhang[1], Huapin Geng, Jiahui Tong, and Mingran Du

[1] Beijing Electromechanical Engineering Institute, Beijing 100074 China
Zhangshuli99@yahoo.com.cn, {geng_sky,mingrandu}@163.com,
tjh80825@126.com

Abstract. From the aspect of management and technique, this paper describes the necessity and basic thinking of making the research on VV&A working system in complex simulation system, and also proposes the specific details and suggestions. This will provide much support for formulating VV&A in complex simulation system.

Keywords: complex simulation system, VV&A, organization mechanism, standard.

1   Introduction

Along with the development of simulation technology, simulation requirements continue to expand in more areas, and the importance of simulation is becoming more obvious. Simulation is even regarded as the third mean or tool to recognize the world, besides the theoretical research and experimental research. With the increasing of the fields of simulation applied and the scale of simulation system, the complex simulation system integrated with field testing, hardware-in-loop simulation and digital simulation occurs.

However, based on similar principium, model theory, system technology, information technology and some M&S relative technology, simulation technology is a kind of comprehensive technology including multiple discipline, using computer and some physics effect instruments as tool, and making use of model to make, analyze and evaluate, and participate in the system. Simulation result cannot be exactly the same as the result of the prototype system, so users concern how much the simulation system can reproduce the prototype system, in other words, the first question in simulation application is simulation credibility. The function of VV&A work is equivalent to the project supervision in simulation development, representing the users and experts, to monitor the quality in the simulation development effectively. VV&A may improve the credibility of simulation system, and make the simulation system more suitable for user requirements. Therefore, the research on VV&A and credibility of simulation system are necessary parts in the simulation application.

In order to better monitor the quality in the developing process and improve the creditability of complex simulation system, developers and users of complex simulation system are clearly aware of the necessity and importance to perform

VV&A in the life cycle. But at present, the effects of VV&A in complex simulation system are not satisfying, the reason can be summarized into the following aspects: (1) The lack of necessary VV&A organization mechanism: No an authoritative VV&A official organization(similar as ISO) is to plan VV&A work process and development strategy, guide and promote VV&A work progress continually, coordinate and supervise how to perform VV&A standard and VV&A;(2) The lack of effective standard: No a universal standard in VV&A for complex simulation system is to guide VV&A work, leading to slow work; (3) The lack of effective incentive policies and necessary funds and effective investment in human resources: The characteristics of multiple levels and large scale and multiple granularities make VV&A highly comprehensive and relative to multi-discipline in complex simulation system. For a long time there is no effective incentive policy and enough funds and even no human resource investment, all these lead to the marginalization of VV&A work. Therefore, VV&A work doesn't really play its proper role in the system; (4) VV&A theoretical research lagged far behind the practical application: VV&A theoretical research is mainly conducted in the universities .The universities have rich human resources and an advantage to track the foreign technology, but they are lack of the support of practical application of military enterprises because of the restriction of privacy in military enterprises. (5) The software tools in VV&A aspects are not enough perfect and mature to complex simulation system, making the development of VV&A slow.

Over the last ten years of VV&A work practices and extensive research and in-depth thinking, it is necessary to restudy VV&A working system of complex simulation system, call for relative departments to establish effective VV&A work mechanism, make feasible standards of complex simulation system, issue correlative incentive mechanism of VV&A work, promote and perfect the developing progress of VV&A theory and technical methods of complex simulation system, and develop effective software tools to ensure the productivity of VV&A work.

2   Establishing Authoritative VV&A Organization

Now, the main reason of slow VV&A work was lack of an Authoritative VV&A Organization(similar as ISO) to guide VV&A work orderly, There is no a sound organization mechanism to promote VV&A technology development and supervise VV&A work progress effectively. Consequently, VV&A technology was not competitive enough. Lack of organization to guarantee, VV&A staff can't participate in the entire development process of complex simulation system, Making VV&A work not synchronous, and they can't understand the development in-depth, verify, validate and evaluate it objectively, so VV&A can't play a key role in the process.

Consequently, it is urgent to strengthen the construction of organization mechanism, found a third party authentication organization of authority, make a long-term VV&A development planning, establish VV&A organization mechanism and standard, and issue the policy of VV&A. For example, simulation system can't be accepted without credibility guarantee, or simulation system can't be used without

qualification certification of VV&A. It is necessary to perform VV&A work in the life cycle by organization. The purpose of establishing a special team of VV&A researchers is , on the one hand, to make standards and promote the development of VV&A theory and technique; on the other hand, to form the normalization mechanism and promote the popularization and application of simulation technology.

