

Yingmin Jia  
Weicun Zhang  
Yongling Fu *Editors*

# Proceedings of 2020 Chinese Intelligent Systems Conference

Volume I

# Lecture Notes in Electrical Engineering

## Volume 705

### Series Editors

Leopoldo Angrisani, Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Naples, Italy

Marco Arteaga, Departament de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico

Bijaya Ketan Panigrahi, Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, Delhi, India  
Samarjit Chakraborty, Fakultät für Elektrotechnik und Informationstechnik, TU München, Munich, Germany  
Jiming Chen, Zhejiang University, Hangzhou, Zhejiang, China

Shanben Chen, Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

Tan Kay Chen, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore

Rüdiger Dillmann, Humanoids and Intelligent Systems Laboratory, Karlsruhe Institute for Technology, Karlsruhe, Germany

Haibin Duan, Beijing University of Aeronautics and Astronautics, Beijing, China

Gianluigi Ferrari, Università di Parma, Parma, Italy

Manuel Ferre, Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain

Sandra Hirche, Department of Electrical Engineering and Information Science, Technische Universität München, Munich, Germany

Faryar Jabbari, Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA

Limin Jia, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Alaa Khamis, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt

Torsten Kroeger, Stanford University, Stanford, CA, USA

Qilian Liang, Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX, USA

Ferran Martín, Departament d'Enginyeria Electrònica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

Tan Cher Ming, College of Engineering, Nanyang Technological University, Singapore, Singapore

Wolfgang Minker, Institute of Information Technology, University of Ulm, Ulm, Germany

Pradeep Misra, Department of Electrical Engineering, Wright State University, Dayton, OH, USA

Sebastian Möller, Quality and Usability Laboratory, TU Berlin, Berlin, Germany

Subhas Mukhopadhyay, School of Engineering & Advanced Technology, Massey University, Palmerston North, Manawatu-Wanganui, New Zealand

Cun-Zheng Ning, Electrical Engineering, Arizona State University, Tempe, AZ, USA

Toyoaki Nishida, Graduate School of Informatics, Kyoto University, Kyoto, Japan

Federica Pascucci, Dipartimento di Ingegneria, Università degli Studi "Roma Tre", Rome, Italy

Yong Qin, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Gan Woon Seng, School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore, Singapore

Joachim Speidel, Institute of Telecommunications, Universität Stuttgart, Stuttgart, Germany

Germano Veiga, Campus da FEUP, INESC Porto, Porto, Portugal

Haitao Wu, Academy of Opto-electronics, Chinese Academy of Sciences, Beijing, China

Junjie James Zhang, Charlotte, NC, USA

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering - quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact [leontina.dicecco@springer.com](mailto:leontina.dicecco@springer.com).

To submit a proposal or request further information, please contact the Publishing Editor in your country:

#### **China**

Jasmine Dou, Associate Editor ([jasmine.dou@springer.com](mailto:jasmine.dou@springer.com))

#### **India, Japan, Rest of Asia**

Swati Meherishi, Executive Editor ([Swati.Meherishi@springer.com](mailto:Swati.Meherishi@springer.com))

#### **Southeast Asia, Australia, New Zealand**

Ramesh Nath Premnath, Editor ([ramesh.premnath@springernature.com](mailto:ramesh.premnath@springernature.com))

#### **USA, Canada:**

Michael Luby, Senior Editor ([michael.luby@springer.com](mailto:michael.luby@springer.com))

#### **All other Countries:**

Leontina Di Cecco, Senior Editor ([leontina.dicecco@springer.com](mailto:leontina.dicecco@springer.com))

**\*\* Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, SCOPUS, MetaPress, Web of Science and Springerlink \*\***

More information about this series at <http://www.springer.com/series/7818>

Yingmin Jia · Weicun Zhang ·  
Yongling Fu  
Editors

# Proceedings of 2020 Chinese Intelligent Systems Conference

Volume I



Springer

*Editors*

Yingmin Jia  
School of Automation Science  
and Electrical Engineering  
Beihang University  
Beijing, Beijing, China

Weicun Zhang  
School of Automation  
and Electrical Engineering  
University of Science  
and Technology Beijing  
Beijing, Beijing, China

Yongling Fu  
School of Mechanical Engineering  
and Automation  
Beihang University  
Beijing, Beijing, China

ISSN 1876-1100

ISSN 1876-1119 (electronic)

Lecture Notes in Electrical Engineering

ISBN 978-981-15-8449-7

ISBN 978-981-15-8450-3 (eBook)

<https://doi.org/10.1007/978-981-15-8450-3>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

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

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

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

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Contents

<b>Distributed Finite-Time Rotating Encirclement Control for Second-Order Agent Dynamics . . . . .</b>	1
Yixin Li, Yingmin Jia, and Tengfei Zhang	
<b>Urban Road Object Detection and Tracking Applications Based on Acoustic Localization . . . . .</b>	10
Zhimin Wang, Chaoli Wang, and Song Shen	
<b>Air Combat Situation Assessment of Multiple UCAVs with Incomplete Information . . . . .</b>	18
Shouyi Li, Qingxian Wu, Mou Chen, and Yuhui Wang	
<b>Detection and Depth Estimation for Objects from Single Monocular Image . . . . .</b>	27
Ziwen Xu and Yingmin Jia	
<b>Leader-Following Consensus of Multi-agent Systems: New Results Based on a Linear Transformation Approach . . . . .</b>	36
Jingyuan Zhan and Yangzhou Chen	
<b>Proportional Step Perturbation Method MPPT for Boost Circuit of TEG . . . . .</b>	48
Feng Ji and John Xu	
<b>Leader-Following Consensus of Second-Order Networks with a Moving Leader and Nonconvex Input Constraints . . . . .</b>	57
Lipo Mo, Zeyang Xu, and Yongguang Yu	
<b>A Novel Deep Learning Ensemble Model with Secondary Decomposition for Short-Term Electricity Price Forecasting . . . . .</b>	69
Na Chen, Xueqing Yang, Yiming Gan, Wuneng Zhou, and Hangyang Cheng	

<b>Disturbance Observer-Based Finite-Time Control for Systems with Nonlinearity and Disturbance . . . . .</b>	78
Xinqing Li and Xinjiang Wei	
<b>Disturbance Observer-Based Disturbance Attenuation Control for Dynamic Positioning System of Ships . . . . .</b>	88
Lihong You and Xinjiang Wei	
<b>Multiple Change Points Detection Method Based on TSTKS and CPI Sliding Window Strategy . . . . .</b>	99
Jialun Liu, Jinpeng Qi, Junchen Zou, and Houjie Zhu	
<b>Steam Pressure Control Based on PLC 200 Smart . . . . .</b>	110
Jing Lv and Zhongsuo Shi	
<b>Event-Triggered Anti-disturbance Tracking Control for Systems with Exogenous Disturbances . . . . .</b>	117
Xiaoli Zhang, Xiang Gu, Yang Yi, and Tianping Zhang	
<b>Parallel Label Consistent KSVD-Stacked Autoencoder for Industrial Process Fault Diagnosis . . . . .</b>	127
Hongpeng Yin, Jiaxin Guo, Guobo Liao, and Yi Chai	
<b>Iterative Learning Formation Control for Multi-agent Systems with Randomly Varying Trial Lengths . . . . .</b>	136
Yimin Fan, Yang Liu, and Na Wang	
<b>Backstepping-Based Adaptive Neural Control of Constrained Nonlinear Systems . . . . .</b>	145
Penghao Chen, Tianping Zhang, Houbin Qian, and Yang Yi	
<b>A Music Generation Model Based on Generative Adversarial Networks with Bayesian Optimization . . . . .</b>	155
Yijie Xu, Xueqing Yang, Yiming Gan, Wuneng Zhou, Hangyang Cheng, and Xuehui He	
<b>Interactive Attention and Position-Aware Mechanism for Aspect-Level Sentiment Analysis . . . . .</b>	165
Xuehui He, Yiming Gan, Xueqing Yang, Wuneng Zhou, and Yuanhong Ren	
<b>Neural Network-Based Exponential Stability of Affine Nonlinear Systems by Event-Triggered Approach . . . . .</b>	175
Fan Liu, Yiming Gan, Xueqing Yang, and Wuneng Zhou	
<b>Adaptive Finite-Time Leader-Following Consensus of Multi-agent Systems Against Time-Varying Actuator Faults . . . . .</b>	185
Yanhui Yin, Fuyong Wang, Zhongxin Liu, and Zengqiang Chen	

<b>Domain Decomposition Strategies for Developing Parallel Unstructured Mesh Generation Software Based on PadMesh . . . . .</b>	194
Fengshun Lu, Xiong Jiang, Xinbiao Bao, Long Qi, and Yongheng Guo	
<b>Dynamic Trajectory Prediction for Continuous Descend Operations Based on Unscented Kalman Filter . . . . .</b>	206
Jun Zhang, Guoqing Wang, and Gang Xiao	
<b>Research on Weighted Multiple Model Adaptive Control Based on U-Model . . . . .</b>	217
Jiayi Li and Weicun Zhang	
<b>Human Action Recognition Method Based on Video-Level Features and Attention Mechanism . . . . .</b>	225
Qiang Cai, Jin Yan, Haisheng Li, and Yibiao Deng	
<b>Application of LSTM in Aeration of Sewage Treatment . . . . .</b>	234
Shaobo Zhang and Qinglin Sun	
<b>Sliding Mode Control on Coordination of Master-Slave Manipulator . . . . .</b>	242
Lijun Wang, Ningxi Liu, Jinkun Liu, Tianyu Cao, and Jiaxuan Yan	
<b>Pose Ambiguity Elimination Algorithm for 3C Components Assembly Pose Estimation in Point Cloud . . . . .</b>	251
Weikun Gu, Xiansheng Yang, Dengwei Dong, and Yunjiang Lou	
<b>Manipulator Control Law Design Based on Backstepping and ADRC Methods . . . . .</b>	261
Lijun Wang, Jiaxuan Yan, Tianyu Cao, and Ningxi Liu	
<b>Manipulator Calibration-Free Hand-Eye Coordination Based on ADRC Under Eye Fixation . . . . .</b>	270
Lijun Wang, Tianyu Cao, Jiaxuan Yan, and Ningxi Liu	
<b>Adaptive Neural Consensus Tracking for Second-Order Nonlinear Multi-agent Systems with Full-State Constraints . . . . .</b>	278
Dan Liu and Lin Zhao	
<b>Adaptive Sliding Mode Control of Mismatched Quantization System . . . . .</b>	287
Qiaoyu Chen, Wuneng Zhou, Dongbing Tong, and Yao Wang	
<b>An Optimal Feedback Control Law for Spacecraft Reorientation with Attitude Pointing Constraints . . . . .</b>	297
Bin Li, Yang Wang, and Kai Zhang	
<b>Active Detection Based Mobile Robot Radioactive Source Localization Method and Data Processing . . . . .</b>	308
Han Gao, Zhengguang Ma, and Yongguo Zhao	

<b>Distributed Adaptive Consensus Tracking Control for Multiple AUVs with State Constraints . . . . .</b>	317
Jingzi Fan and Lin Zhao	
<b>Intelligent Wireless Propagation Model with Environmental Adaptability . . . . .</b>	326
Xiaoyu Qu and Jiangyun Wang	
<b>The Local Navigation and Positioning System of Unmanned Ground Vehicles . . . . .</b>	333
Shaowei Li, Qingquan Feng, Jiangang Wang, Zhiyong Li, Dongxiao Wang, and Shizhao Liu	
<b>A Novel 3D Lidar-IMU Calibration Method Based on Hand-Eye Calibration System . . . . .</b>	342
Lei Ji, Long Zhao, Jingyun Duo, and Chao Wang	
<b>Design of Semi-physical Simulation System for Multi-target Attack Air Defense Missile Weapon System . . . . .</b>	351
Shujun Yang, Jianqiang Zheng, Qinghua Ma, Shuaiwei Wang, Jirong Ma, Haipeng Deng, and Yiming Liang	
<b>Application of Multi-network Fusion in Diagnosis of Chest Diseases . . . . .</b>	358
Shanshan Zhang and Hai Gao	
<b>Distributed Robust <math>H_\infty</math> Containment Control for Fractional-Order Multi-agent Networks . . . . .</b>	367
Xiaolin Yuan, Yongguang Yu, and Lipo Mo	
<b>Refinement and Validation of Humoral Immunity Based on Event-B . . . . .</b>	377
Xuqing Shi, Shengrong Zou, Yudan Shu, and Li Chen	
<b>Accelerated Distributed Algorithm for Solving Linear Algebraic Equations . . . . .</b>	389
Weikang Hu and Aiguo Wu	
<b>CNN-Based Automatic Diagnosis for Knee Meniscus Tear in Magnetic Resonance Images . . . . .</b>	399
Hao Zhou, Liyan Zhang, Bing Zhang, Juan Wang, and Chengyi Xia	
<b>Application of an Effective Fault Localization Prioritization Method to Stereo Matching Software . . . . .</b>	409
Jinfeng Li, Yan Zhang, and Jilong Bian	
<b>Denoising of X-Ray Pulsar Signal Based on Variational Mode Decomposition . . . . .</b>	419
Yong Zhao, Yingmin Jia, and Qiang Chen	

<b>Signal Estimation of Fatigue-Magnetic Properties of 25CrMo4 Based on Stein Algorithm . . . . .</b>	428
Zhenfa Bi and Guobao Yang	
<b>Intelligent Frequency Selection of the Sky-Wave Radar Based on Numerical Ray Tracing . . . . .</b>	437
Runze Li, Jiangyun Wang, and Guanghong Gong	
<b>Design of Multi-port Energy Conversion System of Electric Vehicle Based on Bridge-Type Buck-Boost Topology . . . . .</b>	446
Yunhao Zhang, Xiaonan Xia, Xiaoxing Ge, Wei Tang, and Yu Fang	
<b>Realization of Automatic Zero Calibration of Inductive Proximity Sensor Based on Inductive Increment Detection . . . . .</b>	455
Yuyin Zhao, Yu Fang, Jiajun Yang, Miao Weng, and Xiaonan Xia	
<b>Sliding Mode Control Method of Powered Parafoil Based on Extended State Observer . . . . .</b>	464
Li Yu, Qinglin Sun, and Panlong Tan	
<b>Analysis of Accelerated Vibration-Magnetic Effect of 25CrMo4 . . . . .</b>	475
Zhenfa Bi and Zongkai Wang	
<b>Automated Prediction of Cervical Precancer Based on Deep Learning . . . . .</b>	485
Bing Zhang, Qingyuan Zhang, Hao Zhou, Chengyi Xia, and Juan Wang	
<b>Dynamic Economic Dispatch Considering Wind Based on Adaptive Crisscross Optimization . . . . .</b>	495
Panpan Mei, Lianghong Wu, Hongqiang Zhang, and Zhenzu Liu	
<b>Consensus for Heterogeneous Networked Systems Based on Second-Order Neighbors' Information . . . . .</b>	508
Lei Wang, Huanyu Zhao, Dongsheng Du, and Hongbiao Zhou	
<b>Sliding Mode Control for Neutral-Type Systems with Stochastic Noises and Time-Delay . . . . .</b>	518
Qiaoyu Chen, Wuneng Zhou, and Dongbing Tong	
<b>Reinforcement Learning Adaptive Tracking Control for a Stratospheric Airship . . . . .</b>	527
Kang Wang, Yang Liu, Zewei Zheng, and Ming Zhu	
<b>Model Reference Adaptive Control with Output Constraints . . . . .</b>	541
Yu Hua, Tianping Zhang, Manfei Lin, and Weiwei Deng	
<b>Low-Dose CT Image Denoising Using a Generative Adversarial Network Based on U-Net Network Structure . . . . .</b>	549
Yuan Fang, Guoli Wang, Xianhua Dai, and Xuemei Guo	

<b>Robust Monocular Visual-Inertial SLAM Using Nonlinear Optimization .....</b>	560
Jingyun Duo, Lei Ji, and Long Zhao	
<b>Research on Aerodynamically Assisted Orbit Maneuver Method Based on Feature Model Correction.....</b>	569
Yue Lin, Yingmin Jia, and Songtao Fan	
<b>Fault Estimation of Switched Linear Systems with Actuator and Sensor Faults.....</b>	579
Chunying Su, Jianting Lyu, Xin Wang, and Dai Gao	
<b>Deep Convolutional Neural Network for Real and Fake Face Discrimination .....</b>	590
Yuanyuan Li, Jun Meng, Yaqin Luo, Xinghua Huang, Guanqiu Qi, and Zhiqin Zhu	
<b>Control Design for One-Sided Lipschitz Nonlinear Systems with Actuator Saturation .....</b>	599
Lin Yang, Jun Huang, and Haoran Zhang	
<b>Stochastic Stability of Itô Stochastic Systems with Semi-Markov Jump .....</b>	608
Min Zhang, Jun Huang, and Haoran Zhang	
<b>Agility Detector Designed for Automobile Detection.....</b>	615
Shuang Liu, Xizhong Shen, and Rongfan Leo	
<b>Review of Model Predictive Control Methods for Time-Delay Systems .....</b>	624
Lei Liu, Yi He, and Cunwu Han	
<b>Resistivity Inversion Solving Based on a GA Optimized Convolutional Neural Network .....</b>	634
Peng Wang and Shurong Li	
<b>GECNN-CRF for Prostate Cancer Detection with WSI .....</b>	646
Jinfeng Dong, Xuemei Guo, and Guoli Wang	
<b>Design Method of Robot Welding Workstation Based on Adaptive Planing.....</b>	659
Haofei Dai, Zhaojiang Liu, Yizhong Luan, Jiyang Chen, Wenxu Sun, and Sile Ma	
<b>Road Intersection Path Planning Based on Q-learning for Unmanned Ground Vehicle .....</b>	669
Lingxue Zhao, Chaofang Hu, Yao Guo, and Patrick Tjan	
<b>Univariate ReLU Neural Network and Its Application in Nonlinear System Identification .....</b>	679
Xinglong Liang and Jun Xu	

<b>Development of Multiply Magnetic Field Generator Combined with Living Cell Workstation . . . . .</b>	688
Jiansheng Xu, Chuanfang Chen, Deyu Kong, Linfei Ye, and Ming Xu	
<b>TIO Loss: A Transplantable Inversed One-Hot Loss for Imbalanced Multi-classification . . . . .</b>	696
Lin Wang and Chaoli Wang	
<b>Image Classification Method Based on Generative Adversarial Network . . . . .</b>	708
Longhui Hu and Chaoli Wang	
<b>Design and Development of Integrated Device for Wireless Detection of Flue Gas in Cremation Equipment . . . . .</b>	721
Fengguang Huang, Lin Tian, and Wei Wang	
<b>A Hierarchical Fuzzy Comprehensive Evaluation Algorithm for Running States of a Mine Hoist Synchronous Motor Drive System (MHSS) . . . . .</b>	730
Wei Liu, Fuzhong Wang, Ao Hou, and Sumin Han	
<b>Disturbance Observer-Based Design and Analysis of Iterative Learning Control with Nonrepetitive Uncertainties . . . . .</b>	739
Zirong Guo and Deyuan Meng	
<b>On Scaled Consensus, Bipartite Consensus and Scaled Bipartite Consensus: A Unified Viewpoint . . . . .</b>	749
Yuxin Wu and Deyuan Meng	
<b>Voltage Balancing of Modular Multilevel Converter Based on Cerebellar Model Articulation Controller . . . . .</b>	760
Xiangsheng Liu, Yuanyuan Yang, Lin Ren, Zhengxin Zhou, Yunxia Jiang, Lailong Song, and ZhengLin Jiang	
<b>Consistency of Continuous Multi-agent Systems with Privacy Protection . . . . .</b>	769
Meiyang Yu, Hongyong Yang, Yujiao Sun, and Fei Liu	
<b>Short-Term Prediction of Photovoltaic Power Generation Based on Deep Belief Network with Momentum Factor . . . . .</b>	778
Lai Lei, Jiangzhen Guo, Fuzhong Wang, and Li Zhang	
<b>Improved Adaptive Filter with Unknown Process and Measurement Noise Covariance . . . . .</b>	792
Jirong Ma, Yumei Hu, Qinghua Ma, Shujun Yang, Jianqiang Zheng, and Shuaiwei Wang	
<b>Extended State Observer-Based Sliding Mode Control for Epilepsy . . . . .</b>	801
Wei Wei, Ping Li, and Min Zuo	

<b>Research on General System Level Training Simulation Technology of Mid-And High-End Military UAV . . . . .</b>	810
Qing Zhang, Jiahui Tong, and Haifeng Li	
<b>Event-Based Robust State and Fault Estimation for Stochastic Linear System with Missing Observations and Uncertainty . . . . .</b>	819
Zhidong Xu, Bo Ding, and Tianping Zhang	
<b>Sliding Mode Control for a Constant Force Suspension System . . . . .</b>	832
Yuxin Jia, Yingmin Jia, Kai Gong, Yao Lu, and Meng Duan	
<b>Real-Time Coverage Path Planning of a UAV with Threat and Value Zone Constraints . . . . .</b>	840
Yan Liu, Hao Li, and Zhi Liu	
<b>Author Index . . . . .</b>	849



# Distributed Finite-Time Rotating Encirclement Control for Second-Order Agent Dynamics

Yixin Li, Yingmin Jia<sup>(✉)</sup>, and Tengfei Zhang

The Seventh Research Division and the Center for Information and Control,  
School of Automation Science and Electrical Engineering,  
Beihang University (BUAA), Beijing 100191, China  
[liyixin96@163.com](mailto:liyixin96@163.com)

**Abstract.** Rotating encirclement control problem is a important part in coordinated control and are widely used. In this paper we will solve this problem for second-order agents in finite-time. We establish three estimators for each agent to gauge the geometric center, polar radius and polar angle firstly. The reference trajectory for each agent will be obtained by the estimated values, which will make all agents be uniformly distributed on the circumference at the same speed. Then we use the control law to make all agents follow a predetermined trajectory, which ensure that in finite time, the agents uniformly encircle the targets. The simulation results are also showed to prove the feasibility of the methods.

**Keywords:** Encirclement control · Finite-time stable · Distributed estimator · Multi-agent

## 1 Introduction

Recently, multi-agent technology has developed rapidly, and coordinated control has attracted a great deal of interest from a large number of experts and scholars. Coordinated control problems can be used for UAV formations, robot rescue and so on. Enclosing control of multi-agent system is an achievement in the coordinated control, which has great application value and significance.

D. V. Dimarogonas et al. proposed the enclosing control of multi-agent in [1]. A definition of multi-agent surrounding control is given in [4]. A method of finite-time rotation encirclement control for first-order systems with non convex input constraints is proposed in [2]. [5] solved the problem of surrounding control involving a group of leaders and followers. A control schemes for nonholonomic mobile robots is designed in [7] by only using the relative bearing measurements. Encirclement control for multiple leaders which are stationary or dynamic are investigated in [8]. Collective rotations for second-order multi-agents are investigated in [6] with the help of Lyapunov theory for complex systems. [3] proposed a method which aims to analyze finite time stability. [9,10] investigated the containment control under switching topology.

The organization of this paper is as follows: basic theory about graph theory, encirclement control, and finite-time stability is described in Sect. 2. Control design will be shown in Sect. 3. And the simulation of our method will be shown in Sect. 4. Our conclusion is shown in Sect. 5.

## 2 Preliminaries

In this section, basis theory of graph theory, stability in finite-time, encirclement control with second-order agent dynamics will be present.

### 2.1 Graph Theory

Supposed that  $G(\mathcal{V}, \mathcal{E}, \mathcal{A})$  is an undirected graph of order  $n$ ,  $\mathcal{V}$  is the set of nodes,  $\mathcal{E}$  is the set of edges, and  $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{n \times n}$  is the weighted adjacency matrix. If  $(i, j) \in \mathcal{E}$ ,  $a_{ij} = a_{ji} > 0$  and if  $(i, j) \notin \mathcal{E}$ ,  $a_{ij} = 0$ , for all  $i \in \mathcal{V}$   $a_{ii} = 0$ .

The neighbor set of node  $i$  is defined as  $\mathcal{N}_i = \{j \in \mathcal{N} : (i, j) \in \mathcal{E}\}$ . The graph Laplacian induced by the information flow  $G$  is defined as  $\mathcal{L}(G) = [l_{ij}]$ , where  $l_{ij} = -a_{ij}$  for all  $i \neq j$ ,  $l_{ii} = \sum_{j=1}^n l_{ij}$ .

### 2.2 Stability in Finite-Time

**Definition 1.** *The system can be considered as follows*

$$\dot{x} = f(x), x \in \mathbb{R}^m \quad (1)$$

If  $f_i(\epsilon^{r_1}x_1, \dots, \epsilon^{r_m}x_m) = \epsilon^{\sigma+r_a}f_a(x)$ ,  $\epsilon > 0$ ,  $i = 1, \dots, m$ , then we can called the continuous vector field  $f(x)$  is homogeneous, and  $\sigma \in \mathbb{R}$  is called degree,  $r = (r_1, \dots, r_k)$  is called dilation. If  $f(x)$  is homogeneous, then we can get that the system (1) is homogeneous.

**Lemma 1.** *Provided that in system (1),  $f$  is continuous and is homogeneous with degree  $\sigma$  and dilation  $(r_1, \dots, r_k)$ , its asymptotically stable equilibrium is  $x = 0$ . If degree  $\sigma < 0$ , then the equilibrium of system (1) can stabilize in finite-time.*

**Lemma 2.** *Provided that  $V: \mathbb{R}^n \rightarrow \mathbb{R}_+ \cup \{0\}$  is positive definite function with a continuous radially unbounded such that 1)  $V(x) = 0 \Rightarrow x \in M$  2)  $\dot{V}(x) \leq -(\alpha V^p(x) - \beta V^q(x))^b$ ,  $p > 0, q > 0$ ,  $\alpha > 0, \beta > 0, b = 1, pb < 1, qb > 1$ . then the system is finite-time stable, and  $\forall x \in \mathbb{R}^n$ ,  $T(x_0)$  represents the settling time and it satisfies  $T(x_0) \leq (1/\alpha^k(1-pk)) + (1/\beta^k(qk-1))$ .*

### 2.3 Encirclement Control with Second-Order Agent Dynamics

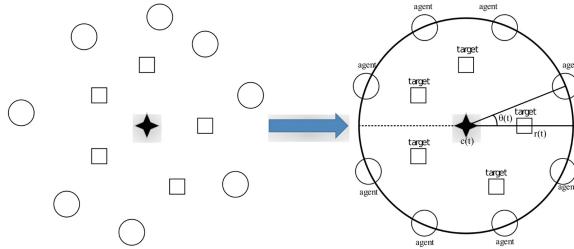
We consider each node of the graph is a dynamic agent

$$\dot{x}_i = v_i, \dot{v}_i = u_i, i \in \mathcal{V} \quad (2)$$

where  $x_i \in \mathbb{R}^2$  donates the position information of agent  $i$ ,  $v_i \in \mathbb{R}^2$  donates the velocity information of the  $i$ -th agent,  $u_i \in \mathbb{R}^2$  donates the control input of the  $i$ -th agent. In order to express this problem more directly, we introduce the polar coordinate transformation, which can be expressed:

$$x_i(t) = c(t) + [r_i(t)\cos\theta_i(t), r_i(t)\sin\theta_i(t)]^T \quad (3)$$

where  $c(t)$  donates the geometric center of all targets,  $p_a(t)$  denotes the position of the  $a$ -th target,  $c(t) = \frac{1}{m} \sum_{a=1}^m p_a(t) \in \mathbb{R}^2$ .  $r_i(t)$  donates the polar radius, which means the distance from  $c(t)$  to  $p_a(t)$ .  $\theta_i(t)$  donates the polar angle, in rectangular coordinates, it is expressed as the angle between the positive direction of the X axis and the connection between the  $p_a(t)$  and  $c(t)$ . In this paper, we can't get the specific information of the  $c(t)$ ,  $r_i(t)$  and  $\theta_i(t)$ , so we need to estimate these values only by the neighbor's information. We give an example to show the polar coordinate expression more vividly in Fig. 1.



**Fig. 1.** Encirclement control problem

**Definition 2.** For all  $i \in \mathcal{V}$ , if system (2) and (3) meet conditions (4) by a control law  $u_i(t)$  in a finite time  $T > 0$ , then the system can achieve finite-time rotating encirclement control.

$$\begin{cases} \theta_i(T) - \theta_j(T) - \frac{2\pi(i-j)}{n} = 0 \\ r_i(T) - kl(T) = 0 \\ \dot{\theta}_i(T) - \delta = 0 \end{cases} \quad (4)$$

where  $l(t) = \max_{a \in \mathcal{T}} \{\|p_a(t) - c(t)\|\}$ ,  $k > 1$ , the desired angular velocity is represented by  $\delta$ .

**Assumption 1.**  $l(t) \leq l^*$ ,  $l^*$  is a constant. The speeds of all targets satisfies  $\|\dot{p}_a(t)\| \leq p^*$ ,  $a \in \mathcal{T}$ .

## 2.4 Notations

In this paper, we consider the encirclement problem with  $n$  agents and  $m$  targets.  $\mathcal{N}_j^{\mathcal{V}} \subseteq \mathcal{V}$  is a set of agents, whose elements can get the information of target  $j$ .  $|\mathcal{N}_i^{\mathcal{T}}|$  donates the quantity of elements in  $\mathcal{N}_j^{\mathcal{V}}$ .  $\mathcal{N}_i^{\mathcal{T}} \subseteq \mathcal{T}$  donates the set of targets whose information can be obtained by the agent  $i$ .  $\|x\|$  donates the Euclidean norm, where  $x \in \mathbb{R}^n$  is a real vector.

**Assumption 2.** *The graph is connected in this paper.  $|\mathcal{N}_i^{\mathcal{T}}| \geq 1, \forall j \in \mathcal{T}$ .*

## 3 Main Results

In this section, we design three estimators for  $c_i(t)$ ,  $r_i(t)$ ,  $\theta_i(t)$ . We will design a reference trajectory for by these estimated value. Stability analysis is also given.

### 3.1 Estimation of the Geometric Center( $c(t)$ )

In this part, we will establish the distributed estimator for all agents to obtain the estimated geometric center  $\tilde{c}_i(t)$ .

$$\begin{cases} \eta_{ij}(t) = \tilde{c}_i(t) - \tilde{c}_j(t), j \in \mathcal{N}_i \\ \dot{\phi}_i(t) = -\alpha_{i1} \sum_{j \in \mathcal{N}_i} \frac{\eta_{ij}(t)}{\|\eta_{ij}(t)\|} - \alpha_{i2} \sum_{j \in \mathcal{N}_i} \eta_{ij}(t) \|\eta_{ij}(t)\| \\ \tilde{c}_i(t) = \phi_i(t) + \frac{n}{m} \sum_{j \in \mathcal{N}_i^T} \frac{1}{|\mathcal{N}_j^{\mathcal{V}}|} p_j(t) \end{cases} \quad (5)$$

where  $\phi_i(t) \in \mathbb{R}^2$  is a vector,  $\phi_i(0) = 0$ .  $\alpha_{i1}, \alpha_{i2}$  are positive design parameters. Using estimators (5), we analyze the finite-time stability of the estimation of the geometric center.

**Lemma 3.** *Using Assumptions 1. We choose that  $\alpha_{i1} > (n-1)p^*$ ,  $\alpha_{i2} > 0, \forall i \in \mathcal{V}$ ,  $\tilde{c}(t)$  will stabilize to  $c(t)$  in finite-time.*

*Proof.* Define  $\bar{c}(t) = \frac{1}{n} \sum_{k=1}^n \tilde{c}_k(t)$ , we choose a Lyapunov function.

$$V_1(t) = \frac{1}{2} \sum_{i=1}^n \left\{ [\tilde{c}_i(t) - \bar{c}(t)]^T [\tilde{c}_i(t) - \bar{c}(t)] \right\} \leq \frac{n}{2} p(t)^2 \quad (6)$$

where  $p(t) = \max_{i,j \in \mathcal{V}} \{\|\tilde{c}_i(t) - \tilde{c}_j(t)\|\}$ .

$$\begin{aligned} \dot{V}_1(t) = & \sum_{i=1}^n \left\{ [\tilde{c}_i(t) - \bar{c}(t)]^T \left[ -\alpha_{i1} \sum_{j \in \mathcal{N}_i} \frac{\eta_{ij}(t)}{\|\eta_{ij}(t)\|} \right. \right. \\ & \left. \left. - \alpha_{i2} \sum_{j \in \mathcal{N}_i} \eta_{ij}(t) \|\eta_{ij}(t)\| + \frac{n}{m} \sum_{j \in \mathcal{N}_i^T} \frac{1}{|\mathcal{N}_j^{\mathcal{V}}|} \dot{p}_j(t) - \dot{\bar{c}}(t) \right] \right\} \end{aligned} \quad (7)$$

Using Assumption 1, Assumption 2, and Lemma 2, we can get that

$$\begin{aligned}
& -\alpha_{i1} \sum_{i=1}^n \left\{ [\tilde{c}_i(t) - \bar{c}(t)]^T \sum_{j \in \mathcal{N}_i} \frac{\eta_{ij}(t)}{\|\eta_{ij}(t)\|} \right\} \\
& = -\frac{\alpha_{i1}}{2} \sum_{i=1}^n \sum_{j \in \mathcal{N}_i} [\tilde{c}_i(t) - \bar{c}(t)]^T \frac{\eta_{ij}(t)}{\|\eta_{ij}(t)\|} - \frac{\alpha_{i1}}{2} \sum_{i=1}^n \sum_{j \in \mathcal{N}_i} [\tilde{c}_i(t) - \bar{c}(t)]^T \frac{\eta_{ij}(t)}{\|\eta_{ij}(t)\|} \\
& = -\frac{\alpha_{i1}}{2} \sum_{i=1}^n \sum_{j \in \mathcal{N}_i} [\tilde{c}_i(t) - \bar{c}(t)]^T \frac{\eta_{ij}(t)}{\|\eta_{ij}(t)\|} + \frac{\alpha_{i1}}{2} \sum_{i=1}^n \sum_{j \in \mathcal{N}_i} [\tilde{c}_j(t) - \bar{c}(t)]^T \frac{\eta_{ij}(t)}{\|\eta_{ij}(t)\|} \\
& = -\frac{\alpha_{i1}}{2} \sum_{i=1}^n \sum_{j \in \mathcal{N}_i} \frac{\eta_{ij}(t)^T \eta_{ij}(t)}{\|\eta_{ij}(t)\|} \leq -\alpha_{i1} s(t)
\end{aligned} \tag{8}$$

$$-\alpha_{i2} \sum_{i=1}^n \left\{ [\tilde{c}_i(t) - \bar{c}(t)]^T \sum_{j \in \mathcal{N}_i} \eta_{ij}(t) \|\eta_{ij}(t)\| \right\} \leq \frac{-4\alpha_{i2}}{n^4} s(t)^3 \tag{9}$$

$$\frac{n}{m} \sum_{i=1}^n \left\{ [\tilde{c}_i(t) - \bar{c}(t)]^T \sum_{j \in \mathcal{N}_i^T} \frac{1}{|\mathcal{N}_j^T|} \dot{p}_j(t) \right\} \leq \frac{(n-1)}{m} \max_{i,j \in \mathcal{V}} (\|\eta_{ij}(t)\|) \sum_{i=1}^n \sum_{j \in \mathcal{N}_i^T} \frac{p^*}{|\mathcal{N}_j^T|} = (n-1)p^* s(t) \tag{10}$$

According to Eqs. (8)–(10),

$$\dot{V}_1(t) \leq -[\alpha_{i1} - (n-1)p^*] p(t) - \frac{4\alpha_{i2}}{n^4} p(t)^3 \leq -\frac{\sqrt{2}(\alpha_{i1} - (n-1)p^*)}{\sqrt{n}} V_1^{\frac{1}{2}} - \frac{2\sqrt{2}\alpha_{i2}}{n^5\sqrt{n}} V_1^{\frac{3}{2}} \tag{11}$$

According to the Lemma 2, we can get that the estimated geometric center  $\tilde{c}_i(t)$  will converge to the real geometric center  $c(t)$  with a finite-time  $T_1$ , satisfying that  $T_1 \leq \frac{\sqrt{2n}}{\alpha_{i1} - (n-1)r_d^*} + \frac{n^5\sqrt{2n}}{2\alpha_{i2}}$ .

### 3.2 Estimation of the Polar Radius $r_i(t)$

In this part, we will establish the distributed estimator for all agents to obtain  $\tilde{r}_i(t)$  the polar radius which means the estimated polar radius.

$$\begin{aligned}
\dot{\tilde{r}}_{i1}(t) & = -\beta_{i1} \frac{\zeta_{i1}(t)}{|\zeta_{i1}(t)|} - \beta_{i2} \zeta_{i1}(t) |\zeta_{i1}(t)|, \zeta_{i1}(t) = \tilde{r}_{i1}(t) - \max_{j \in \mathcal{N}_i^T} (z_{ij}(t)) \\
\dot{\tilde{r}}_{i2}(t) & = -\beta_{i1} \frac{\zeta_{i2}(t)}{|\zeta_{i2}(t)|} - \beta_{i2} \zeta_{i2}(t) |\zeta_{i2}(t)|, \zeta_{i2}(t) = \tilde{r}_{i2}(t) - \max_{j \in \mathcal{N}_i \cup \{i\}} (\tilde{r}_{j1}(t)) \\
& \vdots \\
\dot{\tilde{r}}_{iD}(t) & = -\beta_{i1} \frac{\zeta_{iD}(t)}{|\zeta_{iD}(t)|} - \beta_{i2} \zeta_{iD}(t) |\zeta_{iD}(t)|, \zeta_{iD}(t) = \tilde{r}_{iD}(t) - \max_{j \in \mathcal{N}_i \cup \{i\}} (\tilde{r}_{j(D-1)}(t)) \\
\tilde{r}_i(t) & = \kappa \tilde{r}_{iD}(t)
\end{aligned} \tag{12}$$

where  $\beta_{i1}, \beta_{i2}$  are positive design parameters.  $z_{ij}(t) = \|p_j(t) - \tilde{c}_i(t)\|, D = \max d(i, j)$  Using a Lyapunov function and the Lemma 2, we can get that the estimated polar radius  $\tilde{r}_i(t)$  will converge to the expected polar radius  $\theta_i(t)$  with a finite-time  $T_2$ , satisfying that  $T_2 \leq T_1 + \frac{\sqrt{2}M}{2\beta_{i2}} + \frac{\sqrt{2}M}{(\beta_{i1}-2p^*)}$ .

### 3.3 Estimation of the Polar Angle $\theta_i(t)$

In this part, we will establish the distributed estimator to get  $\tilde{\theta}_i(t)$  for all agents.

$$\begin{cases} \dot{\tilde{\theta}}_i(t) = -\gamma_{i1} \sum_{j \in \mathcal{N}_i} a_{ij} \xi_{ij}(t) |\xi_{ij}(t)| + \delta(t) - \gamma_{i2} \sum_{j \in \mathcal{N}_i} a_{ij} \frac{\xi_{ij}(t)}{|\xi_{ij}(t)|} \\ \xi_{ij}(t) = -\frac{2\pi(i-j)}{n} + \tilde{\theta}_i(t) - \tilde{\theta}_j(t) \end{cases} \quad (13)$$

where  $\gamma_{i1}, \gamma_{i2}$  are positive design parameters. Using a Lyapunov function and the Lemma 2, we can get that the estimated polar angle  $\tilde{\theta}_i(t)$  will converge to the expected polar angle  $\theta_i(t)$  with a finite-time  $T_3$ , satisfying that  $T_3 \leq \frac{4\sqrt{2}}{\gamma_{i1}\sqrt{\lambda_2(\mathcal{L}(\mathcal{A}_1))}} + \frac{2\sqrt{2}n}{\gamma_{i2}\lambda_2(\mathcal{L}(\mathcal{A}_2))^{\frac{3}{2}}}$ , where  $\mathcal{A}_1 = [(a_{ij})^2] \in \mathbb{R}^{n \times n}, \mathcal{A}_2 = [(a_{ij})^{\frac{2}{3}}] \in \mathbb{R}^{n \times n}$ .

### 3.4 Design of Control Algorithm

In this part, we will introduce the control algorithm which uses the estimated value. According to Eq. (3), we can get

$$\tilde{x}_i(t) = \tilde{c}(t) + [\tilde{r}_i(t) \cos \tilde{\theta}_i(t), \tilde{r}_i(t) \sin \tilde{\theta}_i(t)]^T \quad (14)$$

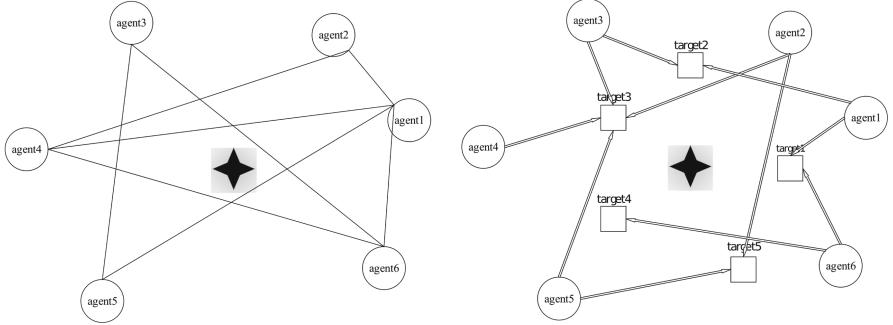
According to the reference trajectory, we design the control law (15)

$$u_i = \mu_1 \text{sig}((x_i - \tilde{x}_i)^{\tau_1} + \sum_{j=1}^n a_{ij} (\mu_2 \text{sig}((v_j - v_i)^{\tau_2})) \quad (15)$$

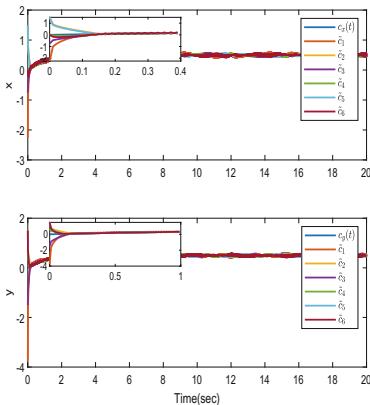
where  $\text{sig}(\omega)^n = |\omega|^n \text{sgn}(\omega), n > 0, \mu_1, \mu_2, \tau_1, \tau_2$  are positive design parameters. Using a Lyapunov function and the Lemma 1, we can get that the real trajectory will converge to the reference trajectory in finite-time.

## 4 Simulation

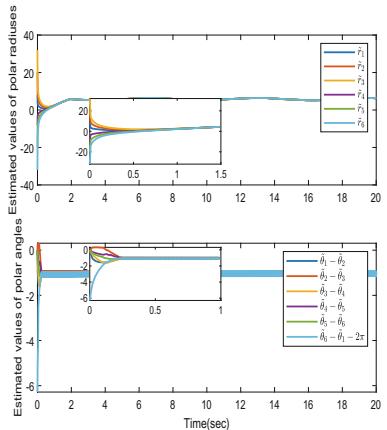
In this section, the result of simulation will be showed to prove the effectiveness of the control algorithm. The topology between and targets and agents of a multi-agent system which we use in this paper is shown in Fig. 2.



**Fig. 2.** Communication topology between targets and agents

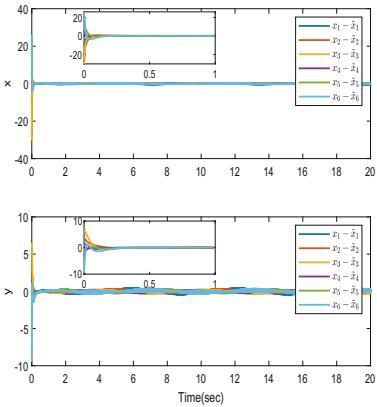


**Fig. 3.** Estimation of the geometric center

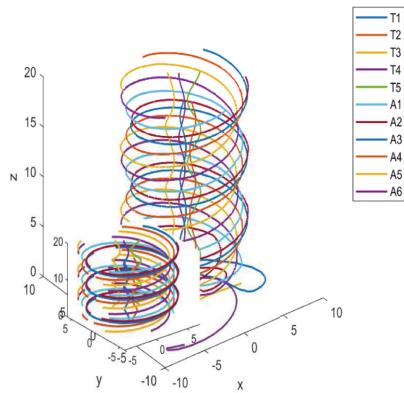


**Fig. 4.** Estimation of the polar radius and polar angle

Figure 3 shows that estimators can estimate the geometric center accurately in finite-time. From Fig. 4, we can easily find that in finite time, the polar radius and the polar angle also can converges to expectation. All agents will move following the reference trajectory in finite-time, which is shown in Fig. 5. The trajectory of all targets and agents are shown in Fig. 6.



**Fig. 5.** Estimation of the reference trajectory



**Fig. 6.** Movement of the targets and agents

## 5 Conclusion

The finite-time rotating encirclement control problem for second-order agent dynamics is studied. Firstly, we establish three estimators to gauge the geometric center, polar radius and polar angle. Secondly, we use the estimated value to design the reference trajectory for each agent. Thirdly, we design a control law for the system. Lastly, the simulation illustrates that the algorithm is effective.

## References

- Dimarogonas, D.V., Egerstedt, M., Kyriakopoulos, K.J.: A leader-based containment control strategy for multiple unicycles. In: Proceedings of the 45th IEEE Conference on Decision and Control, pp. 5968–5973 (2006). <https://doi.org/10.1109/CDC.2006.376700>
- Zhang, T., Ling, J., Mo, L.: Distributed finite-time rotating encirclement control of multiagent systems with nonconvex input constraints. IEEE Access **7**, 102477–102486 (2019). <https://doi.org/10.1109/ACCESS.2019.2930869>
- Wang, X., Hong, Y.: Finite-time consensus for multi-agent networks with second-order agent dynamics. IFAC Proc. Vol. **41**(2), 15185–15190 (2008). <https://doi.org/10.3182/20080706-5-kr-1001.02568>
- Ji, M., Ferrari-Trecate, G., Egerstedt, M., Buffa, A.: Containment control in mobile networks. IEEE Trans. Autom. Control **53**(8), 1972–1975 (2008). <https://doi.org/10.1109/TAC.2008.930098>
- Chen, F., Ren, W., Cao, Y.: Surrounding control in cooperative agent networks. Syst. Control Lett. **59**(11), 704–712 (2010). <https://doi.org/10.1016/j.sysconle.2010.08.006>
- Lin, P., Jia, Y.: Distributed rotating formation control of multi-agent systems. Syst. Control Lett. **59**(11), 587–595 (2010). <https://doi.org/10.1016/j.sysconle.2010.06.015>

7. Zheng, R., Liu, Y., Sun, D.: Enclosing a target by nonholonomic mobile robots with bearing-only measurements. *Automatica* **53**, 400–407 (2015). <https://doi.org/10.1016/j.automatica.2015.01.014>
8. Liu, H., Xie, G., Wang, L.: Necessary and sufficient conditions for containment control of networked multi-agent systems. *Automatica* **48**(7), 1415–1422 (2012). <https://doi.org/10.1016/j.automatica.2012.05.010>
9. Wang, F., Ni, Y., Liu, Z., Chen, Z.: Containment control for general second-order multiagent systems with switched dynamics. *IEEE Trans. Cybern.* **50**(2), 550–560 (2020). <https://doi.org/10.1109/TCYB.2018.2869706>
10. Guo, J., Yan, G., Lin, Z.: Local control strategy for moving-target-enclosing under dynamically changing network topology. *Syst. Control Lett.* **59**(10), 654–661 (2010). <https://doi.org/10.1016/j.sysconle.2010.07.010>



# Urban Road Object Detection and Tracking Applications Based on Acoustic Localization

Zhimin Wang<sup>1(✉)</sup>, Chaoli Wang<sup>1</sup>, and Song Shen<sup>1,2</sup>

<sup>1</sup> University of Shanghai for Science and Technology, Shanghai 200093, China  
934909661@qq.com

<sup>2</sup> China Orient Institute of Noise & Vibration, Beijing 100085, China

**Abstract.** The detection and tracking of urban road traffic videos based on acoustic positioning is used to detect and track target vehicles in urban road environments of this paper. In order to speed up the overall detection and tracking operation rate, the innovation of this paper is the combination of acoustics and images, which can achieve the effect of real-time detection on ordinary hardware. This paper is generally divided into three modules: acoustic positioning, target detection, and target tracking. This article can realize the positioning, classification, detection and tracking of the target vehicle, and provide an effective basis for the traffic management department to timely and effectively grasp the abnormal situation of urban roads.

**Keywords:** Acoustic positioning · Object detection · Object tracking · Urban road

## 1 Introduction

With the improvement of people's living standards, the number of cars in our country has increased year by year. This has also become one of the important reasons that lead to traffic jams and frequent accidents. Rescue and deal with traffic vehicles that have serious violations of the law. An algorithm for real-time detection of vehicle traffic has been widely concerned and studied by scholars at home and abroad. With the development of artificial intelligence and the improvement of computer computing power, deep learning is widely used in computer vision. The field has been further developed, especially in the fields of target detection and target tracking with more extensive research and development [1]. Because of the limitations in the current hardware conditions and the many functions of the programs that the hardware needs to execute at the same time, most of the current target detection algorithms cannot achieve the effect of real-time detection on ordinary hardware, and each program itself consumes certain hardware memory resources. Object detection is another program that occupies a lot of memory resources. When the program runs, it seriously affects

the operation of other programs in the project, causing the entire project to fail to run normally. Therefore, in order to solve the problem of excessive utilization of hardware memory resources occupied by the target detection algorithm program, a new detection mode is proposed. The image and acoustic positioning are combined, and the vehicle is detected and tracked after accurate acoustic positioning to avoid redundant detection. The detection time is shortened, so that the algorithm can realize the real-time detection and tracking function while ensuring the accuracy on ordinary hardware.

## 2 Research Status

Noise source identification methods based on microphone arrays and beamforming have been greatly developed in recent years, and have been increasingly applied to the sound source localization of aircraft, train, automobile and other equipment [2]. The microphone array is composed of several microphones arranged according to a certain geometry, and simultaneously measures multiple signals with different phase delays, and then uses beamforming to suppress noise and strengthen the sound source signal to achieve the purpose of spatial sound source positioning. Because it can measure transient sound sources or mobile sound sources, it is called ‘acoustic photography’ technology, which is mainly used in radar, sonar, and communication fields [3], and then gradually applied to speech recognition and sound source positioning etc. With the continuous improvement of computing power of computer hardware, artificial intelligence has been further developed and researched. In 2014, foreign scholar Ross Girshick and others applied deep learning convolutional networks to the field of target detection for the first time, and achieved a significantly improved accuracy than traditional target detection algorithms [4]. The target detection and tracking technology has been further developed, which can be used to help urban road traffic to perform monitoring, identification, and other functions, and to accurately grasp the situation of traffic vehicles in real time. Although the target detection algorithm can theoretically achieve the effect of real-time detection, the author of the article YOLOv3 [5] studied the results of the program running on a high-performance GPU such as TianX. However, it is still difficult to achieve fast real-time detection in practical applications. The reason that the real-time detection cannot be achieved is because of too many network layers, global detection redundancy, and hardware condition limitations. Among them, the biggest problem is detection redundancy. The conventional method is to manually select the required detection and tracking in the video frame. The target object requires a lot of manpower, and the process is tedious. This article makes the following innovations and improvements based on actual requirements: in combination with acoustic positioning, directly locates the target that needs to be detected and tracked, and transmits its related parameters to the target detection module. The target detection directly detects the target. In this way, the target detection can reduce redundant detection, thereby greatly reducing the algorithm running time, and transmitting the correct image to the target tracking module for real-time target tracking.