3   Strengthening the Research and Establishment on VV&A Standards of Complex Simulation System

Along with development process of complex simulation system (the requirements analysis, project design, system design, system development, system integration and the maintenance), VV&A is a very important work in life cycle. Only by establishing VV&A standards and specifications of complex simulation system in life cycle, can it? guide the work effectively and orderly. Therefore, strengthening the research on VV&A standard of complex simulation system is not only necessary for VV&A work, but also for perfecting and developing the application.

Currently, there are two relative VV&A standards of complex simulation system. One is IEEE 1278.4, it is the VV&A standard based on DIS; the other is IEEE 1516.4, it is the VV&A standard of distributed interactive simulation system based on HLA. Both are aimed at standards performing the VV&A work for typical system with special structure, so neither of them is universal and can guide VV&A work of complex simulation system effectively. It is necessary to establish a series of universal VV&A work standards of complex simulation system, standardize VV&A work process and content, improve the credibility of complex simulation system and control the quality in the development of simulation system effectively.

One characteristic of VV&A is the end-to-end process with simulation system development. Therefore, the premise of establishing VV&A standards is to make development process of complex simulation system clear, then the corresponding VV&A work is made according to the process, Finally the purposes, principle, process, content, VV&A methods and tools suggested in VV&A standards are confirmed, under the foundation of analyzing the character of the object that to be verified, validated and evaluated in the development process.

The other characteristic of VV&A is to supervise the quality of the development process of the complex simulation system effectively, find the problem and solve it in time, reduce the development costs, and shorten the development period. Therefore, in the process of establishing VV&A standards, the related concepts, methods and technology of the simulation model, software development, system engineering and quality management should be used for reference, and integrated with the quality certification system perfectly to form a complete set of VV&A woking system of complex simulation system so as to guide VV&A work effectively.

The research and establishment on VV&A standards is not only the important content in the simulation technology, but also the problem to be solved urgently in our national defense and military simulation field. It is the work of great significance, and it can improve the level of creditability of complex simulation system, accelerate the standardization, intelligence, integration and automation of the reliability assessment,

speed up the process of normalization and standardization of modeling and simulation, and satisfy the demands of VV&A standards of complex simulation system in our national defense and military areas.

4   Strengthening the Research on VV&A Theory of Complex Simulation System

With the simulation requirements expanding continually, it is urgent to extend simulation technology and human resources. Unfortunately, simulation is not the first level discipline, but the expert group has been committed to the job application for first level discipline and achieved remarkable results. But considering the long-term development of simulation discipline, there must be a complete working system to support it and promote its development. The development of simulation theory is the engine of leading and promoting the rapid development of this discipline, but VV&A is the important factor that supports the development and application of simulation discipline. Judging from the research situation from home and abroad, achievements in VV&A theory are mostly in the concept, significance, principles and process and so on, but VV&A theory and creditability evaluation of complex simulation system are urgent to develop and perfect. Therefore, it is necessary to strengthen the research on VV&A theory, and increase the innovation of VV&A theory, breakthrough the key technology, such as verification and validation in system-level and creditability evaluation and so on, help simulation technology further development, and promote the construction of simulation discipline.

There are some limitations on the current VV&A theory for supporting VV&A work of complex simulation system, such as the lack of the theories of VV&A and credibility evaluation aiming at qualitative simulation, uncertainty simulation and different granularity model ; the lack of the awareness of quality control of simulation system; and merely paying attention to the precision of the model; ignorance of the concept of system quality control, etc..

VV&A theory is an important support for VV&A work. Therefore, under the leadership of special authentication institutions, it is necessary to strengthen the research on VV&A theory, methodology and standard and so on, form the standard theory system of VV&A, and make VV&A work more effectively.

5   Extending the Research on VV&A Technique and Methods of Complex Simulation System

At present, a lot of research results of VV&A technique and method from home and abroad was about verification and validation on the model. VV&A work of complex simulation system is much more complicated than that of the model, Due to complex simulation system have the following characteristics: multiple levels, multiple structure, multiple relationship, multiple models, large scale and so on.