### 3 Research Process

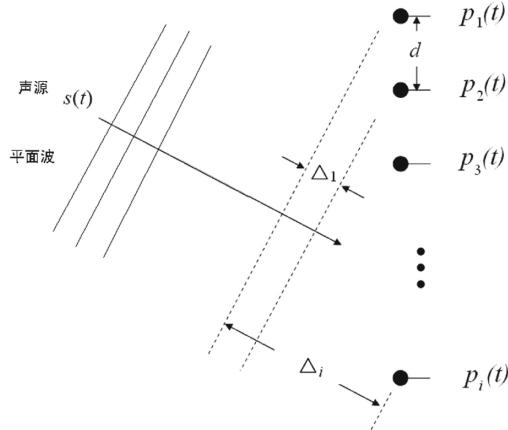
#### 3.1 Acoustic Position

Acoustic positioning is a passive positioning technology that can be applied in the fields of traffic, sound source tracking and speech enhancement. The working principle is that the microphone array collects sound signals and converts the analog sound signals into digital signals through an analog-to-digital converter. The beam forming method is used. The beam forming method in noise source identification is based on the microphone array. Accepting the time difference of the sound waves and the relative position difference between the individual microphones, the orientation of the sound source can be determined. In the calculation process, a certain beamforming algorithm is needed. The most basic method is Delay-Sum [2]. As shown in Fig. 1, the signals measured on the array elements in the microphone array are weighted, delayed, The operations such as summation make the signal delay caused by the position difference of each array element can be compensated, forming a result with spatial directivity, indicating the orientation of the sound source. Beamforming is equivalent to a spatial filter, which can enhance signals in a specified direction and filter out interference signals in other directions. The time domain calculation formula is as follows:

$$b(\bar{r}, t) = \frac{1}{N} \sum_{i=1}^N w_i p_i(t - \Delta_i(\bar{r})) \quad (1)$$

Where,  $N$  is the number of microphones,  $w_i$  is the weighting coefficient of the  $i$  microphones. If all microphones are in an equal position, it can be simply taken as  $w_i = 1$ ,  $p_i(t)$  is the sound pressure signal measured by the first microphones at the moment, and  $\Delta_i(\bar{r})$  is the delay of the first microphones when  $i$  calculated in the specified direction, which  $\bar{r}$  is determined by the position vector from the microphones to the specified direction. In the actual measurement, the result calculated by the above formula with the direction of the actual sound source as the specified direction will be the maximum.

This article uses a 32-channel spiral array of acoustic array sensor equipment, equipped with MEMS sensors suitable for outdoor and waterproof performance. Generally, the engine frequency of automobile vehicles is generally 200 Hz–500 Hz, and the whistle frequency is 300 Hz–4000 Hz. In order to avoid spatial aliasing [6], according to the Shannon sampling theorem formula, the spiral spacing should be no less than 4.25 cm. The advantage of performing acoustic positioning first is that after the target vehicle is successfully positioned, it can save the time of detecting each frame of video for the next target detection module, and the coordinate position of the positioning can reduce redundant detection of target detection. Detects objects, thereby further reducing runtime. The acoustic positioning video in this article is provided by the Beijing Oriental Institute of Vibration and Noise Technology, Chengdu Transportation 2 s video library, which is dedicated to the study of national traffic road vehicles.



**Fig. 1.** Delay and sum beamformere

### 3.2 Object Detection

The deep learning network used in this paper is YOLOv3 [4]. YOLOv3 adopts a new network architecture Darknet-53. The YOLOv3 network consists of three parts: the input layer, the convolution layer, and the output layer. The network structure is shown in Fig. 2. The input layer in this paper is the original frame picture after acoustic localization. The image is cut into a size of  $256 * 256$  pixels and input into the network. The convolution layer uses  $3 * 3$  and  $1 * 1$  convolution kernels to eliminate the pooling layer. The size of the tensor is achieved by changing the step size of the convolution kernel. If the step size is set to 2, after convolution, Image side length is reduced by half. Use Leaky Relu as the activation function. To prevent overfitting and the disappearance of gradients, residual layers are used to make up after every two convolutional layers. The residual layer is shown in Fig. 3. The output layer is a tensor output from  $8 * 8 * 1024$  after the convolution layer and the residual block. A global pooling is performed for each  $8 * 8$  feature map (take one (Mean) to get 1024 scalars, then enter a 1000 nodes, and finally classify the final result by SoftMax. The final result used is not only the category label, but also the center coordinates  $x$ ,  $y$  and length and width  $w$ ,  $h$  of the object framed by a rectangle. Record the acoustically located object's coordinate position in the frame picture and send it to the convolution network with the video stream. The convolution network before the improvement requires the global detection of each object for each frame of the video stream. It takes a long time, which is 10 times longer than the original video, and finally needs to artificially pick out the required object. The improved convolutional network only needs to detect the object at the coordinate position [12] of the frame of the picture after successful acoustic localization, identify the object type, select the object with a rectangular frame, and record its more accurate coordinate position to provide more accurate coordinate parameters for target tracking. This reduces the running time, gets rid

Type	Filters	Size	Output
1x	Convolutional	32	$3 \times 3$
	Convolutional	64	$3 \times 3 / 2$
	Convolutional	32	$1 \times 1$
	Convolutional	64	$3 \times 3$
2x	Residual		$128 \times 128$
	Convolutional	128	$3 \times 3 / 2$
	Convolutional	64	$1 \times 1$
	Convolutional	128	$3 \times 3$
8x	Residual		$64 \times 64$
	Convolutional	256	$3 \times 3 / 2$
	Convolutional	128	$1 \times 1$
	Convolutional	256	$3 \times 3$
8x	Residual		$32 \times 32$
	Convolutional	512	$3 \times 3 / 2$
	Convolutional	256	$1 \times 1$
	Convolutional	512	$3 \times 3$
4x	Residual		$16 \times 16$
	Convolutional	1024	$3 \times 3 / 2$
	Convolutional	512	$1 \times 1$
	Convolutional	1024	$3 \times 3$
	Residual		$8 \times 8$
	Avgpool		Global
	Connected		1000
	Softmax		

**Fig. 2.** Network structure diagram

of the constraints of high hardware conditions, and does not need to manually pick out the required frame pictures, which greatly improves the work efficiency and achieves the effect of real-time detection.

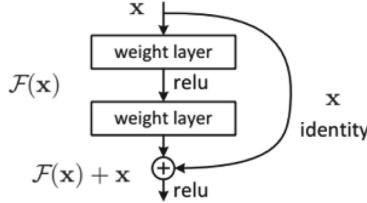
### 3.3 Object Tracking

This paper uses four common opencv tracking algorithms, and uses 2 s urban road traffic video as an example for comparison. Table 1 is a comparison of four tracking algorithms for testing.

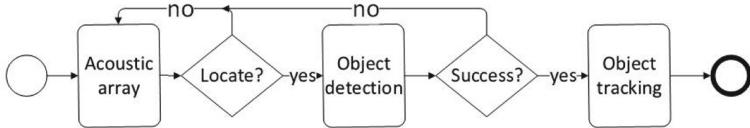
Through the experimental comparison of time and accuracy and the requirements of work efficiency, this paper selects KCF as the object detection algorithm. The coordinates of the rectangular frame and the current frame number after the object detection are successfully used as the input for the original tracking image. Tracking can track the position of the object in the full frame.

### 3.4 Detection Process

The overall detection process is shown in Fig. 4. The whole process of detection [11] is that the video stream after acoustic localization successfully sends the

**Fig. 3.** Residual block**Table 1.** Comparison of four tracking algorithms

Year	Time (s)	Accuracy [8]
MIL [7]	4.804	47.5%
KCF [8]	1.669	73.2%
TLD [9]	3.126	60.8%
MOSSE [10]	1.648	43.1%

**Fig. 4.** Detection process

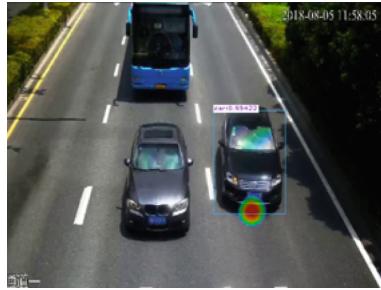
positioned picture and coordinate position to the object detection module. The object detection module quickly and accurately determines the type of the object and rectangles it out. Finally, the rectangular frame position and video stream are sent. To the object tracking module, real-time tracking of the object.

### 3.5 Analysis of Experimental Results

The experimental environment configuration of this article is as follows: Intel Core i5-8300H 8G CPU, software environment Visual Studio 2015 C++, operating system Win10. The detection source of the video and sound array in this article is provided by the Beijing Oriental Institute of Vibration and Noise Technology. The video of ChengDu traffic is selected for 2 s (2 s). In this paper, two functions of object detection and tracking are realized. The theoretical time is 4s. Due to the computer hardware conditions, if no acoustic positioning point is added, according to work requirements, the detection time is more than 10 times that of the original video. If the acoustic positioning point is added, the running time and video playback time are basically the same. As shown in Fig. 5. The small red dot in the picture is the landmark of the acoustic anchor point. Record the acoustic positioning coordinate points and the current frame number of successful positioning, and transfer the parameters to the YOLOv3 object



**Fig. 5.** Acoustic positioning



**Fig. 6.** YOLOv3 detection

**Table 2.** Comparison of acoustic positioning time

File index	Time (s)	No acoustic positioning (s)	Have acoustic positioning (s)
-1	2	26.053	2.86
-2	2	25.324	2.899
-3	2	25.055	2.782

detection module. At this time, YOLOv3 only needs to detect the current frame image of the successful positioning. After the classification detection is successful (see Fig. 6) and returns the bounding box  $x$ . The four coordinate values  $y$ ,  $w$ ,  $h$  and the parameters of the current frame are given to the object tracking program. For practical work efficiency, this paper chooses KCF as the object tracking algorithm. Table 2 is a comparison of the time with and without acoustic localization. Due to objective reasons such as hardware conditions, the time consumed by the non-acoustic positioning points is significantly greater than the time of the acoustic positioning points.

## 4 Summary

This paper proposes a new detection mode that combines acoustics and object detection. It solves the problem that current object detection cannot be detected

in real-time on ordinary hardware, greatly speeding up the running time, and timely and effectively grasping urban roads for traffic management departments. The situation provides an effective basis.

## References

1. Li, Y.P., Hou, L.Y., Wang, C.: Moving objects detection in automatic driving based on YOLOv3. *Comput. Eng. Des.* (4), 38 (2019)
2. Shen, S., Ying, H.Q., Liu, J.M.: The application and comparison of noise source identification technique based on beamforming. In: Abstracts of the Ninth National Conference on Vibration Theory and Application, p. 257 (2007)
3. Steys, H.: Digital beamforming basic. *J. Electron. Defense* **7**, 50–56 (1996)
4. Girshick, R., Donahue, J., Darrell, T.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014). <https://doi.org/10.1109/CVPR.2014.81>
5. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. In: Computer Vision and Pattern Recognition (CVPR), pp. 126–134 (2018)
6. Ying, H.Q., Dong, S.W., Liu, J.M., Ying, M.: The “wonderful” phenomenon of aliasing in signal processing. In: Proceedings of the 19th and 20th National Conference on High Technology and Application of Vibration And Noise, pp. 384–389 (2007)
7. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 983–990. IEEE (2009). <https://doi.org/10.1109/CVPR.2009.5206737>
8. Henriques, J.F., Caseiro, R., Martins, P.: High-speed tracking e kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(3), 583–596 (2014). <https://doi.org/10.1109/TPAMI.2014.2345390>
9. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(7), 1409–1422 (2012). <https://doi.org/10.1109/TPAMI.2011.239>
10. KBolme, D.S., Beveridge, J.R., Draper, B.A.: Visual object tracking using adaptive correlation filters. In: Computer Vision and Pattern Recognition (CVPR), pp. 2544–2550 (2010). <https://doi.org/10.1109/CVPR.2010.5539960>
11. Jia, Y.M.: Robust control with decoupling performance for steering and traction of 4WS vehicles under velocity-varying motion. *IEEE Trans. Control Syst. Technol.* **8**(3), 554–569 (2000). <https://doi.org/10.1109/87.845885>
12. Jia, Y.M.: Alternative proofs for improved LMI representations for the analysis and the design of continuous-time systems with polytopic type uncertainty: a predictive approach. *IEEE Trans. Autom. Control* **48**(8), 1413–1416 (2003). <https://doi.org/10.1109/TAC.2003.815033>



# Air Combat Situation Assessment of Multiple UCAVs with Incomplete Information

Shouyi Li, Qingxian Wu<sup>(✉)</sup>, Mou Chen, and Yuhui Wang

College of Automation Engineering,  
Nanjing University of Aeronautics and Astronautics, Nanjing 211100, China  
[wuqingxian@nuaa.edu.cn](mailto:wuqingxian@nuaa.edu.cn)

**Abstract.** An air combat situation assessment method is developed for multiple unmanned combat air vehicles (UCAVs) based on interval data in this paper. Due to the complexity and uncertainty of the air combat environment, it is difficult to obtain the accurate data of the air combat. Thus, interval number is considered to represent the inaccurate data. An air combat situation index system based on interval numbers is established, which consists of angle advantage, velocity advantage, height advantage, distance advantage and performance advantage. Then, an interval number eigenvector method is developed to give the final situation assessment. An example is given to illustrate the air combat situation assessment method.

**Keywords:** Air combat situation assessment · UCAV · Interval number · Interval number eigenvector method

## 1 Introduction

When unmanned combat air vehicles (UCAVs) is engaged in air combat, due to the complexity and diversity of air combat missions, a single UCAV has limited combat capabilities and it is difficult to complete air combat missions independently. Therefore, UCAVs mostly deploy forces and organize battles in a cluster manner to maximize the overall combat effectiveness. In air combat of multiple UCAVs, different UCAVs have different attack-defense abilities due to their different positions, different weapons and equipments. Therefore, a reasonable situation assessment of the multiple UCAVs can characterize the effects of different attack missions, which is the basis of firepower allocation for multiple UCAVs air combat.

In the existing literature, situation assessment methods include fuzzy comprehensive evaluation method [1], analytic hierarchy process [2], Bayes network [3], Topsis method [4], etc. A dynamic Bayesian network method was proposed in [5], which has better fault tolerance ability. Evidence theory was engaged in air combat situation assessment in [6], but it relied heavily on the rationality of evidence rules. Expert system was used for air combat threat sequencing in [7], but

it is difficult to handle the uncertainty in the actual air combat. In these methods, researchers use accurate air combat data for situation assessment. However, in a complex air combat environment, the air combat data of UCAVs measured by sensors is incomplete due to the different measurement results of various sensors and the measurement error of each sensor. Thus, the air combat situation assessment using accurate data are not in line with the actual situation. What is more, in air combat situation assessment, there are usually multiple assessment indicators, and researchers mainly uses subjective or objective weighting methods to determine the weights of various indicators, which leads to the weights be accurate values. In fact, different experts use different methods, which makes the index weights different and uncertain. Under the incompleteness of decision-making information caused by the complexity of decision-making environment, the irrationality of subjective preferences and emotional thinking of decision-makers establishes a kind of uncertainty problems, which can be well solved by interval numbers. Using interval numbers to describe complex uncertainty problems is more in line with the vague thinking habits of people. The interval analysis method of uncertainty problem has become a hot spot in the field of engineering technology, management decision-making and theory [8]. Especially in air combat decision-making, interval numbers can well describe the uncertainty of air combat information and the subjectivity of expert experience.

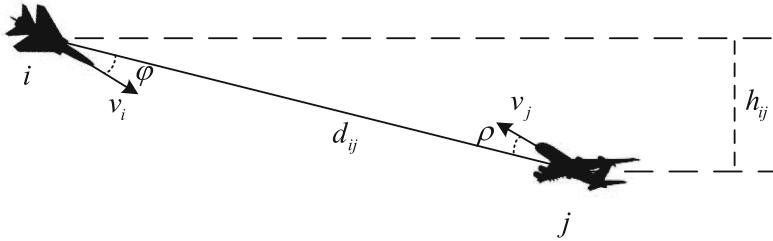
In this paper, we use the battlefield data in the form of interval numbers to establish a situation assessment system of angle advantage, velocity advantage, height advantage, distance advantage, and performance advantage. Then, we use interval number to determine the weight coefficients. Based on Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) method, the situation assessment results in the form of interval numbers are converted into accurate data, and the final situation assessment values are given.

The organization of this paper is as follows: In Sect. 2, the advantage index system is established. In Sect. 3, weight coefficients of different assessment indexes are determined based on the interval number eigenvector method, and the final situation assessment values are given. In Sect. 4, a numerical example is given to illustrate the rationality and effectiveness of the method.

## 2 Problem Statement and Preliminaries

Assuming that both side have multiple UCAVs, the air combat situation of one enemy UCAV and one our UCAV is given in Fig. 1 by borrowing the similar result in [11], which describes the geometric position relationship between the enemy UCAV and our UCAV, where  $i$  represents our UCAV and  $j$  represents the enemy UCAV,  $d_{ij}$  is the distance between the two UCAVs,  $v_i$  is the velocity of  $i$ ,  $v_j$  is the velocity of  $j$ ,  $h_{ij}$  is the height difference between  $i$  and  $j$ ,  $\varphi$  represents the azimuth of  $j$  with respect to  $i$ , that is, the angle between the speed vector and the line of sight,  $\rho$  represents the entry angle of  $j$  with respect to  $i$ , that is, the angle between the speed vector of  $j$  and the extension line of sight.

In a complex air combat environment, assessing air combat situations requires consideration of many factors, which includes not only the geometric positional



**Fig. 1.** Air combat situation of *i* and *j*.

relationship between the enemy UCAVs and our UCAVs, such as the distance, speed, altitude difference, etc., but also the air combat capabilities of the enemy UCAVs and our UCAVs [9]. In this paper, referring to [10] and [11], angle advantage, velocity advantage, height advantage, distance advantage and air combat performance advantage are considered.

Firstly, we give the advantage index system of *i* to *j* as follows.

(1) Angle advantage  $p_{aij}$ .

$$p_{aij} = 1 - (|\varphi| + |\rho|)/360^\circ. \quad (1)$$

(2) Velocity advantage  $p_{vij}$ .

$$p_{vij} = \begin{cases} 0.1 & v_i < 0.6v_j \\ v_i/v_j - 0.5 & 0.6v_j \leq v_i \leq 1.5v_j \\ 1.0 & 1.5v_j < v_i. \end{cases} \quad (2)$$

(3) Height advantage  $p_{hij}$ .

$$p_{hij} = \begin{cases} 0.1 & h_{ij} < -5 \\ 0.5 + 0.1h_{ij} & -5 \leq h_{ij} \leq 5 \\ 1.0 & h_{ij} > 5. \end{cases} \quad (3)$$

(4) Distance advantage  $p_{dij}$ .

Distance advantage is discussed based on the radar detection distance and missile attack distance of the two UVACs which are described as follows.

a) The performance of *j* is better than the performance of *i*:  $rm_i < rm_j < rr_i < rr_j$ , then, we have

$$p_{dij} = \begin{cases} 0 & d_{ij} \leq rm_i, \text{ or } rm_j \leq d_{ij} \leq rr_i, \text{ or } rr_j \leq d_{ij} \\ 0.4 \frac{rm_j - d_{ij}}{rm_j - rm_i} & rm_i < d_{ij} < rm_j \\ 0.2 \frac{rr_j - d_{ij}}{rr_j - rr_i} & rr_i < d_{ij} < rr_j. \end{cases} \quad (4)$$

where  $rr_i$  and  $rr_j$  are the radar detection distance of *i* and *j* respectively,  $rm_i$  and  $rm_j$  are the missile attack distance of *i* and *j* respectively.

- b) The performance of  $i$  is better than the performance of  $j$ :  $rm_j < rm_i < rr_j < rr_i$ , then, we obtain

$$p_{dij} = \begin{cases} 0 & d_{ij} \leq rm_j, \text{ or } rm_i \leq d_{ij} \leq rr_j, \text{ or } rr_i \leq d_{ij} \\ 0.4 \frac{rm_i - d_{ij}}{rm_i - rm_j} & rm_j < d_{ij} < rm_i \\ 0.2 \frac{rr_i - d_{ij}}{rr_i - rr_j} & rr_j < d_{ij} < rr_i. \end{cases} \quad (5)$$

(5) Performance advantage  $p_{cij}$ .

$$p_{cij} = \frac{2}{\pi} \arctan \left( 0.5 \left( \frac{rr_i}{rr_j} + \frac{rm_i}{rm_j} \right) \sqrt{\frac{lm_i}{lm_j}} \right), \quad (6)$$

where  $lm_i$  and  $lm_j$  are the numbers of missiles carried by  $i$  and  $j$  respectively.

Then, in order to analyse the inaccurate data, the following definitions of interval number and their arithmetics are needed.

**Definition 1** [12]. Denote  $\mathbb{R}$  as the real number field, and an interval number is  $\tilde{r} = [r^L, r^U]$ , where  $r^L \in \mathbb{R}$ ,  $r^U \in \mathbb{R}$ , satisfying  $r^L \leq r^U$ . If  $r^L = r^U$ ,  $\tilde{r}$  is a real number.

**Definition 2** [12]. Letting  $\tilde{r} = [r^L, r^U]$  and  $\tilde{s} = [s^L, s^U]$  be two interval numbers, their arithmetic operation are defined as:

$$\tilde{r} + \tilde{s} \triangleq [r^L + s^L, r^U + s^U],$$

$$\tilde{r} - \tilde{s} \triangleq [r^L - s^U, r^U - s^L],$$

$$\tilde{r} \times \tilde{s} \triangleq [\min\{r^L s^L, r^L s^U, r^U s^L, r^U s^U\}, \max\{r^L s^L, r^L s^U, r^U s^L, r^U s^U\}],$$

$$\tilde{r} \div \tilde{s} \triangleq [\min\{r^L / s^L, r^L / s^U, r^U / s^L, r^U / s^U\}, \max\{r^L / s^L, r^L / s^U, r^U / s^L, r^U / s^U\}],$$

and the measure between  $\tilde{r}$  and  $\tilde{s}$  is defined as:

$$m(\tilde{r}, \tilde{s}) \triangleq \sqrt{(r^L - s^L)^2 + (r^U - s^U)^2}.$$

Then, a method of comparing the size of two interval numbers based on the method in [13] is given.

**Definition 3.** Let  $\tilde{r} = [r^L, r^U]$  and  $\tilde{s} = [s^L, s^U]$  be two interval numbers, and denote  $t^L = \max\{r^L, s^L\} + 1$ ,  $t^U = \max\{r^U, s^U\} + 1$ ,  $\tilde{t} = [t^L, t^U]$ , then the size relationship between  $\tilde{r}$  and  $\tilde{s}$  is as follows:

$$\tilde{r} = \tilde{s} \iff m(\tilde{r}, \tilde{t}) = m(\tilde{s}, \tilde{t}),$$

$$\tilde{r} < \tilde{s} \iff m(\tilde{r}, \tilde{t}) > m(\tilde{s}, \tilde{t}),$$

$$\tilde{r} > \tilde{s} \iff m(\tilde{r}, \tilde{t}) < m(\tilde{s}, \tilde{t}).$$

So far, we have studied various factors in the air combat situation assessment. Our goal is to give the situation assessment result that integrates each evaluation indicator by giving the weight coefficient of each indicator with the help of expert experience. The situation assessment result is an accurate value, while the situation data and weight coefficients are inaccurate values in the form of interval numbers.

### 3 Air Combat Situation Assessment

In this section, an air combat situation assessment method with incomplete information is given. Firstly, the weight coefficients of the five indicators are given based on interval number eigenvector method [14], then the interval number results of air combat situation assessment are converted into accurate results based on an idea commonly used in multi-attribute decision-making. For the five indicators of the above air combat situation assessment, a comparison matrix of pairwise weights is given:

$$\tilde{\Omega} = \begin{pmatrix} & \text{Angle} & \text{Velocity} & \text{Height} & \text{Distance} & \text{Performance} \\ \text{Angle} & \tilde{\omega}_{11} & \tilde{\omega}_{12} & \tilde{\omega}_{13} & \tilde{\omega}_{14} & \tilde{\omega}_{15} \\ \text{Velocity} & \tilde{\omega}_{21} & \tilde{\omega}_{22} & \tilde{\omega}_{23} & \tilde{\omega}_{24} & \tilde{\omega}_{25} \\ \text{Height} & \tilde{\omega}_{31} & \tilde{\omega}_{32} & \tilde{\omega}_{33} & \tilde{\omega}_{34} & \tilde{\omega}_{35} \\ \text{Distance} & \tilde{\omega}_{41} & \tilde{\omega}_{42} & \tilde{\omega}_{43} & \tilde{\omega}_{44} & \tilde{\omega}_{45} \\ \text{Performance} & \tilde{\omega}_{51} & \tilde{\omega}_{52} & \tilde{\omega}_{53} & \tilde{\omega}_{54} & \tilde{\omega}_{55} \end{pmatrix}. \quad (7)$$

where  $\tilde{\omega}_{mn} = [\omega_{mn}^L, \omega_{mn}^U]$  ( $m, n = 1, 2, \dots, 5$ ) is an interval number. Denote  $\Omega^L = (\omega_{mn}^L)_{5 \times 5}$ ,  $\Omega^U = (\omega_{mn}^U)_{5 \times 5}$ , then the calculation steps for determining the weight by the interval number eigenvector method are as follows:

*Step 1.* Use the eigenvector method developed in [14] to determine the normalized eigenvectors  $x^L = (x_1^L, x_2^L, x_3^L, x_4^L, x_5^L)$  and  $x^U = (x_1^U, x_2^U, x_3^U, x_4^U, x_5^U)$  corresponding to the maximum eigenvalues of  $\Omega^L$  and  $\Omega^U$ , respectively.