According to the characteristics of the object to be verified, validated and evaluated, the research on VV&A technique of complex simulation system should

make use of the professional knowledge of related fields and experience of VV&A work, abstract the technique and methods suitable for VV&A work of complex simulation system, and guide the efficient implementation of VV&A work

Take a certain simulation system for example, it is a complex simulation system of multi-level and multi-layer, integrated with field testing, hardware-in-loop simulation and digital simulation in signal-level and functional-level. From the aspect of military modeling and simulation, it is composed of engagement simulation(system-level) and engineering simulation(model-level), and it can support the modeling and validation in the mission-level simulation system. "multi-layer"has two meanings. Firstly, it means the system is a three-layers experiments, including field testing, hardware-in-loop simulation and digital simulation. Secondly, it means that the granularity in every layer of the simulation model is different, for example, the granularity in the signal-level simulation model is better than that of functional-level model. According to simulation requirements and purpose ,users can assess the combat process, shoot probability and the ability of anti-jamming for a certain weapon in different levels, and analyze the combat capability and Battle Survivability comprehensively, using the multi-level and multi-layer simulation system structure.

VV&A work of this system includes as follows : requirement analysis verification and validation, project design verification and validation, conceptual model verification and validation , system design verification and validation, implementation verification and validation, results validation and accreditation implementation. Take implementation verification and validation for example, it includes as follows: model verification and validation, simulation equipment verification and test, system integration verification and validation and so on. In every step of VV&A work, proper VV&A methods and technique should be selected and used as the specific operation methods. In addition, in the process of VV&A work, a series of new problems should be considered, for example ,a lot of simulation data involved needing to analysis and processing, the research on support relations of multi-layer simulation experiments, multi-granularity models verification and validation , Space-Time Consistency verification and system interoperability validation and so on. Now, there are not appropriate solutions of these problems, this restricts the development of VV&A work of complex simulation system. Therefore, it is urgent to expand the research on VV&A technique and methods of complex simulation system.

Strengthening the research on VV&A technique and methods of complex simulation system can not only further perfect VV&A theory system, but also establish the foundation for the development of VV&A software tools. At the same time it can solve the problem of single operation methods in VV&A work, and provide the technical support for performing VV&A work in life cycle effectively and comprehensively.

6   The Development of VV&A Automation Software Tools of Complex Simulation System

VV&A work of complex simulation system is a complicated process, it includes as follows: making work objective, making work schedule, confirming work process and

work content, confirming technical methods and managing the resources such as the models, data, documents used in the process. It is difficult and not efficient to carry out VV&A work completely relied on human. Therefore, it is necessary to develop VV&A tools of complex simulation system to make the management more systematic, professional and automatic, to improve the efficiency of VV&A work.

The VV&A automation tools of complex simulation system mainly include the following aspects: Firstly, to improve the extent of standardization of VV&A work, VV&A process should be managed automatically according to VV&A standards .Secondly, to improve the efficiency of VV&A work, the support of methods, technique and calculation should be provided for VV&A work and creditability evaluation. Thirdly, in order to confirm the effective control and usage of the resources including human resource management, document management, dada management, model management, method management, knowledge management and so on, the resources should be managed

At the same time, the automation tools of complex simulation system should fully consider the practicability and universality. Practicability means easy usage in managing and carrying out VV&A work. Universality means the tools should consider VV&A work of all kinds of complex simulation system, not merely be suitable for a certain system.

7   Summarization

To sum up, VV&A work of complex simulation system need not only the necessary organization guarantee, but also strong technical support. It is very urgent and imperative to make research on VV&A working system of complex simulation system.

References

1. IEEE1278.4. Practice for Distributed Interactive Simulations Verification, Validation and Accreditation. The Distributed Interactive Simulation Committee (1997)
2. IEEE Computer Society. 1516.3 IEEE Recommended Practice for High Level Architecture (HLA) Federation Development and Execution Process (2003)
3. IEEE. 1516.4 Recommended Practice for VV&A of a Federation-An Overlay to the HLA FEDEP (2007)
4. Defense Modeling and Simulation Office (DMSO). Verification, Validation and Accreditation Recommended Practice Guides (2000)
5. Jean, G., Youngblood, S., Lewis, R.: A VV&A process of the HLA FEDEP. In: Proceedings of the Summer Computer Simulation Conference (1999)
6. Lewis Robert, O., Dobey Virginial, T.: Verification validation and accreditation (VV&A) process overlay for FEDEP (2003)

Author Index