*Step 2.* Calculate  $\alpha$  and  $\beta$ :

$$\alpha = \left( \sum_{n=1}^5 \frac{1}{\sum_{m=1}^5 \omega_{mn}^U} \right)^{1/2}, \quad \beta = \left( \sum_{n=1}^5 \frac{1}{\sum_{m=1}^5 \omega_{mn}^L} \right)^{1/2}. \quad (8)$$

*Step 3.* Calculate weight vectors for different indicators:

$$\tilde{\omega} = (\tilde{\omega}_1, \tilde{\omega}_2, \tilde{\omega}_3, \tilde{\omega}_4, \tilde{\omega}_5) \quad (9)$$

$$= ([\alpha x_1^L, \beta x_1^U], [\alpha x_2^L, \beta x_2^U], [\alpha x_3^L, \beta x_3^U], [\alpha x_4^L, \beta x_4^U], [\alpha x_5^L, \beta x_5^U]). \quad (10)$$

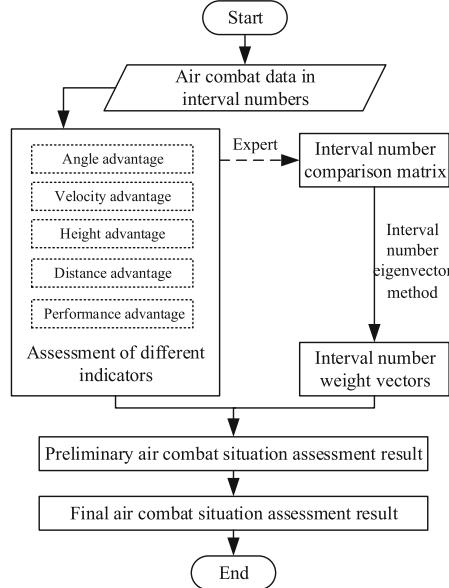
Then, a preliminary air combat situation assessment result can be got:

$$\hat{p}_{ij} = \tilde{\omega}_1 \cdot p_{aij} + \tilde{\omega}_2 \cdot p_{vij} + \tilde{\omega}_3 \cdot p_{hij} + \tilde{\omega}_4 \cdot p_{dij} + \tilde{\omega}_5 \cdot p_{cij}. \quad (11)$$

However,  $\hat{p}_{ij}$  is an interval number contained in  $[0, 1]$ , and we give the final situation assessment result  $p_{ij}$  by the following equation:

$$p_{ij} = \frac{1}{m(\hat{p}_{ij}, [1, 1])}. \quad (12)$$

The air combat situation assessment flowchart based on interval data is given in Fig. 2.



**Fig. 2.** Air combat situation assessment flowchart based on interval data

#### 4 Numerical Example

Supposing there are 2 UCAVs in our UCAV group:  $R_1, R_2$ , and there are 3 UCAVs in the enemy UCAV group:  $B_1, B_2, B_3$ . Their inherent parameters are given in Table 1. We will give situation assessment of our side to illustrate the interval number situation assessment method.

The relative parameters of our side to the enemy side are given by the following matrices, including the relative distance matrix  $d$  (km), the relative height matrix  $h$  (km), the azimuth matrix of our side  $\varphi$  (degree), and the entry angle matrix of the enemy side  $\rho$  (degree).

$$d = \frac{R_1}{R_2} \begin{pmatrix} [32, 33] & [34, 37] & [33, 37] \\ [54, 57] & [22, 26] & [32, 34] \end{pmatrix}, \quad h = \frac{R_1}{R_2} \begin{pmatrix} [8, 9] & [6, 7] & [2, 3] \\ [7, 9] & [5, 6] & [4, 5] \end{pmatrix},$$

**Table 1.** Inherent parameters of both sides.

Symbol	Description	$R_1$	$R_2$	$B_1$	$B_2$	$B_3$	Unit
$rm$	Missile attack distance	[43,55]	[23,24]	[33,35]	[23,25]	[31,33]	km
$rr$	Radar detection distance	[89,95]	[56,64]	[68,71]	[56,67]	[57,66]	km
$v$	Velocity	[453,466]	[512,523]	[465,471]	[356,375]	[312,334]	km/h
$lm$	Number of missiles	[3,4]	[2,2]	[3,3]	[2,3]	[1,2]	N/A

$$\varphi = \frac{R_1}{R_2} \begin{pmatrix} B_1 & B_2 & B_3 \\ [12, 15] & [23, 26] & [41, 43] \\ [34, 37] & [27, 29] & [35, 37] \end{pmatrix}, \quad \rho = \frac{R_1}{R_2} \begin{pmatrix} B_1 & B_2 & B_3 \\ [43, 49] & [36, 38] & [22, 23] \\ [71, 79] & [45, 46] & [47, 51] \end{pmatrix}.$$

By (1)–(6), we can calculate our angle advantage matrix  $p_a$ , our velocity advantage matrix  $p_v$ , our height advantage matrix  $p_h$ , our distance advantage matrix  $p_d$ , and our performance advantage matrix  $p_c$ .

Experts made a pairwise comparison of the weights of different air combat indicators, and concluded that the interval number weight matrix is:

$$\tilde{\Omega} = \begin{array}{c|ccccc} & \text{Angle} & \text{Velocity} & \text{Height} & \text{Distance} & \text{Performance} \\ \hline \text{Angle} & [1, 1] & [1, 2] & [1/3, 1/2] & [2, 3] & [3, 4] \\ \text{Velocity} & [1/2, 1] & [1, 1] & [1, 2] & [1/4, 1/3] & [1, 2] \\ \text{Height} & [2, 3] & [1/2, 1] & [1, 1] & [2, 3] & [1/3, 1/2] \\ \text{Distance} & [1/3, 1/2] & [3, 4] & [1/3, 1/2] & [1, 1] & [1/2, 1] \\ \text{Performance} & [1/4, 1/3] & [1/2, 1] & [2, 3] & [1, 2] & [1, 1] \end{array}.$$

Then we obtain that the maximum eigenvalue of  $\Omega^L$  is 5.0178, and the corresponding normalized eigenvector is  $x^L = (0.2759, 0.1221, 0.2478, 0.1589, 0.1953)$ . Similarly, we obtain that the maximum eigenvalue of  $\Omega^U$  is 7.7089, and the corresponding normalized eigenvector is  $x^U = (0.2528, 0.1691, 0.2258, 0.1645, 0.1878)$ . By Eq. (8), we can calculate that  $\alpha = 0.8063$ ,  $\beta = 1.0179$ . Then the weight vectors for different indicators are obtained:

$$\tilde{\omega} = (\tilde{\omega}_1, \tilde{\omega}_2, \tilde{\omega}_3, \tilde{\omega}_4, \tilde{\omega}_5) = ([0.2224, 0.2573], [0.0985, 0.1722], [0.1998, 0.2298], [0.1281, 0.1675], [0.1575, 0.1911]).$$

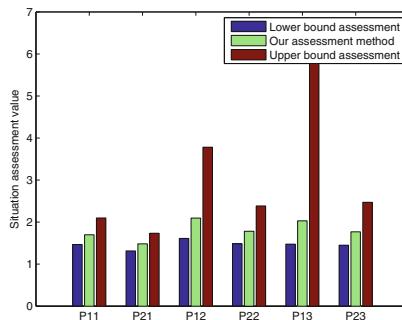
By Eq. (11) and (12), the preliminary air combat situation assessment matrix  $\hat{P}$  are obtained:

$$\hat{P} = \frac{R_1}{R_2} \begin{pmatrix} B_1 & B_2 & B_3 \\ [0.5177, 0.6628] & [0.5613, 0.8130] & [0.5208, 0.8856] \\ [0.4619, 0.5924] & [0.5235, 0.7035] & [0.5123, 0.7139] \end{pmatrix}.$$

By Eq. (12), the final air combat situation assessment matrix  $P$  are got:

$$P = \begin{pmatrix} 1.6993 & 2.0970 & 2.0298 \\ 1.4813 & 1.7818 & 1.7687 \end{pmatrix}$$

By using the assessment methods in [15, 16] with the upper and lower bounds of interval numbers to evaluate the air combat situation, the comparison with our results is presented in Fig. 3. Our results is within the range of upper and lower bound evaluation, which shows that our method is scientific and reasonable, and the uncertainty of the data is considered in our result, which is more valuable.



**Fig. 3.** Comparison of our method with existing methods.

## 5 Conclusion

Air combat situation assessment is the basis of air combat decision-making and weapon resource allocation, and plays a significant role in air combat. Traditional air combat situation assessment methods often use accurate air combat data and accurate weight coefficients to give air combat situation assessment values, ignoring the uncertainties of air combat data and expert experience. Thus, it is difficult to objectively describe the air combat situation. To solve this problem, this paper has proposed an air combat situation assessment method based on interval number theory, which uses interval numbers to describe air combat data and weight coefficients. The method is simple and practical, and the situation assessment results obtained are scientific and reasonable.

**Acknowledgements.** This work was supported by Major Projects for Science and Technology Innovation 2030 (Grant No. 2018AA0100800), Equipment Pre-research Foundation of Laboratory (Grant No. 61425040104) and Joint fund of China Electronics Technology for Equipment Pre-research (Grant No. 6141B0823110a).

## References

1. Ding, H., Chen, J., Song, J.: Fuzzy synthetic evaluation of air defense weapon synergistic efficacy for ship formation. *Fire Control Command Control* **35**, 95–98 (2010). <https://doi.org/10.3969/j.issn.1002-0640.2010.02.026>
2. Li F., Feihu C., Jin X., et al: Research of air combat situation assessment method. In: 2012 Third International Conference on Digital Manufacturing and Automation, GuiLin (2012). <https://doi.org/10.1109/ICDMA.2012.155>
3. Page, S.F., Oldfield, J.P., Thomas, P.: Towards integrated threat assessment and sensor management: Bayesian multi-target search. In: 2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Baden-Baden. IEEE Press (2016). <https://doi.org/10.1109/MFI.2016.7849465>
4. Geng, T., Zhang, A., Hao, X.: Multi-target threat assessment in air combat based on combination determining weights TOPSIS. *Fire Control Command Control* **36**, 16–19 (2011). <https://doi.org/10.3969/j.issn.1002-0640.2011.03.004>

5. Meng, G., Ma, X., Liu, X., Xu, Y.: Situation assessment for unmanned aerial vehicles air combat based on hybrid Bayesion network. *Command Control Simul.* **39**, 1–6+39 (2017). <https://doi.org/10.3969/j.issn.1673-3819.2017.04.001>
6. Wang, L., Kou, Y.: Application of Dempster-Shafer evidence theory in air combat situation assessment. *Electron. Opt. Control* **14**, 155–157 (2007). <https://doi.org/10.3969/j.issn.1671-637X.2007.06.039>
7. Zhao, W., Zhou, D.: Application of expert system in sequencing of air combat multi-target attacking. *Electron. Opt. Control* **15**, 23–26 (2008). <https://doi.org/10.3969/j.issn.1671-637X.2008.02.007>
8. Sun, H., Yao, W.: Comments on method for ranking interval numbers. *J. Syst. Eng.* **25**, 304–312 (2010). CNKI:SUN:XTGC.0.2010-03-005
9. Guo, H., Ren, B., Lv, Y., Luo, Y., Cui, L.: Target threat assessment for air combat based on intervals theory. *Fire Control Command Control* **38**, 31–34 (2013). CNKI:SUN:HLYZ.0.2013-06-011
10. Ou, A., Zhu, Z.: A method of threat assessment based on MADM and results of situation assessment in air to air combat. *Fire Control Radar Technol.* **35**, 64–67+89 (2006). <https://doi.org/10.19472/j.cnki.1008-8652.2006.02.015>
11. Xiao, L., Huang, J., Xu, Z.: Modeling air combat situation assessment based on combat area division. *J. Beijing Univ. Aeronaut. Astronaut.* **39**, 1309–1313 (2013). <https://doi.org/10.13700/j.bh.1001-5965.2013.10.006>
12. Buckley, B., Jowers, L.: Monte Carlo methods in fuzzy optimization. Springer, Heidelberg (2008). <https://doi.org/10.1007/978-3-540-76290-4>
13. Tran, L., Duckstein, L.: Comparison of fuzzy numbers using a fuzzy distance measure. *Fuzzy Sets Syst.* **130**, 331–341 (2002). [https://doi.org/10.1016/S0165-0114\(01\)00195-6](https://doi.org/10.1016/S0165-0114(01)00195-6)
14. Guo, H., Xu, H., Liu, L.: Threat assessment for air combat target based on interval TOPSIS. *Syst. Eng. Electron.* **31**, 2914–2917 (2009). CNKI:SUN:XTYD.0.2009-12-028
15. Dong, Y., Feng, J., Zhang, H.: Cooperative tactical decision methods for multi-aircraft air combat simulation. *J. Syst. Simul.* **14**, 723–725 (2002). <https://doi.org/10.3969/j.issn.1004-731X.2002.06.012>
16. Huo, X., Zhu, H., Shen, L.: Study on decision models of multi-target attack for group-craft cooperative air combat. *J. Syst. Simul.* **18**, 2573–2576 (2006). <https://doi.org/10.3969/j.issn.1004-731X.2006.09.049>



# Detection and Depth Estimation for Objects from Single Monocular Image

Ziwen Xu and Yingmin Jia<sup>(✉)</sup>

The Seventh Research Division and the Center for Information and Control,  
School of Automation Science and Electrical Engineering,  
Beihang University (BUAA), Beijing 100191, China  
[{xuziwen,ymjia}@buaa.edu.cn](mailto:{xuziwen,ymjia}@buaa.edu.cn)

**Abstract.** This paper addresses the problem of detecting and estimating depth for objects given a single monocular RGB image. We propose a integrative network to implement multiple tasks of detection and depth estimation at the same time and realize the rate of 6 fps. We use convolutional neural network to extract features and fully connection network to generate depth straightway and evaluate the performance of our model on KITTY. To adapt the model to multiple range scales of objects, we rectify the loss function and further improve the performance of our model.

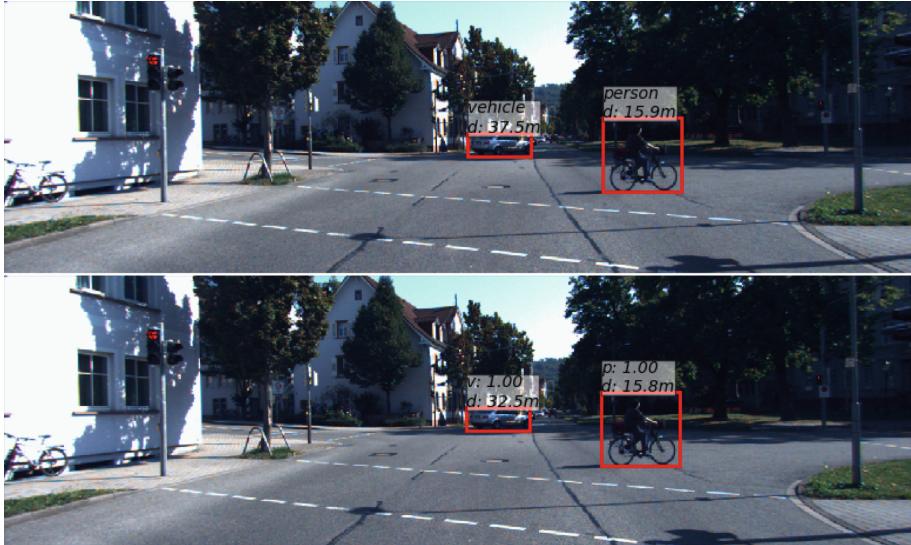
**Keywords:** Depth estimation · Object detection · Monocular vision

## 1 Introduction

Estimating depth and detecting objects from a single monocular image is a fundamental task in computer vision, which has widespread applications in scene understanding, 3D modelling, robotics, *etc.* Especially in the field of autonomous driving, detecting moving targets (vehicles, persons, *etc.*) and estimating distances from them to our vehicle camera is a crucial step which influencing on following path planning and behavior arbitration.

Object detection from 2D image has made significant progress in recent years, and its accuracy and efficiency are gradually improving by using deep convolutional neural network. But the task of depth estimation from single 2D image is usually seen as an ill-posed problem because of the lack of spatial information: Given a monocular image, there are a number of possible real world scenes which can produce it, but most of these are physically nonexistent based on heuristic knowledge. Thus we can still predict depth with considerable accuracy in ways of supervision learning (Fig. 1).

In this paper, we propose a simple and convenient solution to acquire information of location and depth for objects from single monocular RGB image. The task can be divided into two subtasks: object detection based on Faster R-CNN [19] and depth estimation based on our proposed method, which just extract corresponding features from bounding boxes generated by Faster R-CNN and

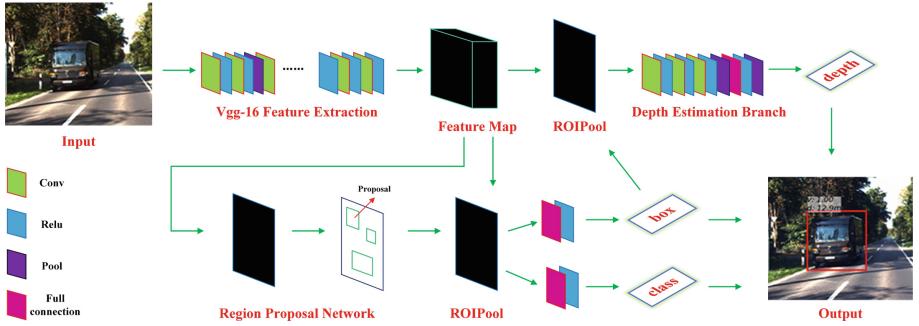


**Fig. 1.** Our experimental results about detection and depth estimation for objects from a single monocular RGB image with KITTY Dataset [13]. **Top:** Ground truth annotations about locations, labels and depth information of objects; **Bottom:** The predictive results where ‘**p**’ represents ‘**person**’, ‘**v**’ represents ‘**vehicle**’ and ‘**d**’ represents ‘**depth**’. The values following the above letters represent probability and length respectively. Only persons and vehicles are detected in this model.

then produce distance information by training a CNN network. To achieve the above objectives, a multi-task neural network is designed and then evaluated in KITTY [13].

## 2 Related Work

**Object Detection:** In deep learning area, object detection can be grouped in two branch: two-stage object detection and one-stage object detection, where the former methods usually generate proposed bounding box in the first stage and refine it in the second stage while the later methods just regress the bounding box in whole step. The typical two-stage frames are as follows: RCNN [5], Fast R-CNN [4], Faster R-CNN [19], FPN [10], etc. The two-stage object detector can achieve higher accuracy at the expense of speed. On the contrary, one-stage frames can proceed real-time detection but are unable to compare with the former with respect with accuracy. Representative one-stage frames can be listed as follows: YOLO and its v2, v3 editions [16–18], SSD [12], RetinaNet [11], etc. In this paper, in consideration of accurate feature map for objects is necessary to predict depth information, Faster R-CNN is applied as the object detector and extended with our proposed depth estimation branch. By using



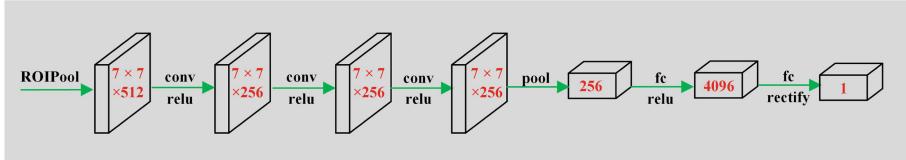
**Fig. 2.** The pipeline for detecting objects and estimating its depth. Given an input RGB image, we use Faster R-CNN to acquire feature map and generate bounding box, and then extract feature using ROI Pooling from each bounding box and perform estimating depth using our proposed module. The final output image combines information about box location, class and depth.

this strategy, the model realizes the synchronization of detecting and distance measurement.

**Monocular Depth Estimation:** One of the earliest work promoting research in monocular depth estimation using deep neural network is by Eigen et al. [2]. This paper introduces two-scale information and directly regresses over pixels for depth estimation in two-stage where the first stage produce a coarse output and refine it in the second stage. This two-scale design is then extended to three-scale for further refinement [1]. Laina et al. [8] replace fully-connected architecture in previous work with a fully convolutional model which can reduce the number of parameters for calculation overhead. Besides that, this work corporate Resnet [7] with a up-projections module to predict high resolution depth maps. Li et al. [9] propose a two-streamed neural network to estimate depth where one is used to predict depth and the other to predict gradient of depth and fuse them at the end. Fu et al. [3] propose a deep ordinal regression network which builds ordinal regression training loss [15] to relieve slow convergence that previous work more often suffer from. In our work, we just use fully convolutional network to extract depth features and then regress depth of objects directly by building logarithmic loss function.

### 3 Method

Our proposed network architecture is conceptually simple: We add a new depth estimation branch to original Faster R-CNN [19] to predict depth for each object directly. But to acquire more refined feature for more accurate prediction, we replace proposals which regress ultimate boxes location and decide what class it belongs to generated in the first stage with boxes generated in the second stage. We will discuss that in the following paragraph.



**Fig. 3. Depth Estimation Branch:** We extend Faster R-CNN with our proposed depth estimation branch. We start with ROI Pool for each bounding box generated by the second stage regression of Faster R-CNN, and adopt three convolutional layers to extract depth feature. Then we use global pooling and two fully connection layers to regress depth. Numbers denote spatial resolution and channels. All convs are  $3 \times 3$ , followed by corresponding ReLU [14] activation layer. The last layer is a special rectified layer which we discuss in Sect. 3.

**Faster R-CNN:** The network architecture begins with basic detector of Faster R-CNN. Firstly, Faster R-CNN adopts Vgg-16 [20] backbone to extract feature and generate feature map for each image (outlined in Fig. 2). Besides, two phases are divided into to regress bounding box: the first phase, called Region Proposal Network (RPN), produces proposal bounding boxes which are coarse locations for ground-truth objects; the second phase, extracts features using ROI Pool from corresponding proposal bounding boxes and performs refined regression and classification. This two-phase bounding box regression network architecture exports accurate locations of boxes at last.

**Depth Estimation Branch:** Some of other typical computer vision tasks, such as instance segmentation, which performs semantic segmentation on the basis of object detection, adopt multi-branch structure to handle corresponding tasks. Mask R-CNN [6] extends Faster R-CNN with a mask branch to output a binary segmentation mask for each object in the first stage (which is after ROI Pool). Our idea is similar with that: we just add a new depth estimation branch to Faster R-CNN to output depth information for each object. But the difference is that the depth estimation branch is extended in the second stage (which is after the second ROI Pool), considering that the feature extracted from the second stage is more accurate than the first.

Each bounding box is represented by feature matrix with same shape (channel, height and width are 512, 7 and 7, respectively) after ROI Pool processing. Then we predict its depth by our proposed depth estimation branch. Details are shown in Fig. 3. Specifically, unlike some other per-pixel depth estimation methods [3, 8, 9] by using fully convolutional neural network, we adopt fully connection layers in the last two layers to estimate depth for each object, which not only simplify the structure but also consider global information of the whole object.

At the end of the last structure of the branch, we adopt a special layer to rectify depth, which can be represented as follows:

$$d = \begin{cases} |x|, & |x| > 0.1, \\ 0.1, & |x| \leq 0.1 \end{cases} \quad (1)$$

where  $x$  is the scalar output of the last fully connection layer, and  $d$  is the final rectified output of depth.

We set a threshold of 0.1 for depth, in view of log-loss function we will take must assure that there is a Non-negative lower bound. Besides that, we adopt absolute value to rectify depth in consideration of the output of depth value may be negative under the condition of parameters are randomly initialized. if we only set threshold but not use absolute value to rectify when output of value is negative, the model may suffer gradient vanishing which will make the model no longer converge.

**Loss Function:** We adopt scale-invariant mean squared error instead of universal mean squared error. The scale-invariant mean squared error can be represented as follows:

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^N (\log d - \log \bar{d})^2 \quad (2)$$

where  $d$  is the output of depth estimation branch,  $N$  is the number of all objects and  $\bar{d}$  is the ground-truth depth value of each object. The reason why we use scale-invariant mean squared error is that we consider the estimation of depth should adapt to various scales. For example, suppose that there are two objects whose true depth value are 20 m and 60 m, respectively, and the prediction of our depth estimation branch are all 40 m. It is obvious that the predictive accuracy of the latter is higher than the former in terms of the ratio between prediction and truth. Beyond that, log-loss function can also accelerate convergence of our model (convergence curve is shown in Fig. 4).

But we found that the model tends to generate smaller value and create fairly big error when estimating distant objects if we just adopt the above-mentioned loss function (as in Eq. 2) to train the model. To alleviate this problem, we add a penalty factor to objects with closer range. The improved loss function can be represented as follows:

$$\text{Loss} = \frac{1}{N} \left( \sum_{d>\bar{d}} (\log d - \log \bar{d})^2 + \lambda \sum_{d\leq\bar{d}} (\log d - \log \bar{d})^2 \right) \quad (3)$$

where  $\lambda$  is the penalty factor and we set that  $\lambda \geq 1$ . If we set that  $\lambda = 1$ , Eq. 3 is in reality same with Eq. 2. We will discuss that how the hyper-parameter  $\lambda$  has influence on our experimental result (see Sect. 4).

## 4 Experiments

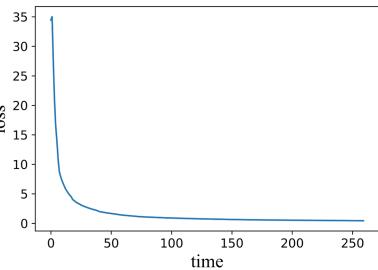
We evaluate our method on KITTY [13] and implement our network on the public deep learning platform Pytorch.

**KITTY:** The KITTY [13] contains over 93 thousand depth maps with corresponding raw Lidar scans and RGB images of resolution about  $375 \times 1375$  captured by cameras and depth sensors in a driving car. Furthermore, it also provide 7481 training images and 7518 test images, comprising a total of 80256 labeled objects for object detection. Since we are short of unified annotations about object detection and depth estimation, we can only pre-train Faster R-CNN [19] network in object detection dataset of KITTY, and then train our depth estimation branch in depth dataset of KITTY.

**Implementation Details:** We firstly pre-train our model of object detection based on Faster R-CNN [19] with 7481 training images provided by KITTY [13]:

**Table 1.** Comparison of depth estimation errors on the KITTY with different penalty factor  $\lambda$  by using our depth estimation branch (**rel** is relative error,  $\log_{10}$  is mean  $\log_{10}$  error and **rms** is root mean squared error; details of evaluation criteria can refer to Eigen et al. [2]).

Penalty factor	Lower is better		
	rel	$\log_{10}$	rms
1.0	0.265	0.158	11.16
2.0	<b>0.238</b>	<b>0.118</b>	<b>8.15</b>
3.0	0.240	0.135	8.49
3.5	0.280	0.146	9.75

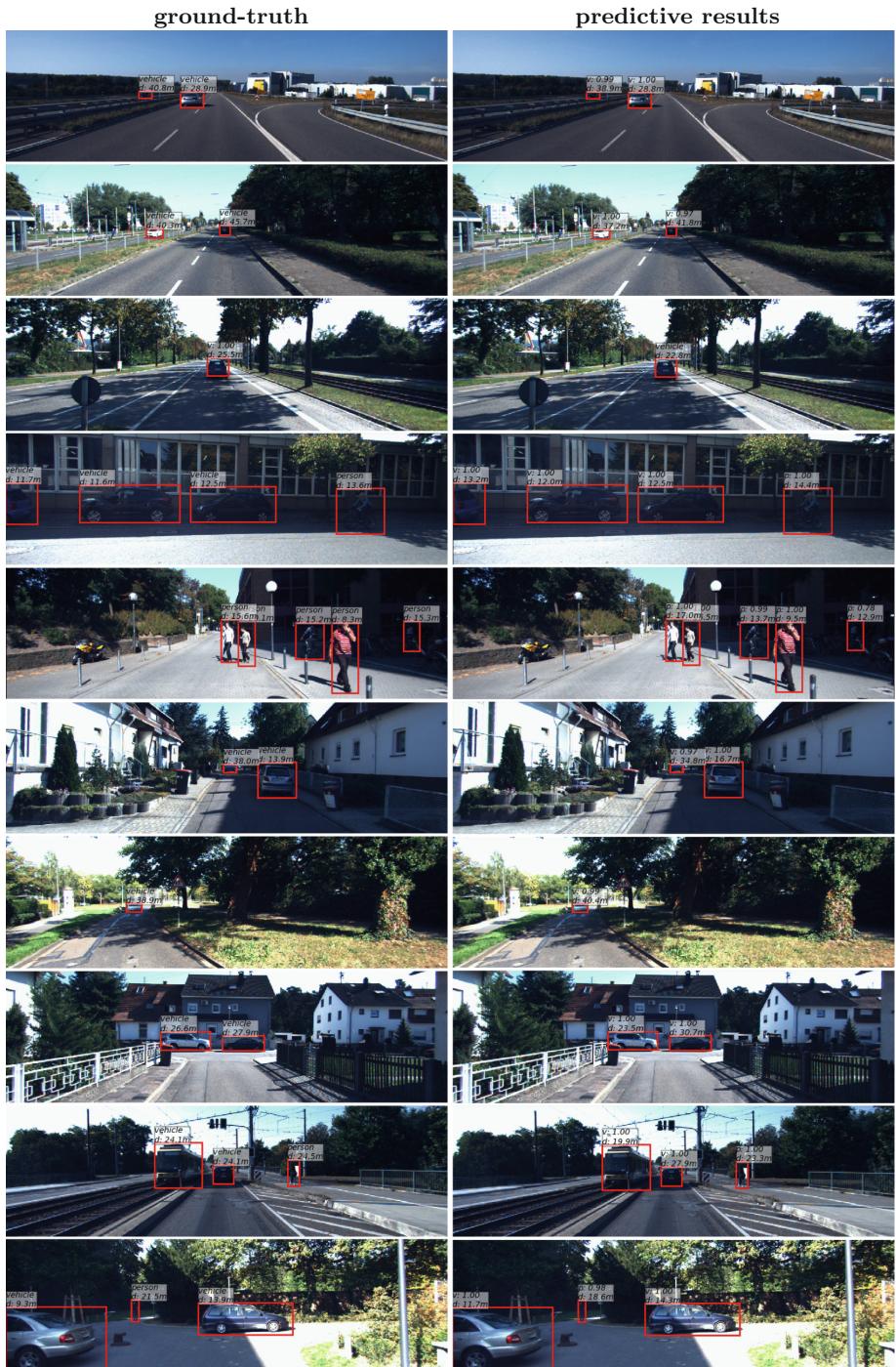


**Fig. 4.** Training loss (**rel**) of our experimental result of depth estimation branch ( $\lambda = 2.0$  in this picture). On account of using of log-loss function, the model converges very quickly in the first few iterations.

We replace all categories of vehicle (car, truck, tram, *etc.*) with label of vehicle and replace bicycle and pedestrian with label of person for the sake of simplicity. After preprocessing training dataset with the mentioned method, we set the following parameters to train the network: learning rate of 0.001 (in the last two epochs we use learning rate decay of 0.1), batch size of 1, weight decay of 0.0005 (regularization coefficient of  $L2$ ) and epoch of 4 (30k iterations approximately).

After obtaining the relatively accurate model of object detecting, we start focusing on the training of the depth estimation branch. We sample 4110 images comprising raw RGB images and corresponding depth maps from KITTY and divide those into 7 to 3 (2877 images for training and 1233 images for test respectively). In view of the information about depth in depth maps is per-pixel and that means we can't obtain the depth of entire object directly, we take the average of depth of all valid pixels that exist depth in the region of one object and regard it as the depth of the object for simplicity.

We frozen all parameters in our pre-trained Faster R-CNN model and initialize parameters in depth estimation branch randomly. What other hyper-parameters we set for training are as follows: learning rate of 0.0025 (in the



**Fig. 5.** More results of the network ( $\lambda = 2$  is set in this picture) on test dataset of KITTY [13].

last epoch we use learning rate decay of 0.1), batch size of 1, weight decay of 0.0005 and epoch of 2 (6k iterations approximately) by using a single device of GTX1060. We found that the training loss will not nearly descend in the second epoch and this is why we let the network train only two rounds. The average training speed of our model is 6 fps (frames per second). The visualization of our training process is shown in Fig. 4.

**Evaluation:** We consider the influence of several values of hyper-parameter  $\lambda$  on the experimental results in the test dataset separated from KITTY (shown in Table 1). The evaluation criteria of **rel** and **log<sub>10</sub>** reflect relative error in test dataset while **rms** reflect absolute error by contrast. Note that 2.0 is not necessarily the most appropriate value of  $\lambda$  and 3.0 or 3.5 could produce better result under some circumstances. But anyway,  $\lambda > 1.0$  can improve the effect of depth estimation than  $\lambda = 1.0$  because of the better adaptation for distant object. We also attempt to continue to increase the value of parameter  $\lambda$  (e.g..  $\lambda = 4.0$ ), but this does not only improve the result but also may cause the model no longer converge. It indicates that the penalty factor  $\lambda$  should be controlled within a reasonable range. More visualization results are shown in Fig. 5.

## 5 Conclusion

In this paper we present a simple but effective approach to the detection and depth estimation for objects from single monocular image. Unlike typical approaches to estimate depth are based on generating depth maps for one total image, our method can predict depth for all objects which are also detected accurately in our network. We extend the approach of object detection based on Faster R-CNN with our proposed generative model of depth estimation branch to realize the learning of all tasks above. The model has a fast rate of training and convergence and achieve a relatively reliable accuracy.

However, the shortage of our method is that the model is not end-to-end (pre-trained in Faster R-CNN) and that is mainly because we lack of datasets with sufficient annotations which should comprise effective information of depth for objects. Besides that, we believe that speculating depth in monocular images should consider more factors of integrity and environment which are also missing from current approaches. We will make improvements on above-mentioned shortcomings in future works.

**Acknowledgement.** This work was supported by the National Basic Research Program of China (973 Program: 2012CB821200, 2012CB821201) and the NSFC (61134005, 61327807, 61521091, 61520106010).

## References

1. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: The IEEE International Conference on Computer Vision (ICCV), December 2015

2. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems 27*, pp. 2366–2374. Curran Associates, Inc. (2014)
3. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
4. Girshick, R.: Fast R-CNN. In: The IEEE International Conference on Computer Vision (ICCV), December 2015
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014
6. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: The IEEE International Conference on Computer Vision (ICCV), October 2017
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
8. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks (2016)
9. Li, J., Klein, R., Yao, A.: A two-streamed network for estimating fine-scaled depth maps from single RGB images. In: The IEEE International Conference on Computer Vision (ICCV), October 2017
10. Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
11. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: The IEEE International Conference on Computer Vision (ICCV), October 2017
12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *Computer Vision – ECCV 2016*, pp. 21–37. Springer, Cham (2016)
13. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
14. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 807–814 (2010)
15. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output CNN for age estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
16. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
17. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
18. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement (2018)
19. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 28*, pp. 91–99. Curran Associates, Inc. (2015)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014)



# Leader-Following Consensus of Multi-agent Systems: New Results Based on a Linear Transformation Approach

Jingyuan Zhan and Yangzhou Chen<sup>(✉)</sup>

College of Artificial Intelligence and Automation,  
Beijing University of Technology, Beijing 100124, China  
[{jyzhan,yzchen}@bjut.edu.cn](mailto:{jyzhan,yzchen}@bjut.edu.cn)

**Abstract.** This paper investigates the leader-following consensus problem of multiple single-integrators by using a novel linear transformation method together with the input-to-state stability property. The multi-agent system is assumed to contain a single leader, and we consider the following three cases. 1) The leader's input is pre-given and known by all following agents. 2) The leader's input is unknown. 3) The leader's input is measurable online and transmitted to some of follower agents. By constructing a transformation matrix based on incidence matrix of a virtual leader-rooted spanning tree, we make an equivalent transformation from leader-following consensus problem to an input-to-state stability problem. Then we give a necessary and sufficient condition, which is the Hurwitz stability of a matrix associated with the communication topology, for ensuring the leader-following consensus. In order to efficiently check whether the matrix is Hurwitz stable, especially for large-scale multi-agent systems, we further employ the Hurwitz stability criteria of the matrix based on Metzler matrix theory. Finally, we give numerical examples to validate the theoretical results.

**Keywords:** Multi-agent system · Leader-following consensus · Linear transformation · Input-to-state stability (ISS) · Hurwitz stable · Metzler matrix

## 1 Introduction

In recent decades, cooperative control of networked multi-agent systems (MASs) has attracted tremendous attentions due to its advantages, such as scalability, robustness and flexibility. As a typical and fundamental problem of cooperative control, consensus of MASs has becoming a hot research topic, and it also finds broad industrial applications in intelligent transportation systems [1], wireless sensor networks [2], smart grids [3], and etc.

Early in 1980s, a distributed consensus algorithm for an MAS was proposed by Tsitsiklis and Athans in [4], where conditions for achieving the asymptotic consensus were also derived. The consensus of agents' headings in discrete-time first-order MASs was studied in [5], and the model is the famous Vicsek model. Later, it

was proved in [6] that if the communication topology in the Vicsek model is jointly connected, consensus is guaranteed. The theoretical framework of studying consensus of MASs with time-delays and switching topology was introduced in [7]. More recently, abundant results were obtained in investigating consensus problems with various complex settings involving time delays [8], intermittent communication [9], packet loss [10], high-order dynamics [11], and so on.

Besides, there has been a great amount of results concerning consensus problems of MASs with a leader, which is the so-called leader-following consensus. Jadbabaie et al. [6] considered the first-order MAS with a fixed leader, firstly pointing out all follower agents' states converge to the leader's asymptotically if the communication topology is jointly connected. By employing the extended LaSalle's invariance principle, a theoretical proof for the leader-following consensus was given in [12]. The controllability of a leader-following MAS was considered from perspective of graph theory in [13]. Furthermore, a velocity-varying leader was taken into account in addressing the leader-following consensus problem of delayed MASs [14]. The leader-following consensus problem of multiple agents with  $n$ -th order linear dynamics was studied in [15], by introducing an approach based on a Riccati-inequality combined with graph theory.

Most of the existing work aforementioned employed tools from Lyapunov functions, graph theory, contraction theory, stochastic matrix analysis, and etc. Distinguishingly, a linear transformation method was developed in [16] to make an equivalent transformation from the consensus problem to a partial stability problem, such that the existing theoretical results of partial stability can be applied directly. Motivated by the effectiveness of the linear transformation approach, we study the leader-following consensus problem by introducing a novel linear transformation approach together with input-to-state stability (ISS) property. Explicitly, we consider a leader-following MAS with single-integrator dynamics in this paper with three cases taken into account. 1) The leader's input is pre-given and known by all following agents. 2) The leader's input is unknown. 3) The leader's input is measurable online and transmitted to some of following agents. By regarding the leader as a root node, we can arbitrarily define a virtual directed spanning tree of the MAS, and we call it the virtual leader-rooted spanning tree. By constructing a transformation matrix based on incidence matrix of a virtual leader-rooted spanning tree, we make an equivalent transformation from leader-following consensus problem to an ISS problem. Then we give a necessary and sufficient condition, which is the Hurwitz stability of a matrix associated with the communication topology, for ensuring the leader-following consensus. In order to efficiently check whether the matrix is Hurwitz stable, especially for large-scale multi-agent systems, we further employ the Hurwitz stability criteria of the matrix based on Metzler matrix theory [17–19].

The main contributions of this study are two-fold. 1) We introduce a novel transformation matrix based on incidence matrix of a virtual leader-rooted spanning tree, and then make an equivalent transformation from leader-following consensus problem to an ISS problem so that the existing results related to ISS can be applied directly. This study partially extends the application of linear

transformation approach in [16] to leader-following consensus. The proposed linear transformation approach is also effective in analyzing the leader-following consensus problem of MASs with more general dynamics, though we merely consider the single-integrator dynamics in this paper. 2) Since the necessary and sufficient condition for ensuring the leader-following consensus is the Hurwitz stability of a matrix associated with the overall communication topology, we further employ the Hurwitz stability criteria of the matrix based on Metzler matrix theory, which are very simple to check especially for large-scale MASs.

The rest of this paper is organized as follows. Section 2 provides preliminaries related to the problem to be studied, and then describes the problem formulation explicitly. The leader-following consensus protocols and analysis under three cases are given in Sect. 3. Several numerical examples are provided in Sect. 4 to verify the correctness of the theoretical results. Conclusions of this paper are finally given in Sect. 5.

## 2 Preliminaries and Problem Formulation

We firstly introduce some mathematical notations. Let  $\mathbb{R}$  denote the set of real numbers,  $\mathbb{R}_*$  denote the set of nonnegative real numbers,  $\mathbb{R}^n$  denote the set of  $n \times 1$ -dimensional real vectors, and  $\mathbb{R}^{n \times m}$  denote the set of  $n \times m$ -dimensional real matrices. For a given matrix  $B$ ,  $B^T$  and  $B^{-1}$  denote the transpose and inverse of  $B$ , respectively.  $B > 0$  if it is positive definite,  $B < 0$  if it is negative definite,  $B \succ 0$  if all elements are positive, and  $B \prec 0$  if all elements are negative.  $\mathbf{1} = [1 \ 1 \ \dots \ 1]^T$ ,  $\mathbf{0} = [0 \ 0 \ \dots \ 0]^T$ , and  $I$  denotes the identity matrix, all of which are with appropriate dimensions if no confusions arise.

Recall that a function  $\gamma : \mathbb{R}_* \rightarrow \mathbb{R}_*$  is a  $\mathcal{K}$ -function if it is continuous and strictly increasing with  $\gamma(0) = 0$ . A function  $\beta : \mathbb{R}_* \times \mathbb{R}_* \rightarrow \mathbb{R}_*$  is a  $\mathcal{KL}$ -function if  $\beta(\cdot, t)$  is a  $\mathcal{K}$ -function for any fixed  $t \geq 0$ , and  $\beta(s, \cdot)$  is decreasing for any fixed  $s \geq 0$  with  $\beta(s, t) \rightarrow 0$  as  $t \rightarrow \infty$ . A matrix  $M \in \mathbb{R}^{n \times n}$  is a Metzler matrix if its off diagonal elements are nonnegative, and the following lemma summarizes the results from [17–19] on Metzler matrix stability.

**Lemma 1.** *For a Metzler matrix  $M \in \mathbb{R}^{n \times n}$ , the following 7 points of statements are equal:*

1.  *$M$  is Hurwitz stable, i.e., all eigenvalues of  $M$  are located in the open left half-plane;*
2. *[17] all sequential principal minors of matrix  $-M$  are positive;*
3. *[17] all coefficients of the characteristic polynomial  $\det(\lambda I - M)$  are positive;*
4. *[17]  $M$  is diagonally stable, i.e., there exists a diagonal matrix  $P > 0$  satisfying  $M^T P + P M < 0$ ;*
5. *[18] there exists a vector  $a \succ 0$  satisfying  $Ma \prec 0$ ;*
6. *[18] there exists a vector  $b \succ 0$  satisfying  $b^T M \prec 0$ ;*
7. *[19] all the diagonal elements of matrices*

$$M_n, M_{n-1}, \dots, M_1$$

are negative, where  $M_n = M$  and  $M_k, k = n-1, n-2, \dots, 1$  are calculated iteratively by  $M_k = \hat{M}_k - b_k c_k^T / d_k$  with

$$M_{k+1} = \begin{bmatrix} \hat{M}_k & b_k \\ c_k^T & d_k \end{bmatrix},$$

$$\hat{M}_k \in \mathbb{R}^{k \times k}, b_k, c_k \in \mathbb{R}^k, d_k \in \mathbb{R}.$$

Consider an MAS with  $N+1$  single-integrator agents labelled with 0 through  $N$ . Agent 0 is regarded as the leader, and the others are follower agents. The leader-following MAS is described by

$$\dot{x}_i(t) = u_i(t), \quad i = 0, 1, \dots, N \quad (1)$$

where  $x_i \in \mathbb{R}$  and  $u_i \in \mathbb{R}$  denote the state and input of agent  $i$ , respectively. Define the overall state and input of followers as  $x = [x_1 \ x_2 \ \dots \ x_N]^T$  and  $u = [u_1 \ u_2 \ \dots \ u_N]^T$ , respectively.

The communication topology among followers is described by a digraph  $\mathcal{G}_f = (\mathcal{V}_f, \mathcal{E}_f, A)$ , in which  $\mathcal{V}_f = \{1, \dots, N\}$  is the vertex set,  $\mathcal{E}_f \subseteq \{(i, j) : i, j \in \mathcal{V}_f, j \neq i\}$  is the edge set, and  $A = [a_{ij}] \in \mathbb{R}^{N \times N}$  is the adjacency matrix with elements denoted by  $a_{ij}$ . An edge  $(j, i) \in \mathcal{E}_f$  means that agent  $i$  can receive the information transmitted from agent  $j$ . Iff  $(i, j) \in \mathcal{E}_f$ ,  $a_{ji} > 0$ ; otherwise,  $a_{ji} = 0$ . The neighbors of agent  $i$  are denoted by  $N_i = \{j \in \mathcal{V}_f : a_{ij} > 0\}$ . Then Laplacian matrix  $L = [l_{ij}]$  of  $\mathcal{G}_f$  is defined as:

$$l_{ij} = \begin{cases} -a_{ij}, & \text{if } i \neq j \\ \sum_{k \in \mathcal{V}} a_{ik}, & \text{if } i = j \end{cases}.$$

Besides, we define  $d_i, i \in \mathcal{V}_f$  to represent the interconnection between agent  $i$  and agent 0, which is unidirectional from agent 0 to agent  $i$  for all  $i \in \mathcal{V}_f$ . If agent  $i$  can receive information transmitted from agent 0, we set  $d_i > 0$ ; otherwise, set  $d_i = 0$ . We also define  $D = \text{diag}\{d_i\}$  and  $d = [d_1, d_2, \dots, d_N]^T$ . Then we define the overall communication topology  $\mathcal{G}$  consisting of  $\mathcal{G}_f$ , the leader and the edges between the leader and other agents.

**Definition 1.** MAS (1) achieves leader-following consensus if there exists a  $\mathcal{KL}$ -function  $\beta$  satisfying

$$\|x(t) - \mathbf{1}x_0(t)\| \leq \beta(\|x(0) - \mathbf{1}x_0(0)\|, t). \quad (2)$$

**Definition 2.** MAS (1) achieves input-to-state leader-following consensus if there exist a  $\mathcal{K}$ -function  $\gamma$  and a  $\mathcal{KL}$ -function  $\beta$  satisfying

$$\|x(t) - \mathbf{1}x_0(t)\| \leq \beta(\|x(0) - \mathbf{1}x_0(0)\|, t) + \gamma(\|u_0\|_\infty), \quad (3)$$

where  $\|u_0\|_\infty = \sup_{t \geq 0} \|u_0(t)\|$ .

We are going to study the leader-following consensus problem of MAS (1) with the following three cases. 1) The leader's input is pre-given and known by all following agents. 2) The leader's input is unknown. 3) The leader's input is measurable online and transmitted to some of following agents.

### 3 Leader-Following Consensus

In this section, we mainly consider the leader-following consensus problem of MAS (1) in the three cases as mentioned before. By constructing a transformation matrix based on incidence matrix of a virtual leader-rooted spanning tree, we make an equivalent transformation from leader-following consensus problem to an ISS problem. Then we derive a necessary and sufficient condition, which is the Hurwitz stability of  $-(L + D)$ , for ensuring the leader-following consensus. In order to efficiently check whether  $-(L + D)$  is Hurwitz stable, especially for large-scale multi-agent systems, we further employ the Hurwitz stability criteria of  $-(L + D)$  based on Metzler matrix theory.

#### 3.1 Leader-Following Consensus with Leader's Input Given

When the input of the leader is pre-given and known by all the followers, the consensus protocol for follower agents is

$$u_i = u_0 + k \sum_{j \in N_i} a_{ij}(x_j - x_i) + kd_i(x_0 - x_i), \quad (4)$$

where  $k$  is the gain to be designed. MAS (1) under protocol (4) can be organized into the compact form:

$$\begin{bmatrix} \dot{x}_0 \\ \dot{x} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{0}^T \\ kd & -k(L + D) \end{bmatrix} \begin{bmatrix} x_0 \\ x \end{bmatrix} + \mathbf{1} u_0. \quad (5)$$

We introduce a transformation matrix denoted by  $U$ , which is constructed as follows:

$$U = [E_1 \ E_2 \ \cdots \ E_N \ e_1]^T = \begin{bmatrix} P^T \\ e_1^T \end{bmatrix} = \begin{bmatrix} p & \tilde{P}^T \\ 1 & \mathbf{0}^T \end{bmatrix},$$

where  $e_1 = [1 \ 0 \ 0 \ \dots \ 0]^T \in \mathbb{R}^{N+1}$  and

$$P = [E_1 \ E_2 \ \cdots \ E_N] = \begin{bmatrix} p^T \\ \tilde{P} \end{bmatrix}$$

is the incidence matrix of a virtual leader-rooted spanning tree of  $\mathcal{G}$ . For example, one can construct a chain rooted from the leader as the simplest virtual directed spanning tree.  $E_i = [E_{1,i} \ E_{2,i} \ \dots \ E_{N+1,i}]^T \in \mathbb{R}^{N+1}$  corresponds to an edge  $(n_i, i)$  in the virtual spanning tree with  $E_{n_i+1,i} = 1$ ,  $E_{i+1,i} = -1$  and the other elements equal to 0. It is straightforward that  $P^T \mathbf{1} = \mathbf{0}$ .  $p \in \mathbb{R}^N$  is with all entries equal to 0 except only one entry 1 in the  $i$ -th row iff the leader is connected to agent  $i$  in the virtual spanning tree.

The transformation matrix  $U$  is an invertible matrix, and we have

$$U^{-1} = [(P^T)^+ \ \mathbf{1}] = \begin{bmatrix} \mathbf{0}^T & 1 \\ (\tilde{P}^T)^{-1} & \mathbf{1} \end{bmatrix}, \quad (6)$$

where  $(P^T)^+ = \begin{bmatrix} \mathbf{0}^T \\ (\tilde{P}^T)^{-1} \end{bmatrix}$  satisfying  $P^T(P^T)^+ = I$ .

Considering the linear transformation

$$z = \begin{bmatrix} \tilde{x} \\ x_0 \end{bmatrix} = U \begin{bmatrix} x_0 \\ x \end{bmatrix},$$

system (5) is transformed into

$$\dot{z} = \begin{bmatrix} p \tilde{P}^T \\ 1 \end{bmatrix} \begin{bmatrix} 0 & \mathbf{0}^T \\ kd - k(L + D) & \end{bmatrix} \begin{bmatrix} \mathbf{0}^T & 1 \\ (\tilde{P}^T)^{-1} & \mathbf{1} \end{bmatrix} z + \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} u_0. \quad (7)$$

Notice that  $kd - k(L + D)\mathbf{1} = 0$  and

$$\begin{bmatrix} p \tilde{P}^T \\ 1 \end{bmatrix} \begin{bmatrix} 0 & \mathbf{0}^T \\ kd - k(L + D) & \end{bmatrix} \begin{bmatrix} \mathbf{0}^T \\ (\tilde{P}^T)^{-1} \end{bmatrix} = -k\tilde{P}^T(L + D)(\tilde{P}^T)^{-1}.$$

Then we have

$$\dot{\tilde{x}} = -k\tilde{P}^T(L + D)(\tilde{P}^T)^{-1}\tilde{x} \quad (8)$$

**Theorem 1.** MAS (1) under protocol (4) achieves leader-following consensus iff Metzler matrix  $-(L + D)$  satisfies any one item of 1–7 in Lemma 1.

*Proof.* (Necessity.) Assume MAS (1) under protocol (4) achieves leader-following consensus. Then there exists a  $\mathcal{KL}$ -function  $\beta_1$  satisfying (2), according to Definition 1.

Define  $Q \triangleq P^T$ ,  $\bar{x} = [x_0 \ x^T]^T$ , and  $\bar{e} = \bar{x} - \mathbf{1}x_0$ . Then

$$\begin{aligned} \tilde{x} &= Q\bar{x} = Q(\bar{x} - \mathbf{1}x_0) = Q\bar{e}, \\ \bar{e} &= U^{-1}U\bar{e} = [(Q)^+Q + \mathbf{1}e_1^T]\bar{e} = (Q)^+Q\bar{e} = (Q)^+\tilde{x}. \end{aligned}$$

Therefore,

$$\begin{aligned} \|\tilde{x}(t)\| &\leq \|Q\| \cdot \|\bar{e}(t)\| \\ &= \|Q\| \cdot \|x(t) - \mathbf{1}x_0(t)\| \\ &\leq \|Q\| \cdot \beta_1(\|x(0) - \mathbf{1}x_0(0)\|, t) \\ &= \|Q\| \cdot \beta_1(\|\bar{e}(0)\|, t) \\ &= \|Q\| \cdot \beta_1((Q)^+\tilde{x}(0)\|, t) \\ &\leq \|Q\| \cdot \beta_1((Q)^+\|\tilde{x}(0)\|, t) \\ &= \beta_2(\|\tilde{x}(0)\|, t), \end{aligned}$$

where  $\beta_2(\|\tilde{x}(0)\|, t) = \|Q\| \cdot \beta_1((Q)^+\|\tilde{x}(0)\|, t)$  is also a  $\mathcal{KL}$ -function. It implies  $\tilde{x}(t) \rightarrow 0$  when  $t \rightarrow \infty$ , which means system (8) is stable. It's straightforward that system (8) is stable iff  $-(L + D)$  is Hurwitz stable. Obviously,  $-(L + D)$  is a Metzler matrix. From Lemma 1, we further obtain that system (1) under protocol (4) achieves leader-following consensus only if  $-(L + D)$  satisfies any one item of 1–7 in Lemma 1.

(Sufficiency.) Assume  $-(L + D)$  is Hurwitz stable, and we can easily obtain that (2) holds following the similar arguments in the proof for necessity such that MAS (1) under protocol (4) achieves leader-following consensus.

**Remark 1.** It is obvious that the Hurwitz stability of  $-(L + D)$  is equivalent to the existence of a leader-rooted spanning tree in  $\mathcal{G}$ . By employing the equivalent criteria w.r.t. a Metzler matrix as stated in Lemma 1, one can efficiently check whether  $-(L + D)$  is Hurwitz stable, especially for large-scale MASs.

### 3.2 Leader-Following Consensus with Leader's Input Unknown

When the leader's input is unknown, the consensus protocol for follower agents is

$$u_i = k \sum_{j \in N_i} a_{ij}(x_j - x_i) + kd_i(x_0 - x_i). \quad (9)$$

MAS (1) under protocol (9) can be organized into the compact form:

$$\begin{bmatrix} \dot{x}_0 \\ \dot{x} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{0}^T \\ kd - k(L + D) & \end{bmatrix} \begin{bmatrix} x_0 \\ x \end{bmatrix} + e_1 u_0. \quad (10)$$

Considering the linear transformation

$$z = \begin{bmatrix} \tilde{x} \\ x_0 \end{bmatrix} = U \begin{bmatrix} x_0 \\ x \end{bmatrix},$$

system (10) is transformed into

$$\dot{z} = \begin{bmatrix} p \tilde{P}^T \\ 1 \mathbf{0}^T \end{bmatrix} \begin{bmatrix} 0 & \mathbf{0}^T \\ kd - k(L + D) & \end{bmatrix} \begin{bmatrix} \mathbf{0}^T & 1 \\ (\tilde{P}^T)^{-1} & \mathbf{1} \end{bmatrix} z + \begin{bmatrix} p \\ 1 \end{bmatrix} u_0. \quad (11)$$

Then we have

$$\dot{\tilde{x}} = -k\tilde{P}^T(L + D)(\tilde{P}^T)^{-1}\tilde{x} + pu_0. \quad (12)$$

**Theorem 2.** System (1) under protocol (9) achieves input-to-state leader-following consensus if and only if  $-(L + D)$  satisfies any one item of 1–7 in Lemma 1.

*Proof.* (Necessity.) Assume system (1) under protocol (9) achieves input-to-state leader-following consensus. According to Definition 2, there exist a  $\mathcal{K}$ -function  $\gamma_1$  and a  $\mathcal{KL}$ -function  $\beta_1$  satisfying (3).

Define  $Q \triangleq P^T$ ,  $\bar{x} = [x_0 \ x^T]^T$ , and  $\bar{e} = \bar{x} - \mathbf{1}x_0$ . Then

$$\tilde{x} = Q\bar{x} = Q(\bar{x} - \mathbf{1}x_0) = Q\bar{e},$$

$$\bar{e} = U^{-1}U\bar{e} = (Q)^+Q\bar{e} = (Q)^+\tilde{x}.$$

Therefore,

$$\begin{aligned} \|\tilde{x}(t)\| &\leq \|Q\| \cdot \|\bar{e}(t)\| \\ &= \|Q\| \cdot \|x(t) - \mathbf{1}x_0(t)\| \\ &\leq \|Q\| \cdot (\beta_1(\|x(0) - \mathbf{1}x_0(0)\|, t) + \gamma_1(\|u_0\|)) \\ &= \|Q\| \cdot (\beta_1(\|\bar{e}(0)\|, t) + \gamma_1(\|u_0\|)) \\ &= \|Q\| \cdot (\beta_1(\|(Q)^+\tilde{x}(0)\|, t) + \gamma_1(\|u_0\|)) \\ &\leq \|Q\| \cdot \beta_1(\|(Q)^+\| \cdot \|\tilde{x}(0)\|, t) + \|Q\| \cdot \gamma_1(\|u_0\|) \\ &= \beta_2(\|\tilde{x}(0)\|, t) + \gamma_2(\|u_0\|), \end{aligned}$$

where  $\beta_2(\|\tilde{x}(0)\|, t) = \|Q\| \cdot \beta_1(\|(Q)^+\| \cdot \|\tilde{x}(0)\|, t)$  is also a  $\mathcal{KL}$ -function, and  $\gamma_2(\|u_0\|) = \|Q\| \gamma_1(\|u_0\|)$  is also a  $\mathcal{K}$ -function. It implies that system (12) is ISS. It's straightforward that system (12) is ISS iff  $-(L + D)$  is Hurwitz stable. Obviously,  $-(L + D)$  is a Metzler matrix. From Lemma 1, we further obtain that system (1) under protocol (9) achieves input-to-state leader-following consensus only if  $-(L + D)$  satisfies any one item of 1–7 in Lemma 1.

(Sufficiency.) Assume  $-(L + D)$  is Hurwitz stable, and we can easily obtain that (3) holds following the similar arguments in the proof for necessity such that system (1) under protocol (9) achieves input-to-state leader-following consensus.

### 3.3 Leader-Following Consensus with Leader's Input Measurable Online

When the input of the leader is measurable online, the consensus protocol for follower agents is

$$u_i = k \sum_{j \in N_i} a_{ij}(x_j - x_i) + kd_i(x_0 - x_i) + g_i u_0, \quad (13)$$

where  $g_i > 0$  if agent  $i$  can receive the measured input of the leader, and  $g_i = 0$  otherwise. MAS (1) under protocol (13) can be organized into the compact form:

$$\begin{bmatrix} \dot{x}_0 \\ \dot{x} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{0}^T \\ kd - k(L + D) & \end{bmatrix} \begin{bmatrix} x_0 \\ x \end{bmatrix} + \begin{bmatrix} 1 \\ g \end{bmatrix} u_0, \quad (14)$$

where  $g = [g_1, g_2, \dots, g_N]^T$ .

Considering the linear transformation

$$z = \begin{bmatrix} \tilde{x} \\ x_0 \end{bmatrix} = U \begin{bmatrix} x_0 \\ x \end{bmatrix},$$

system (14) is transformed into

$$\dot{z} = \begin{bmatrix} p \tilde{P}^T \\ 1 \mathbf{0}^T \end{bmatrix} \begin{bmatrix} 0 & \mathbf{0}^T \\ kd - k(L + D) & \end{bmatrix} \begin{bmatrix} \mathbf{0}^T & 1 \\ (\tilde{P}^T)^{-1} \mathbf{1} & \end{bmatrix} z + \begin{bmatrix} p + \tilde{P}^T g \\ 1 \end{bmatrix} u_0. \quad (15)$$

Then we have

$$\dot{\tilde{x}} = -k\tilde{P}^T(L + D)(\tilde{P}^T)^{-1}\tilde{x} + (p + \tilde{P}^T g)u_0. \quad (16)$$

Similar as Theorem 2, it's straightforward for us to present the following theorem, whose proof is not included in this paper.

**Theorem 3.** *MAS (1) under protocol (13) achieves leader-following consensus iff  $-(L + D)$  satisfies any one item of 1–7 in Lemma 1 and  $p + \tilde{P}^T g = \mathbf{0}$ .*

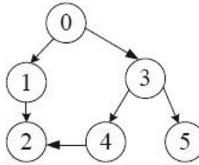
**Remark 2.** *According to the equality  $p + \tilde{P}^T g = \mathbf{0}$  in Theorem 3,  $g = -(\tilde{P}^T)^{-1}p$  is uniquely determined, and the property that  $P^T \mathbf{1} = \mathbf{0}$  further implies  $g = \mathbf{1}$ , which means that the input of the leader needs to be transmitted to all following agents.*

## 4 Simulation Results

We will present several numerical examples to demonstrate the theoretical results in this section. Consider a leader-following MAS (1) with  $N = 5$ , whose communication topology is as shown in Fig. 1. Assume the weights of all edges are 1, and then  $-(L + D)$  for the graph in Fig. 1 is

$$\begin{bmatrix} -1 & 0 & 0 & 0 & 0 \\ 1 & -2 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{bmatrix}.$$

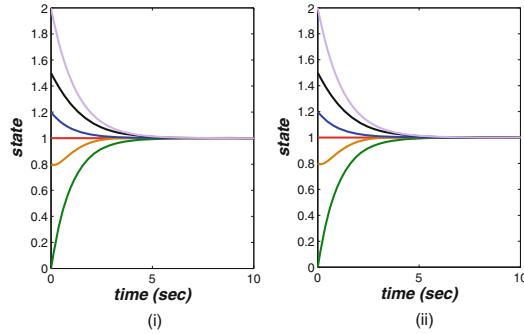
By checking item 7 in Lemma 1, we can easily know that  $-(L + D)$  is Hurwitz stable.



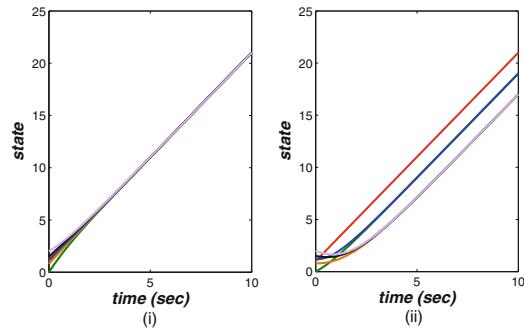
**Fig. 1.** Communication topology  $\mathcal{G}$

*Example 1:* Assume the leader is fixed with  $u_0 = 0$ , the initial state of the leader is  $x_0(0) = 1$ , and those of all followers are  $x(0) = [0 \ 0.8 \ 1.2 \ 1.5 \ 2]^T$ . When the inputs of follower agents follow protocol (4), the agents' states are depicted in Fig. 2 (i). When the inputs of follower agents follow protocol (9), the agents' states are depicted in Fig. 2 (ii). In both Fig. 2 (i) and (ii), the red line corresponds to the state of the leader. It is revealed that following agents' states converge to the leader's asymptotically under both protocols. We can easily obtain that when the leader is fixed with zero input, the leader-following consensus is achieved no matter whether follower agents know the leader's input or not.

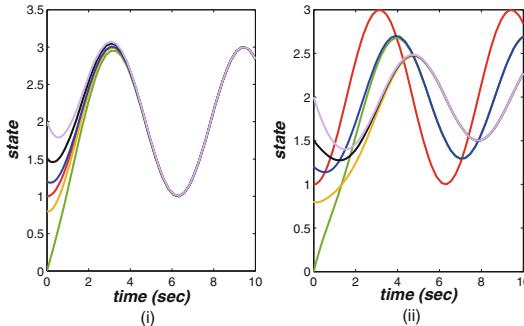
*Example 2:* Assume the input of the leader is fixed with  $u_0 = 2$ , the initial state of the leader is  $x_0(0) = 1$ , and those of all followers are  $x(0) = [0 \ 0.8 \ 1.2 \ 1.5 \ 2]^T$ . When the inputs of follower agents follow protocol (4), the agents' states are depicted in Fig. 3 (i) with red line corresponding to the leader's state, showing that following agents' states converge to the leader's asymptotically. When the inputs of follower agents follow protocol (9), the agents' states are depicted in Fig. 3 (ii) with red line corresponding to the leader's state as well, revealing the differences between follower agents' states and the leader's are bounded.



**Fig. 2.** States of the agents under (i) protocol (4) (ii) protocol (9)



**Fig. 3.** States of the agents under (i) protocol (4) (ii) protocol (9)



**Fig. 4.** States of the agents under (i) protocol (4) (ii) protocol (9)

*Example 3:* Assume the input of the leader is time-varying with  $u_0(t) = \sin(t)$ ,  $x_0(0) = 1$ , and those of all followers are  $x(0) = [0 \ 0.8 \ 1.2 \ 1.5 \ 2]^T$ . When the inputs of follower agents follow protocol (4), the agents' states are depicted in Fig. 4 (i) with red line corresponding to the leader's state, showing that following agents' states converge to the leader's asymptotically. When the inputs of follower agents

follow protocol (9), the agents' states are depicted in Fig. 4 (ii) with red line corresponding to the leader's state as well, revealing the differences between follower agents' states and the leader's are bounded.

## 5 Conclusion

The leader-following consensus problem of multiple single-integrator agents has been studied in this paper, by using a novel linear transformation method together with ISS property. We have considered three cases: 1) the leader's input is pre-given and known by all following agents; 2) the leader's input is unknown; 3) the leader's input is measurable online and transmitted to some of follower agents. We have equivalently transformed the leader-following consensus problem into an ISS problem, by introducing a transformation matrix based on incidence matrix of a virtual leader-rooted spanning tree. Then a necessary and sufficient condition for ensuring the leader-following consensus has been given, which is the Hurwitz stability of  $-(L + D)$ . In order to efficiently check whether  $-(L + D)$  is Hurwitz stable, especially for large-scale multi-agent systems, we have further employed the Hurwitz stability criteria of  $-(L + D)$  based on Metzler matrix theory. Several numerical examples have also been given to verify the correctness of the theoretical results. An important issue for further research is to apply the proposed linear-transformation-based ISS approach to studying the leader-following consensus problem of MASs with more general dynamics.

**Acknowledgement.** This work was supported by the National Natural Science Foundation of China (Nos. 61803007, 61573030), the Rail Transit Joint Funds of Beijing Natural Science Foundation and Traffic Control Technology (No. L171001), and Beijing Municipal Education Commission.

## References

1. Li, Y., Tang, C., Li, K., He, X., Peeta, S., Wang, Y.: Consensus-based cooperative control for multi-platoon under the connected vehicles environment. *IEEE Trans. Intell. Transp. Syst.* **20**(6), 2220–2229 (2019)
2. Li, T., Mallick, M., Pan, Q.: A parallel filtering-communication based cardinality consensus approach for real-time distributed PHD filtering. *IEEE Sensors J.* (2020). <https://doi.org/10.1109/JSEN.2020.3004068>
3. Simpson-Porco, J.W.: On stability of distributed-averaging proportional-integral frequency control in power systems. *IEEE Control Syst. Lett.* **5**(2), 677–682 (2021)
4. Tsitsiklis, J., Athans, M.: Convergence and asymptotic agreement in distributed decision problems. *IEEE Tran. Autom. Control* **29**(1), 42–50 (1984)
5. Vicsek, T., Czirok, A., Ben-Jacob, E., Cohen, I., Shochet, O.: Novel type of phase transition in a system of self-derived particles. *Phys. Rev. Lett.* **75**(6), 1226–1229 (1995)
6. Jadbabaie, A., Lin, J., Morse, A.S.: Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Trans. Autom. Control* **48**(6), 988–1001 (2003)

7. Olfati-Saber, R., Murray, R.M.: Consensus problems in networks of agents with switching topology and time-delays. *IEEE Trans. Autom. Control* **49**(9), 1520–1533 (2004)
8. Cui, Q., Sun, J., Zhao, Z., Zheng, Y.: Second-order consensus for multi-agent systems with time-varying delays based on delay-partitioning. *IEEE Access* **8**, 91227–91235 (2020)
9. Hou, M.W., Wang, J.A., Zhao, Z.C.: Distributed consensus of multi-agent via aperiodic intermittent sampled-data control. In: Chinese Automation Congress, Hangzhou, China, (2019) <https://doi.org/10.1109/CAC48633.2019.8996697>
10. Zhang, Y., Tian, Y.P.: Consensus of data-sampled multi-agent systems with random communication delay and packet loss. *IEEE Trans. Autom. Control* **55**(4), 939–943 (2010)
11. You, X., Hua, C.C., Li, K., Jia, X.: Fixed-time leader-following consensus for high-order time-varying nonlinear multi-agent systems. *IEEE Trans. Autom. Control* (2020). <https://doi.org/10.1109/TAC.2020.3005154>
12. Cheng, D., Wang, J., Hu, X.: An extension of Lasall's invariance principle and its application to multi-agent consensus. *IEEE Trans. Autom. Control* **53**(7), 1765–1770 (2008)
13. Rahmani, A., Ji, M., Mesbahi, M., Egerstedt, M.: Controllability of multi-agent systems from a graph-theoretic perspective. *SIAM J. Control Optim.* **48**(1), 162–186 (2009)
14. Peng, K., Yang, Y.: Leader-following consensus problem with a varying-velocity leader and time-varying delays. *Phys. A* **388**(2–3), 193–208 (2009)
15. Ni, W., Cheng, D.: Leader-following consensus of multi-agent systems under fixed and switching topologies. *Syst. Control Lett.* **59**(3–4), 209–217 (2010)
16. Qu, X., Chen, Y., Aleksandrov, A.Y., Dai, G.: Distributed consensus of large-scale multi-agent systems via linear-transformation-based partial stability approach. *Neurocomputing* **222**(11), 54–61 (2017)
17. Farina, L., Rinaldi, S.: Positive Linear Systems: Theory and Applications. Series on pure and applied mathematics. Wiley-Interscience, New York (2000)
18. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. SIAM, Philadelphia (1994)
19. Narendra, K.S., Shorten, R.: Hurwitz stability of Metzler matrices. *IEEE Trans. Autom. Control* **55**(6), 1484–1487 (2010)



# Proportional Step Perturbation Method MPPT for Boost Circuit of TEG

Feng Ji<sup>(✉)</sup> and John Xu

University of Nottingham, Ningbo, China  
[Feng.Ji@nottingham.edu.cn](mailto:Feng.Ji@nottingham.edu.cn)

**Abstract.** In order to improve the working efficiency of thermoelectric generation (TEG), maximum power point tracking (MPPT) is usually used in the system. The perturbation method is a traditional method for MPPT. The perturbation of the step length is normally fixed. It will lead to be not accurate enough if the step size is too big. If the step size is too small, it will lead to take a long time. Therefore, the proportional step perturbation method MPPT is proposed in the paper. This method can be implemented by using large steps at the beginning to save time. When it is near the maximum power point, small steps can be used to achieve the accurate purpose. The derivation of the method has been described in details in the paper. The simulation model is established based on Matlab. The simulation results show that the method is feasible.

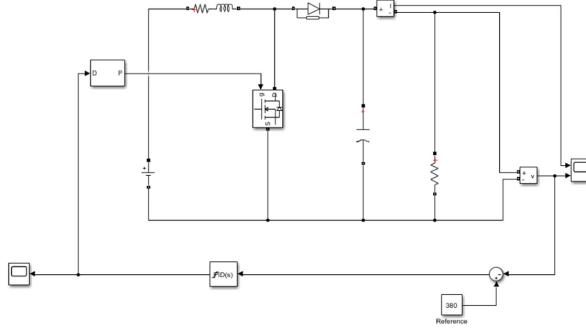
**Keywords:** Boost · MPPT · Proportional step perturbation method

## 1 Introduction

At present, the MPPT methods include fixed voltage method, disturbance observation method, conductance increment method, etc. The constant voltage tracking method is a kind of approximate maximum power tracking method. The constant voltage tracking method has certain power loss. When the temperature changes, the open circuit voltage of the thermoelectric chip will change accordingly. Therefore, the tracking efficiency is not high enough. And the MPPT fixed point is not accurate enough [1–3]. The incremental conductance method is accurate in control and quick in response. It is suitable for the situation where atmospheric conditions change rapidly. However, the hardware requirements are high such as the electronic sensor accuracy. And the response speed of each part of the system is required to be fast. Therefore, the hardware cost of the whole system is also relatively high. When the accuracy of the sensor is limited, there will be errors in the calculation of the increment and instantaneous conductance of the thermoelectric module by the processor. It will inevitably lead to inaccurate tracking [4–6]. For the traditional perturbation method with fixed step length, the output of the thermoelectric panel will be floating around maximum power point if  $\Delta U$  value is too large. If  $\Delta U$  value is small, it can

guarantee the tracking accuracy. But this will need more time. When the maximum power point change frequently, the result will be worse [7–9]. Therefore, the proportional step perturbation method MPPT is proposed in this paper. It can overcome the shortcomings of the traditional method and ensure fast and accurate MPPT.

## 2 Boost Circuit with PID Controller



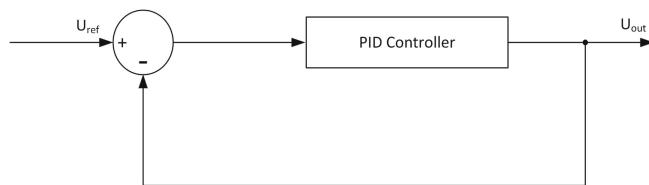
**Fig. 1.** Boost circuit

When the IGBT conducts in the circuit, the current forms a circuit by  $E_0$  through the boost inductance  $L$ . And the inductance  $L$  stores energy. When the IGBT is turned off, the reverse electromotive force generated by the inductance and the DC power supply are superimposed on the working load. Therefore, it can obtain a higher voltage on the load. The function of the diode is to ensure a single direction pass. And the current  $i_L$  continues to flow through the diode  $D$ . By adjusting the on-off period of the switch device IGBT, the output current and voltage on the load side can be adjusted.

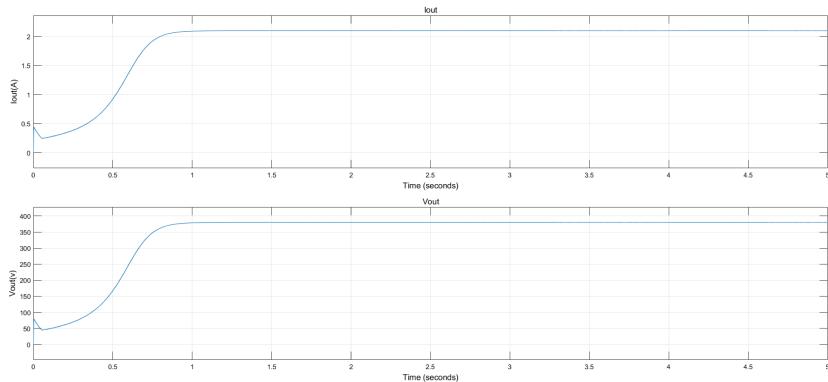
$$U_0 = \frac{t_{on} + t_{off}}{t_{off}} E_0 \quad (1)$$

In the above Eq. (1),  $T$  is the switching period,  $t_{on}$  is the conduction time and  $t_{off}$  is the turn-off time. The circuit adopts closed-loop voltage control as shown in the Fig. 2. It is the schematic diagram of the closed-loop voltage control system. The voltage loop is controlled by the constant value. The voltage controller adopts PID controller. Therefore, the output voltage can obtain accurate and stable value.

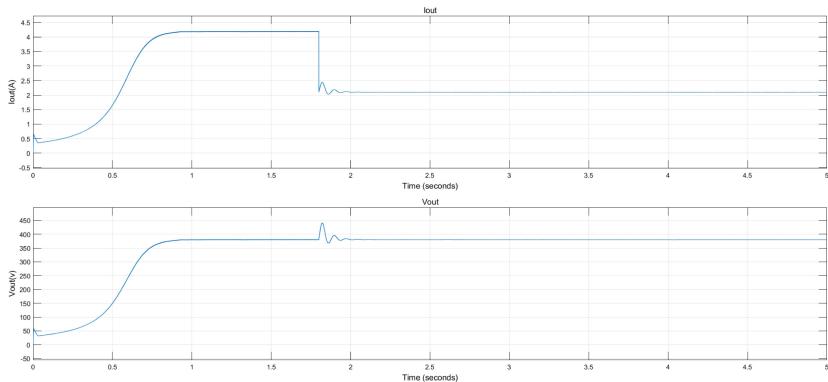
Figure 3 is the output under normal conditions for the Boost circuit. And Fig. 4 is the output under sudden load change for the Boost circuit. The waveform shows that it can obtain stable output under different working load.



**Fig. 2.** Voltage PID control mode



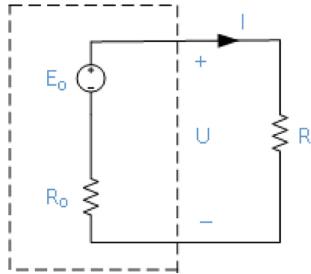
**Fig. 3.** Output under normal conditions of Boost circuit



**Fig. 4.** Output under sudden load change of Boost circuit

### 3 Proportional Step Perturbation MPPT

Because of the slow temperature changing speed, it can be considered that the temperature at both ends of the thermoelectric generator is constant in a control cycle. The circuit model can be equivalent to a voltage source and a fixed resistance in series as shown in Fig. 5.



**Fig. 5.** Thermoelectric circuit model

Figure 5 is a simple closed circuit,  $R_L$  is the load resistance of the external circuit.

The terminal voltage is:

$$U_0 = E_0 - RE_0 \quad (2)$$

Multiply the current  $I$  on both sides of the equation:

$$IU = IE_0 - I^2R_0 \quad (3)$$

Assuming the load is pure resistance,  $I * E_0$  is the total power of the circuit,  $I^2R_0$  is the power of internal resistance and  $I * U$  is the power of output.

$$P = IU = I^2RL = R_L[\frac{E_0}{(R_L + R_0)}]^2 = \frac{R_L E_0^2}{(R_L + R_0)^2} \quad (4)$$

Because of:

$$(R_L + R_0)^2 = (R_L - R_0)^2 + 4R_L R_0 \quad (5)$$

So:

$$P = \frac{R_L E_0^2}{(R_L + R_0)^2} = \frac{R_L E_0^2}{(R_L - R_0)^2 + 4R_L R_0} = \frac{E_0^2}{\frac{(R_L - R_0)^2}{R_L} + 4R_0} \quad (6)$$

Since the EMF  $E_0$  and internal resistance  $R_0$  of the circuit are independent of the external circuit, they can be regarded as constant. When the load resistance

$R_L$  is equal to the internal resistance  $R_0$ , the output power of the circuit reaches the maximum value. Its maximum value is:

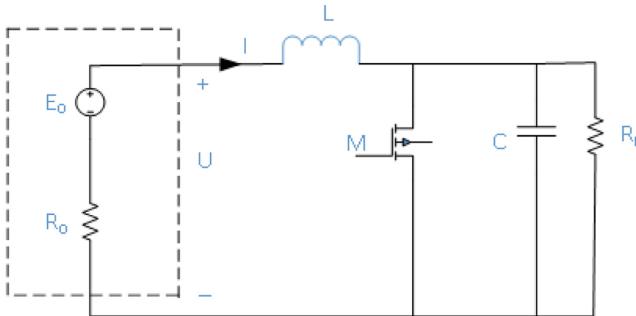
$$P_m = \frac{E_0^2}{4R_0} = \frac{E_0^2}{4R_L} \quad (7)$$

The short-circuit current is  $I_S$ . When  $R_0=R_L$ , the equivalent circuit of the thermoelectric generation is working at the maximum power point (MPP). Then the following two conditions are also met with.

$$U = \frac{E_0}{2} \quad (8)$$

$$I_s = \frac{E_0}{R_0} = \frac{2E_0}{R_0 + R_L} = 2I \quad (9)$$

The thermoelectric generator is connected to DC/DC boost circuit as shown in the Fig. 6. There are two modes in normal operation: MOSFET on mode and MOSFET off mode.



**Fig. 6.** Boost circuit of thermoelectric generation

As shown in the Fig. 6, when MOSFET is on, the voltage  $E_0$  is:

$$E_0 = R_0 I + L \frac{dI}{dt} \quad (10)$$

Divided on both sides of the equation by  $R_0$ :

$$I_s = I + \frac{L}{R_0} \frac{dI}{dt} \quad (11)$$

When the circuit works in MPPT,

$$U_{sc} = \frac{E_0}{2} = I_s \frac{R_0}{2} \quad (12)$$

Because  $E_0$  will change by different temperature, the short-circuit current of thermoelectric generator can be calculated according to the inductance current

value and its changing rate. Then the value of  $E_0$  can be calculated. So as long as the value of control port voltage U is  $E_0/2$ , MPPT control strategy can be achieved.

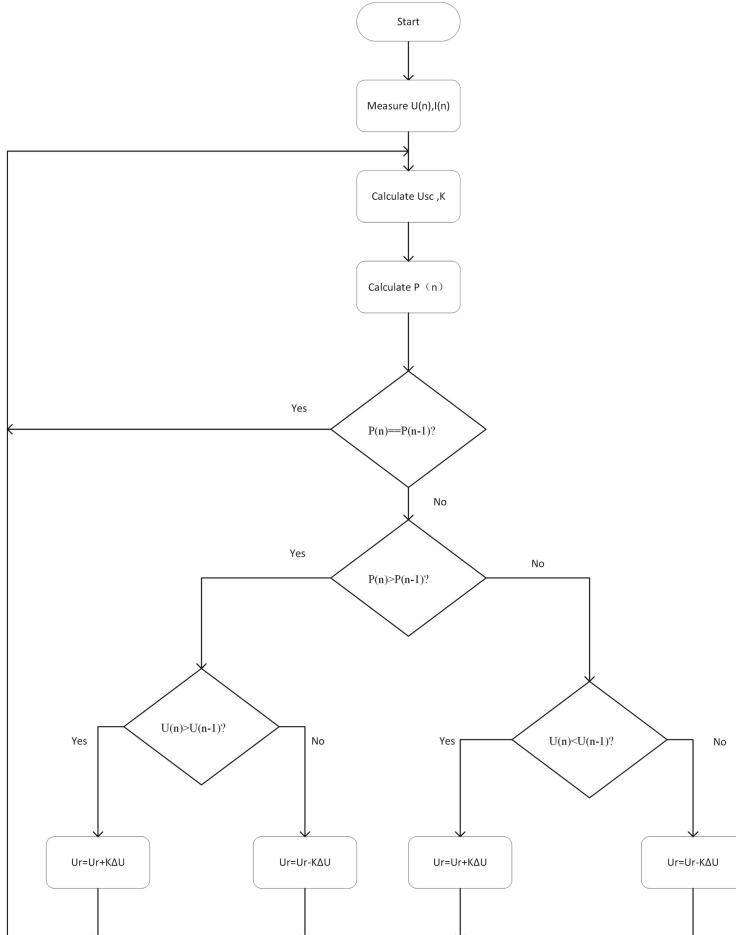
$$K = \frac{|U - U_{sc}|}{U_{sc}} \quad (13)$$

When the working point is on the left of the maximum power point.

$$\frac{dP}{dU} > 0 : U = U + K\Delta U \quad (14)$$

When the working point is the maximum power point.

$$\frac{dP}{dU} = 0 : U = U_m \quad (15)$$



**Fig. 7.** Proportional step perturbation MPPT flow chart

When the working point is on the right of the maximum power point.

$$\frac{dP}{dU} < 0 : U = U - K \Delta U \quad (16)$$

The Fig. 7 is the proportional step perturbation MPPT flow chart. As shown in the Fig. 7, the coefficient K corrects the new value in each cycle. Therefore, it can be ensured that the step size is larger when the working point is farther away from the MPPT point. And the step size is smaller when the working point is closer to the MPPT point.

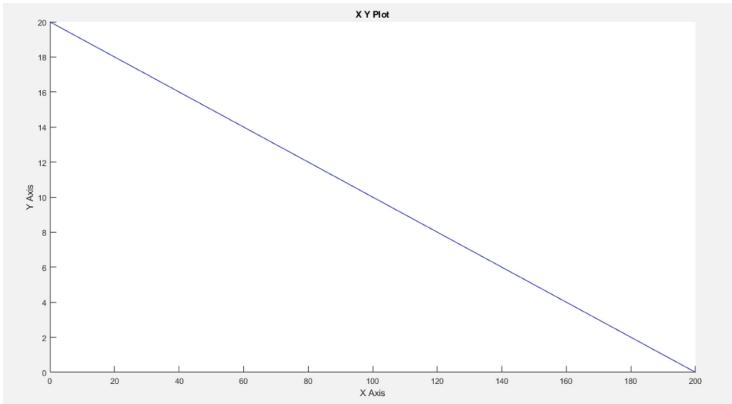
## 4 Simulation and Conclusion

The Fig. 8 is the I/U Waveform of the thermoelectric cell. And the Fig. 9 is the P/U Waveform of the thermoelectric cell.

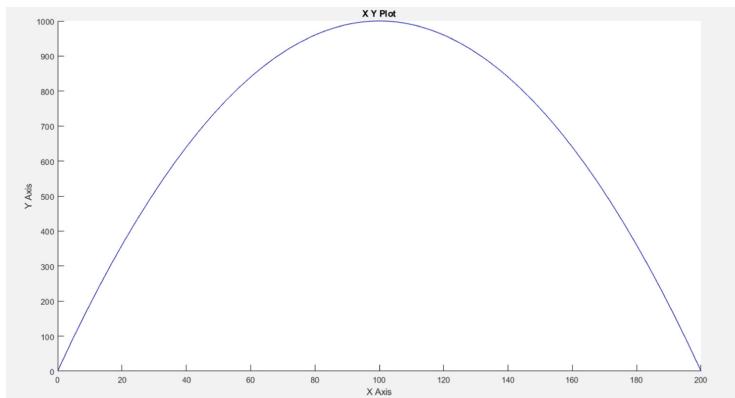
As shown in the Fig. 9, the maximum power point is located near the midpoint of open circuit voltage. Therefore, the closer the voltage is to the midpoint, the smaller voltage step is to achieve accurate control. The further away from the midpoint of the open-circuit voltage, the voltage step can be increased for the purpose of quickly MPPT.

The Fig. 10 is the simulation waveform of MPPT. The temperature changes at 0.03 s. And the maximum power changes from 1,400 W to 3,250 W.

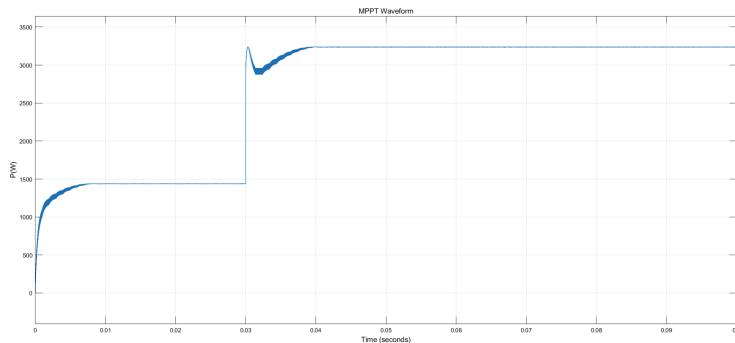
It can be seen from the waveform that the MPPT algorithm based on proportional step perturbation method can lock the maximum power point quickly, accurately and stably. The simulation results show that the design is feasible.



**Fig. 8.** I/U waveform



**Fig. 9.** P/U waveform



**Fig. 10.** MPPT waveform

## References

1. Bai, Y., Kou, B., Chan, C.C.: A simple structure passive MPPT standalone wind turbine generator system. *IEEE Trans. Magn.* **51**(11), 1–4 (2015)
2. Stanzione, S., van Liempd, C., van Schaijk, R., Naito, Y., Yazicioglu, F., Van Hoof, C.: A high voltage self-biased integrated DC-DC buck converter with fully analog MPPT algorithm for electrostatic energy Harvesters. *IEEE J. Solid-State Circuits* **48**(12), 3002–3010 (2013)
3. Murdoch, S., Reynoso, S.: Design and implementation of a MPPT circuit for a Solar UAV. *IEEE Lat. Am. Trans.* **11**(1), 108–111 (2013)
4. Pradhan, R., Subudhi, B.: Double integral sliding mode MPPT control of a photovoltaic system. *IEEE Trans. Control Syst. Technol.* **24**(1), 285–292 (2016)
5. Jiang, R., Han, Y., Zhang, S.: Wide-range, high-precision and low-complexity MPPT circuit based on perturb and observe algorithm. *Electron. Lett.* **53**(16), 1141–1142 (2017)
6. Bond, M., Park, J.: Current-sensorless power estimation and MPPT implementation for thermoelectric generators. *IEEE Trans. Ind. Electron.* **62**(9), 5539–5548 (2015)

7. Lopez-Santos, O., et al.: Analysis, design, and implementation of a static conductance-based MPPT method. *IEEE Trans. Power Electron.* **34**(2), 1960–1979 (2019)
8. Liu, X., Huang, L., Ravichandran, K., Sánchez-Sinencio, E.: A Highly Efficient Reconfigurable Charge Pump Energy Harvester With Wide Harvesting Range and Two-Dimensional MPPT for Internet of Things. *IEEE J. Solid-State Circuits* **51**(5), 1302–1312 (2016)
9. Abdullah, M.A., Al-Hadhrami, T., Tan, C.W., Yatim, A.H.: Towards Green Energy for Smart Cities: Particle Swarm Optimization Based MPPT Approach. *IEEE Access* **6**, 58427–58438 (2018)



# Leader-Following Consensus of Second-Order Networks with a Moving Leader and Nonconvex Input Constraints

Lipo Mo<sup>1(✉)</sup>, Zeyang Xu<sup>1</sup>, and Yongguang Yu<sup>2</sup>

<sup>1</sup> School of Mathematics and Statistics, Beijing Technology and Business University,  
Beijing 100048, People's Republic of China  
[beihangmlp@126.com](mailto:beihangmlp@126.com)

<sup>2</sup> School of Science, Beijing Jiaotong University, Beijing 10044, People's Republic of China

**Abstract.** This paper considers the constrained leader-following consensus problem of second-order multi-agent networks with nonconvex input constraints, where the leader moves with a constant velocity. It is assumed that each agent can only perceive its own nonconvex constraint set and the joint communication graph has a directed spanning tree whose root is the leader. By introducing the constraint operator and the estimator of the leader, a new distributed algorithm is designed. Then, it is proved that the leader-following consensus can be reached under some mild conditions by constructing some auxiliary functions. Finally, a simulation example is given to examine the effectiveness of our results.

**Keywords:** Second-order networks · Multi-agent · Leader-following consensus · Nonconvex constraints

## 1 Introduction

Since the Vicsek model was proposed in 1995, consensus problem of multi-agent networks has became a hot point of the network science field [1]. Motivated by the movement behavior of biological groups, many distributed algorithms were designed to force all agents in the network to reach a consensus and the corresponding convergence was analyzed [2–6]. In physical systems, it is always required that the states of agents are constrained to some constraint sets due to the restriction of realistic environment and objective conditions. In [7, 8], the constrained consensus problems were considered for multi-agent networks with convex constraints. While, some constraints might not be convex, such as the velocity or acceleration of quadrotor. When nonconvex constraints were taken into account, the convergence analysis became more complicated. In [9–11], the nonconvex constrained consensus problems were studied for the multi-agent networks with nonconvex velocity or input constraints by adding a constraint operator.

In reality, all agents in the network are often required to not only reach a consensus, but also track a target, which is called the leader-following problem, such as car-following system [12]. The leader-following problems were considered in [13–15] for multi-agent networks and it was proved that the tracking control could be completed

if there existed at least one directed path from the leader to each follower in the union graph. However, these leader-following consensus results were built under the assumption that all agents were worked in ideal environment and no constraint was considered. When the nonconvex constraints are taken into account, the distributed algorithms and the analysis methods in [13–15] would not work anymore due to the strong nonlinearity of nonconvex constraint operator. Though the nonconvex input constraints were considered in our previous works for the situation of stationary leaders [16, 17], it is still unclear when the leader is moving.

In this paper, we focus on the nonconvex constraints leader-following consensus problem of second multi-agent networks, where the leader moves with a constant velocity and the inputs are constrained to the different nonconvex sets, which can only known by each follower. This paper has the following two contributions. First, a new distributed algorithm is designed, which is composed of a collaborative protocol between agents and a distributed estimate for the velocity of the leader by each agent. Second, because the system matrix doesn't be stochastic anymore when the leader is moving, the methods based on stochastic matrix theory in [9–11, 16, 17] are invalid. We develop a new method to analyze the convergence of the system variables after two coordination transformations. Then, the sufficient conditions for consensus are obtained.

**Notations.**  $\mathbf{R}^n$  represents the real vector space with dimension  $n$ ; Given a matrix  $A$ ,  $A^T$  represents its transpose,  $[A]_{ij}$  represents the  $(i, j)$  entry of matrix  $A$ ; Given vector  $x$ ,  $\|x\|$  means its Euclidean norm; Let  $0 \in Y$  be a bounded closed set, define

$$S_Y(x) = \frac{x}{\|x\|} \max_{0 \leq a \leq \|x\|} \{a | b \frac{ax}{\|x\|} \in Y, b \in [0, 1]\}$$

if  $x \neq 0$  and  $S_Y(0) = 0$ .

## 2 Problem Statement

Consider a second-order multi-agent network with one leader. Each agent is regarded as a node of a directed graph  $G(k) = (V, E(k))$ , where  $V = \{0, 1, 2, \dots, n\}$  is the node set, 0 represents the leader,  $1, 2, \dots, n$  represent the follower agents and  $E(k)$  represents the edge set of  $G(k)$  at time  $k$ . Let  $1 > \bar{\eta} > 0$  be a constant, if the follower agent  $i$  can receive the information of the follower agent  $j$  at time  $k$ , then the edge  $(j, i) \in E(k)$  and its weight  $a_{ij}(k) \geq \bar{\eta}$ , otherwise  $a_{ij}(k) = 0$ . If the follower agent  $i$  can receive the information of the leader, then  $(0, i) \in E(k)$  and its weight  $b_i(k) > \bar{\eta}$ , otherwise  $b_i(k) = 0$ . Suppose the dynamics of the follower agents are described by

$$\begin{aligned} p_i(k+1) &= p_i(k) + v_i(k)T, \\ v_i(k+1) &= v_i(k) + S_{U_i}[u_i(k)]T, \end{aligned} \quad (1)$$

where  $p_i(k), v_i(k), u_i(k) \in \mathbf{R}^r$  are the position, velocity and input of the  $i$  th agent at time  $k$ .  $0 \in U_i \subset \mathbf{R}^r$  is the nonconvex constraint set, which can only be known by agent  $i$ ,  $i = 1, 2, \dots, n$ ,  $T$  is the sampling time. Suppose the dynamics of the leader is as follows

$$p_0(k+1) = p_0(k) + v_0 T, \quad (2)$$

where  $p_0(k) \in \mathbf{R}^r$  is the position of the leader,  $v_0 \in \mathbf{R}^r$  is a constant, representing the velocity of the leader,  $T$  is the sampling time.

*Remark 1.* In system (1), the constraint operator  $S_{U_i}[\cdot]$  is used to describe the nonconvex constraints of the inputs. Any physical system, whose inputs have different maximum magnitude in different directions, can be modeled by system (1), such as quadrotor.

The main task of this paper is to design a distributed algorithm to force all followers track the active leader who moves with a constant velocity. In this paper, we design the following distributed algorithm

$$u_i(k) = S_{Q_i}[w_i(k)] + l_i(\tilde{v}_i(k) - v_i(k)), \quad (3)$$

$$\tilde{v}_i(k+1) = S_{\tilde{V}_i}[\tilde{v}_i(k) + \sum_{j \in N_i(k)} a_{ij}(k)(\tilde{v}_j(k) - \tilde{v}_i(k))T + b_i(k)(v_0 - \tilde{v}_i(k))T], \quad (4)$$

where  $l_i > 0$  is the feedback gain,  $w_i(k) = \sum_{j \in N_i(k)} a_{ij}(k)(p_j(k) - p_i(k)) + b_i(k)(p_0(k) - p_i(k))$ ,  $Q_i = \{x \in \mathbf{R}^r \mid \|x\| \leq M_i\}$  is a constraint set with  $M_i > 0$ ,  $\tilde{V}_i = \{x \in \mathbf{R}^r \mid \|x\| \leq \beta_i\}$  is a constraint set with  $\beta_i > 0$ ,  $N_i(k) = \{j = 1, 2, \dots, n \mid (j, i) \in E(k)\}$  is the neighbor set of agent  $i$ ,  $\tilde{v}_i(k)$  is the estimation of the leader's velocity by agent  $i$  at time  $k$ . In this paper, we assume that  $v_0 \in \tilde{V}_i$  for all  $i$ , which can guarantee that all estimated velocities  $\tilde{v}_i(k)$  could converge to  $v_0$ .

*Remark 2.* In algorithm (3), constraint operator  $S_{Q_i}[\cdot]$  is introduced to guarantee the boundedness of  $w_i(k)$ . In view of the physics, this term must be bounded if the leader-following consensus can be achieved. Equation (4) is used to estimate the velocity of the leader by each agent.

**Assumption 1.** Suppose that  $0 = k_0 < k_1 < k_2 < \dots < k_m < \dots$  is a time sequence, and there exists at least one directed path from the leader to any follower in the union graph of  $G(k_m), G(k_m + 1), \dots, G(k_{m+1} - 1)$ , and  $k_{m+1} - k_m \leq K$  for all  $m$ , where  $K \geq 1$  is a constant.

**Assumption 2.** Suppose  $0 \in U_i$  is the nonconvex bounded closed set,  $\sup_{x \in U_i} \|S_{U_i}[x]\| = \bar{\rho}_i$  and  $\inf_{x \notin U_i} \|S_{U_i}[x]\| = \underline{\rho}_i$ , where  $\bar{\rho}_i > \underline{\rho}_i > 0$  for all  $i$ .

*Remark 3.* Assumption 1 is a standard condition for reaching consensus [13]. Assumption 2 guarantees that the inputs of all followers can vary along arbitrary directions and are bounded [9].

### 3 Main Results

To analyze the consensus of network (1), we first do two model transformations. Let

$$e_i(k) = \begin{cases} \|S_{U_i}[u_i(k)]\| / \|u_i(k)\|, & u_i(k) \neq 0 \\ 1, & u_i(k) = 0, \end{cases}$$

$$\delta_i(k) = \begin{cases} \|S_{Q_i}[w_i(k)]\| / \|w_i(k)\|, & w_i(k) \neq 0 \\ 1, & w_i(k) = 0. \end{cases}$$

Define  $x_i(k) = p_i(k) - p_0(k)$  and  $\bar{y}_i(k) = v_i(k) - v_0$ . Then the closed-loop system can be written as

$$x_i(k+1) = x_i(k) + \bar{y}_i(k)T,$$

$$\begin{aligned} \bar{y}_i(k+1) &= (1 - e_i(k)l_iT)\bar{y}_i(k) + [-e_i(k)\delta_i(k)(b_i(k) + \sum_{j \in N_i(k)} a_{ij}(k))T] \times \\ &\quad x_i(k) + e_i(k)\delta_i(k)\sum_{j \in N_i(k)} a_{ij}(k)Tx_j(k) + e_i(k)l_iTz_i(k), \end{aligned} \quad (5)$$

where  $z_i(k) = \tilde{v}_i(k) - v_0$ . Furthermore, define  $y_i(k) = x_i(k) + \frac{1}{c_i}\bar{y}_i(k)$ ,  $c_i > 0$ . Then the closed-loop system can be changed into

$$\begin{aligned} x_i(k+1) &= (1 - c_iT)x_i(k) + c_iTy_i(k) \\ y_i(k+1) &= h_{ix}(k)x_i(k) + h_{iy}(k)y_i(k) \\ &\quad + \sum_{j \in N_i(k)} h_{ij}(k)x_j(k) + \frac{1}{c_i}e_i(k)l_iTz_i(k), \end{aligned} \quad (6)$$

where  $h_{ix}(k) = e_i(k)l_iT - c_iT - \frac{1}{c_i}e_i(k)\delta_i(k)T(b_i(k) + \sum_{j \in N_i(k)} a_{ij}(k))$ ,  $h_{iy}(k) = 1 + c_iT - e_i(k)l_iT$ ,  $h_{ij}(k) = \frac{1}{c_i}e_i(k)\delta_i(k)a_{ij}(k)T$ ,  $j \in N_i(k)$ .

**Assumption 3.** Suppose  $\bar{\eta} < \frac{1}{T}$ ,  $[L(k)]_{ii} < d_{1i}$ ,  $b_i(k) < d_{2i}$  and  $1 - d_{1i}T - d_{2i}T > 0$ , where  $d_{1i} > 0$  and  $d_{2i} > 0$  for all  $i$ , and  $L(k)$  is the Laplacian of all followers at time  $k$ .

*Remark 4.* Assumption 3 is easy to be satisfied by selecting appropriate sampling time.

**Lemma 1.** Let  $Y \subset \mathbf{R}^r$  be a closed ball centered at the origin,  $x \in \mathbf{R}^r$  and  $y \in Y$ . Then  $\|S_Y[x] - y\|^2 \leq \|x - y\|^2 - \|S_Y[x] - x\|^2$  and  $\|S_Y[x] - y\| \leq \|x - y\|$ .

**Proof.** If  $x \in Y$ , then  $S_Y[x] = x$  and  $\|S_Y[x] - y\| = \|x - y\|$ . If  $x \notin Y$ , from the convexity of  $Y$ , we have the angle between the vectors  $x - S_Y[x]$  and  $y - S_Y[x]$  lies in  $[\frac{\pi}{2}, \pi]$ . Then,  $[S_Y[x] - x]^T [S_Y[x] - y] \leq 0$ . Hence,  $[S_Y[x] - x]^T [x - y] = [S_Y[x] - x]^T [x - S_Y[x]] + [S_Y[x] - x]^T [S_Y[x] - y] \leq -\|S_Y[x] - x\|^2$ . Therefore,  $\|S_Y[x] - y\|^2 = \|S_Y[x] - x\|^2 + \|x - y\|^2 + 2[S_Y[x] - x]^T [x - y] \leq \|x - y\|^2 - \|S_Y[x] - x\|^2$  and  $\|S_Y[x] - y\| \leq \|x - y\|$ .

**Lemma 2.** Suppose Assumptions 1, 2 and 3 hold. Then  $\lim_{k \rightarrow \infty} \|z_i(k)\| = 0$  for all  $i$ .

**Proof.** It follows from Assumption 3 that  $1 - \sum_{j \in N_i(k)} a_{ij}(k)T - b_i(k)T = 1 - [L(k)]_{ii}T - b_i(k)T \geq 1 - d_{1i}T - d_{2i}T > 0$  and  $a_{ij}(k)T \geq \bar{\eta}T$  for all  $j \in N_i(k)$ . Recall that  $z_i(k) = \tilde{v}_i(k) - v_0$ . Let  $V_1(k) = \max_i \|z_i(k)\|$ . According to Lemma 1, we have

$$\|z_i(k+1)\| \leq (1 - b_i(k)T)V_1(k) \leq V_1(k).$$

Hence,  $V_1(k+1) \leq V_1(k)$ , which implies that  $\lim_{k \rightarrow \infty} V_1(k)$  must exist.

If there exists a follower agent  $i_1$ , who can access directly to the information of the leader at  $k' \in [k_m, k_{m+1})$ , then  $b_{i_1}(k') \geq \bar{\eta}$ . Hence,

$$\|z_{i_1}(k'+1)\| \leq (1 - \bar{\eta}T)V_1(k_m) = \lambda_1 V_1(k_m),$$

where  $\lambda_1 = 1 - \bar{\eta}T \in (0, 1)$ .

If there exist a follower agent  $i_2$  and  $k'' > k_m$ , such  $\|z_{i_2}(k'' + 1)\| \leq \lambda_2 V(k_m)$  for some  $\lambda_2 \in (0, 1)$ . Then we can conclude by recursion that

$$\|z_{i_2}(k'' + s)\| \leq [1 - (1 - \lambda_2)(1 - \bar{\eta}T)^s]V_1(k_m)$$

for all  $s \geq 1$ . Take  $\lambda_3 = 1 - (1 - \lambda_2)(1 - \bar{\eta}T)^{nK}$ , then  $\|z_{i_2}(k)\| \leq \lambda_3 V_1(k_m)$  for all  $k'' < k \leq k_{m+n}$ .

If there exist two follower agents  $i_3, i_4$  and  $0 < \lambda_4 < 1$ , such that  $a_{i_3 i_4}(k''') > 0$  for some  $k''' > k_m$  and  $\|z_{i_4}(k''')\| \leq \lambda_4 V_1(k_m)$ . Then

$$\|z_{i_3}(k'' + 1)\| \leq [1 - (1 - \lambda_4)\bar{\eta}T]V_1(k_m) = \lambda_5 V_1(k_m),$$

where  $\lambda_5 = 1 - (1 - \lambda_4)\bar{\eta}T \in (0, 1)$ .

Noted that  $\lambda_1, \lambda_2, \lambda_3, \lambda_5$  are independent on  $m$  and each agent, by Assumption 1, there must exist a follower  $i_1$ , who can access to the information of the leader at some time in  $[k_m, k_{m+1}]$ . Hence, based on the above discussion, there exist  $\bar{\lambda}_1 \in (0, 1)$ , such that  $\|z_{i_1}(k_{m+s})\| \leq \bar{\lambda}_1 V_1(k_m)$  for all  $1 \leq s \leq n$ . Similarly, there must exist a follower  $i_2 \neq i_1$ , who can access to the information of agent  $i_1$  or the leader at some moment in  $[k_{m+1}, k_{m+2}]$ . Hence, based on the above discussion, there exists  $0 < \bar{\lambda}_2 < 1$ , such that  $\|z_{i_2}(k_{m+s})\| \leq \bar{\lambda}_2 V_1(k_m)$  for all  $2 \leq s \leq n$ . Similarly, for each  $3 \leq t \leq n$ , there exists  $0 < \bar{\lambda}_t < 1$ , such that  $\|z_{i_t}(k_{m+s})\| \leq \bar{\lambda}_t V_1(k_m)$  for all  $t \leq s \leq n$ . Take  $\lambda = \max\{\bar{\lambda}_1, \bar{\lambda}_2, \dots, \bar{\lambda}_n\}$ , which is independent on  $m$ , then  $\|z_i(k_{m+n})\| \leq \lambda V_1(k_m)$  for all  $i$ . Therefore,  $V_1(k_{m+n}) \leq \lambda V_1(k_m)$ . By the incursion, we can conclude that  $V_1(k_{m+sn}) \leq \lambda^s V_1(k_m)$ , which implies that  $\lim_{s \rightarrow \infty} V_1(k_{m+ns}) = 0$ . Since  $\lim_{k \rightarrow \infty} V_1(k)$  exists, we can conclude that  $\lim_{k \rightarrow \infty} V_1(k) = 0$ . Hence,  $\lim_{k \rightarrow \infty} \|z_i(k)\| = 0$  for all  $i$ .

*Remark 5.* It follows from Lemma 2 that the estimated velocities of the leader by all followers would converge to the velocity of the leader.

**Assumption 4.** Suppose that  $(\underline{\rho}_i - 2M_i)/(2\beta_i) < l_i < (\underline{\rho}_i - M_i)/\beta_i$  and  $[(d_{1i} + d_{2i})/\mu_i] + 1 < l_i < 1/T$  for all  $i$ , where  $\mu_i = (\underline{\rho}_i - \bar{M}_i)/\bar{M}_i$ ,  $\bar{M}_i = M_i + l_i\beta_i$ .

*Remark 6.* Assumption 4 is a technical condition. In the designing process of the parameters, we can first select  $l_i$  to guarantee Assumption 4 is satisfied, then choosing proper sampling time  $T$  to guarantee that Assumption 3 is satisfied.

**Lemma 3.** If Assumptions 1, 2, 3 and 4 hold, then there exists  $K_1 > 0$ , such that  $\mu_i \leq e_i(k) \leq 1$  for all  $k \geq K_1$ .

**Proof.** Note that  $\|\tilde{v}_i(k)\| \leq \beta_i$ , we have

$$\|v_i(k+1)\| \leq (1 - e_i(k)l_iT)\|v_i(k)\| + \bar{M}_i T$$

and

$$e_i(k) \geq \frac{\underline{\rho}_i}{\bar{M}_i + l_i\|v_i(k)\|},$$

when  $S_{U_i}[S_{Q_i}[w_i(k)] + l_i(\tilde{v}_i(k) - v_0)] \neq S_{Q_i}[w_i(k)] + l_i(\tilde{v}_i(k) - v_0)$ . Hence,

$$\|v_i(k+1)\| - \|v_i(k)\| \leq -\left(\frac{l_i T \underline{\rho}_i}{\bar{M}_i / \|v_i(k)\| + l_i} - \bar{M}_i T\right).$$

If  $\frac{l_i T \rho_i}{\bar{M}_i / \|v_i(k)\| + l_i} - \bar{M}_i T > \varepsilon_0$ , i.e.,  $\|v_i(k)\| > \frac{\bar{M}_i(\bar{M}_i T + \varepsilon_0)}{l_i(\underline{\rho}_i T - \bar{M}_i T - \varepsilon_0)}$ , where  $\varepsilon_0 > 0$  is small enough, then

$$\|v_i(k+1)\| - \|v_i(k)\| < -\varepsilon_0.$$

Hence, there must exist  $K_1 > 0$ , such that

$$\|v_i(k)\| \leq \frac{\bar{M}_i(\bar{M}_i T + \varepsilon_0)}{l_i(\underline{\rho}_i - \bar{M}_i T - \varepsilon_0)}$$

for all  $k \geq K_1$ . Therefore,

$$e_i(k) \geq \frac{\underline{\rho}_i T - \bar{M}_i T - \varepsilon_0}{\bar{M}_i T}.$$

Since  $\varepsilon_0$  can be selected small arbitrarily, we have  $e_i(k) \geq \mu_i$ . On the other hand,  $e_i(k) = 1$  when  $S_{U_i}[S_{Q_i}[w_i(k)] + l_i(\tilde{v}_i(k) - v_0)] = S_{Q_i}[w_i(k)] + l_i(\tilde{v}_i(k) - v_0)$ . Therefore,  $\mu_i \leq e_i(k) \leq 1$  for all  $i$  and  $k \geq K_1$ .

**Lemma 4.** Under Assumptions 1, 2, 3 and 4. There exists  $0 < \bar{\mu}_i < 1$ , such that  $\bar{\mu}_i \leq \delta_i(k) \leq 1$  for all  $i$  and  $k$ .

**Proof.** Let  $V_2(k) = \max_i \{\|x_i(k)\|, \|y_i(k)\|\}$ . Then

$$\|x_i(k+1)\| \leq (1 - c_i T) \|x_i(k)\| + c_i T \|y_i(k)\| \leq V_2(k).$$

By applying Lemma 2, similar to the proof of Lemma 2.1.2 in [18], we can prove that there exist  $0 < \eta_1 < 1$  and  $C_1 > 0$ , such that  $\|z_i(k)\| \leq C_1 \eta_1^k$  for all  $i$  and  $k$ . Hence,

$$\|y_i(k+1)\| \leq V_2(k) + \frac{l_i T}{c_i} C_1 \eta_1^k \leq V_2(k) + \bar{C}_1 \eta_1^k,$$

where  $\bar{C}_1 = \max_i \{\frac{l_i T}{c_i} C_1\}$ . Therefore,

$$V_2(k+1) \leq V_2(k) + \bar{C}_1 \eta_1^k \leq V_2(0) + \sum_{i=1}^k \bar{C}_1 \eta_1^i \leq V_2(0) + \frac{\bar{C}_1 \eta_1}{1 - \eta_1}.$$

Thus,

$$\begin{aligned} & \left\| \sum_{j \in N_i(k)} a_{ij}(p_j(k) - p_i(k)) + b_i(k)(p_0(k) - p_i(k)) \right\| \\ &= \left\| \sum_{j \in N_i(k)} a_{ij}(x_j(k) - x_i(k)) - b_i(k)x_i(k) \right\| \\ &\leq (2d_{1i} + d_{2i})(V_2(0) + \frac{\bar{C}_1 \eta_1}{1 - \eta_1}). \end{aligned}$$

If  $\|w_i(k)\| \leq M_i$ , then  $\delta_i(k) = 1$ . Otherwise,

$$\delta_i(k) \geq \frac{M_i}{(2d_{1i} + d_{2i})(V_2(0) + \frac{\bar{C}_1 \eta_1}{1 - \eta_1})}.$$

Take

$$\bar{\mu}_i = \min\left\{1, \frac{M_i}{(2d_{1i} + d_{2i})(V_2(0) + \frac{\bar{C}_1 \eta_1}{1 - \eta_1})}\right\}.$$

Then  $\bar{\mu}_i \leq \delta_i(k) \leq 1$  for all  $i$  and  $k$ .

To analyze the stability of (6), we need to consider the asymptotic behavior of the following auxiliary system.

$$\begin{aligned}\tilde{x}_i(k+1) &= (1 - c_i T) \tilde{x}_i(k) + c_i T \tilde{y}_i(k) \\ \tilde{y}_i(k+1) &= h_{iy}(k) \tilde{y}_i(k) + h_{ix}(k) \tilde{x}_i(k) + \sum_{j \in N_i(k)} h_{ij}(k) \tilde{x}_j(k),\end{aligned}\quad (7)$$

where  $\tilde{x}_i(k), \tilde{y}_i(k) \in \mathbf{R}^r$  are the auxiliary variables.

**Lemma 5.** Under Assumptions 1, 2, 3 and 4. Consider system (7). Then  $\lim_{k \rightarrow \infty} \|\tilde{x}_i(k)\| = 0$  and  $\lim_{k \rightarrow \infty} \|\tilde{y}_i(k)\| = 0$ .

**Proof.** Taking  $c_i = d_{1i} + d_{2i}$ , it follows from Assumption 4 that  $h_{iy}(k) \geq 1 - c_i T - l_i T > 0$ ,  $h_{ix}(k) \geq \mu_i l_i T - c_i T - \mu_i(d_{2i} + d_{1i})T/c_i > 0$  and  $h_{ij}(k) \geq 0$ . Define  $V_3(k) = \max_i \{\|\tilde{x}_i(k)\|, \|\tilde{y}_i(k)\|\}$ . Then for  $k \geq K_1$ , we have

$$\begin{aligned}\|\tilde{x}_i(k+1)\| &\leq (1 - c_i T) \|\tilde{x}_i(k)\| + c_i T \|\tilde{y}_i(k)\| \leq V_3(k) \\ \|\tilde{y}_i(k+1)\| &\leq (1 - \frac{1}{c_i} e_i(k) \delta_i(k) b_i(k) T) V_3(k) \leq V_3(k).\end{aligned}$$

Hence,  $V_3(k+1) \leq V_3(k)$ , which suggests that  $\lim_{k \rightarrow \infty} V_3(k)$  must exist.

In each interval  $[k_m, k_{m+1})$ , there must exists a follower agent  $i_1$ , who can access directly to the information of the leader at  $k' \in [k_m, k_{m+1})$ , then  $b_{i_1}(k') \geq \bar{\eta}$  and

$$\|\tilde{y}_{i_1}(k'+1)\| \leq \sigma_1 V_3(k_m),$$

where  $\mu = \min_i \{\mu_i\}$ ,  $\bar{\mu} = \min_i \{\bar{\mu}_i\}$ ,  $c = \max_i \{c_i\}$  and  $\sigma_1 = 1 - \mu \bar{\mu} \bar{\eta} T / c \in (0, 1)$ . Then, by recursion, we have

$$\|\tilde{y}_{i_1}(k'+s)\| \leq [1 - (1 - \xi_1)^{s-1} (1 - \sigma_1)] V_3(k_m)$$

for all  $s \geq 1$ , where  $\xi_1 = (l - \bar{c})T \in (0, 1)$ ,  $l = \max_i \{l_i\}$ ,  $\bar{c} = \min_i \{c_i\}$ . In addition,

$$\|\tilde{x}_{i_1}(k'+2)\| \leq [1 - \bar{c}T(1 - \sigma_1)] V_3(k_m).$$

Then, by recursion, we have

$$\|\tilde{x}_{i_1}(k'+s)\| \leq [1 - \bar{c}T(1 - \xi_1)^{s-2} (1 - \sigma_1)] V_3(k_m)$$

for all  $s \geq 2$ . Take  $\bar{\sigma}_1 = 1 - \bar{c}T(1 - \xi_1)^{(2n+1)K} (1 - \sigma_1) \in (0, 1)$ , which is independent on  $m$  and  $i_1$ , then  $\|\tilde{x}_{i_1}(k)\| \leq \bar{\sigma}_1 V_3(k_m)$  and  $\|\tilde{y}_{i_1}(k)\| \leq \bar{\sigma}_1 V_3(k_m)$  for all  $k \in [k_{m+2}, k_{m+2n}]$ .

If there exists  $i_2 \neq i_1$ , such that  $a_{i_2 i_1}(k'') > 0$  for some  $k' < k'' < k_{m+3}$ . Then

$$\|\tilde{y}_{i_2}(k''+1)\| \leq [1 - (1 - \bar{\sigma}_1) \frac{1}{c_{i_2}} e_{i_2}(k'') \delta_{i_2}(k'') a_{i_2 i_1}(k'')] V_3(k_m) \leq \sigma_2 V_3(k_m),$$

where  $\sigma_2 = 1 - (1 - \bar{\sigma}_1) \frac{1}{c} \mu \bar{\mu} \bar{\eta} \in (0, 1)$ . By recursion, we can arrive at

$$\|\tilde{y}_{i_2}(k''+s)\| \leq [1 - (1 - \xi_1)^{s-1} (1 - \sigma_2)] V_3(k_m)$$

for all  $s \geq 1$  and

$$\|\tilde{x}_{i_2}(k'' + s)\| \leq [1 - \bar{c}T(1 - \xi_1)^{s-2}(1 - \sigma_2)]V_3(k_m)$$

for all  $s \geq 2$ . Take  $\bar{\sigma}_2 = 1 - \bar{c}T(1 - \xi_1)^{(2n+1)K}(1 - \sigma_2) \in (0, 1)$ , which is independent on  $m$  and  $i_2$ . Then  $\|\tilde{x}_{i_2}(k)\| \leq \bar{\sigma}_2 V_3(k_m)$  and  $\|\tilde{y}_{i_2}(k)\| \leq \bar{\sigma}_2 V_3(k_m)$  for all  $k \in [k_{m+4}, k_{m+2n}]$ . Similar to the analysis of Lemma 2, we can conclude that there must exist  $\bar{\sigma} \in (0, 1)$ , such that  $\|\tilde{x}_i(k_{m+2n})\| \leq \bar{\sigma} V_3(k_m)$  and  $\|\tilde{y}_i(k_{m+2n})\| \leq \bar{\sigma} V_3(k_m)$ . Hence,  $V(k_{m+2n}) \leq \bar{\sigma} V_3(k_m)$ . By incursion, we have  $V_3(k_{m+2ns}) \leq \bar{\sigma}^s V_3(k_m)$ , which implies that  $\lim_{s \rightarrow \infty} V_3(k_{m+2ns}) = 0$ . Note that  $\lim_{k \rightarrow \infty} V_3(k)$  exists, we arrive at  $\lim_{k \rightarrow \infty} V_3(k) = 0$ . Therefore,  $\lim_{k \rightarrow \infty} \|\tilde{x}_i(k)\| = 0$  and  $\|\tilde{y}_i(k)\| = 0$  for all  $i$ .

Next, we will prove the stability of system (6) based on the above lemmas.

**Theorem 1.** Under Assumptions 1, 2, 3 and 4. The constrained leader-following consensus of systems (1) and (2) can be achieved by the distributed algorithms (3) and (4).

**Proof.** Let  $\Phi(k)$  be a  $2n \times 2n$  matrix, where  $[\Phi(k)]_{2i-1, 2i-1} = 1 - c_i T$ ,  $[\Phi(k)]_{2i-1, 2i} = c_i T$ ,  $[\Phi(k)]_{2i, 2i-1} = h_{ix}(k)$ ,  $[\Phi(k)]_{2i, 2i} = h_{iy}(k)$ ,  $[\Phi(k)]_{2i, 2j-1} = h_{ij}(k)$ ,  $j \in N_i(k)$ ,  $i = 1, 2, \dots, n$ , and other entries are zeros. Define

$$\theta(k) = [x_1(k)^T, y_1(k)^T, x_2(k)^T, y_2(k)^T, \dots, x_n(k)^T, y_n(k)^T]^T$$

and

$$\phi(k) = [0, \frac{1}{c_1} e_1(k) l_1 T z_1(k)^T, \dots, 0, \frac{1}{c_n} e_n(k) l_n T z_n(k)^T]^T.$$

Then the closed-loop system (6) can be rewritten as

$$\theta(k+1) = [\Phi(k) \otimes I_r] \theta(k) + \phi(k).$$

Let  $\Psi(k, s) = \Phi(k) \Phi(k-1) \cdots \Phi(s)$ . By Lemma 5, we can conclude that  $\lim_{k \rightarrow \infty} \Psi(k, s) = 0$  for all fixed  $s$ . It is easy to see that all nonzero entries of  $\Phi(k)$  have a strictly positive lower bound. Similar to the proof of Lemma 2.1.2 in [18], we can prove that there exist  $0 < \eta_2 < 1$  and  $C_2 > 0$ , such that  $\|[\Psi(k, s)]_{ij}\| \leq C_2 \eta_2^{k-s}$  for all  $i$  and  $j$ . It follows from Lemma 2 that  $\lim_{k \rightarrow \infty} \|\phi_i(k)\| = 0$ . For any  $\varepsilon > 0$ , there must exist  $s > 0$ , such that  $\|\phi_i(k)\| < \varepsilon$  for all  $k \geq s$  and  $i$ . Thus,

$$\|\theta_i(k)\| \leq \sum_{j=1}^{2n} C_2 \eta_2^{k-s} \|\theta_j(s)\| + \sum_{j=1}^{2n} C_2 \varepsilon \frac{1 - \eta_2^{k-s-1}}{1 - \eta_2} + \varepsilon.$$

Hence,  $\lim_{k \rightarrow \infty} \|\theta_i(k)\| \leq \frac{2nC_1}{1-\eta_2} \varepsilon + \varepsilon$ . By the arbitrariness of  $\varepsilon$ , we have  $\lim_{k \rightarrow \infty} \|\theta_i(k)\| = 0$ , i.e.,  $\lim_{k \rightarrow \infty} \|x_i(k)\| = \lim_{k \rightarrow \infty} \|y_i(k)\| = 0$ , which means

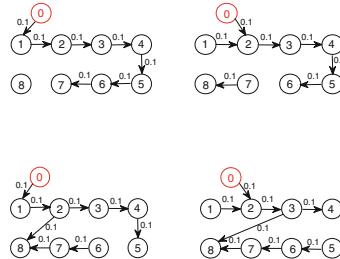
$$\lim_{k \rightarrow \infty} \|p_i(k) - p_0(k)\| = \lim_{k \rightarrow \infty} \|v_i(k) - v_0\| = 0$$

for all  $i$ . Therefore, the constrained leader-following consensus of systems (1) and (2) is achieved.

*Remark 7.* In [19], the leader-following consensus problem was considered for discrete-time multi-agent network with saturation input by constructing square Lyapunov function, which was the convex constraint in essence. In addition, it was assumed that the communication graph was fixed and undirected. However, this paper considers the situation of nonconvex input constraints and switching directed communication graph. Due to the strong nonlinearity and time-variation characteristic of the system matrix, the method of Lyapunov function in [19] can not be used directly to analyze the stability of our system. We directly analyze the convergence of variables after two coordination transformation by introducing some auxiliary functions.

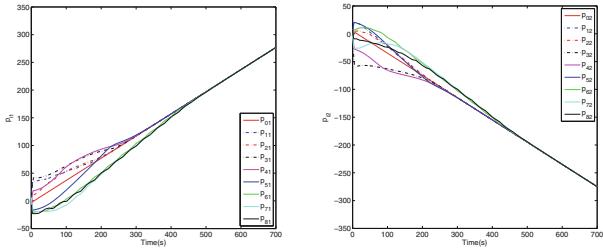
## 4 Simulations

A numerical example is given to show the correctness of the theoretical results in this section. Consider the constrained leader-following consensus problem of second-order multi-agent networks (1) and (2) with the distributed algorithm (3) and (4) over the switching topology that shown in Fig. 1, where node 0 represents the leader, node 1–8 represent the followers. Take  $T = 0.2s$ ,  $l_i = 1.8$ ,  $Q_i = \{x \in \mathbf{R}^2 | \|x\| \leq 0.5\}$ ,  $\tilde{V}_i = \{x \in \mathbf{R}^2 | \|x\| \leq 0.6\}$ , i.e.,  $M_i = 0.5$ ,  $\beta_i = 0.6$ , and take  $\rho_i = 2.5$ ,  $\bar{\rho}_i = 3$ ,  $U_i = \{x = (x_1, x_2)^T \in \mathbf{R}^2 | \|x\| \leq 2.5, |\frac{x_2}{x_1}| \geq 1.2\} \cup \{x = (x_1, x_2)^T \in \mathbf{R}^2 | \|x\| \leq 3, |\frac{x_2}{x_1}| \leq 1.2\} \cup \{(0, 0)\}$ ,  $i = 1, 2, \dots, 8$ , it is easy to verify that Assumptions 1, 2, 3 and 4 are satisfied.

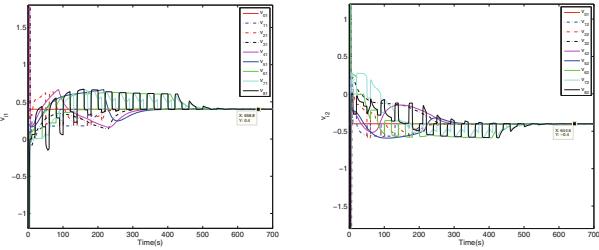


**Fig. 1.** The switching communication topologies

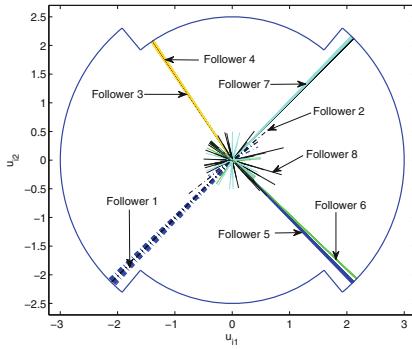
The position states trajectories of all agents are given in Fig. 2, from which we could see that  $\lim_{k \rightarrow \infty} \|p_i(k) - p_0(k)\| = 0$  for all  $i$ ; The velocities trajectories of all agents are given in Fig. 3, from which we could see that  $\lim_{k \rightarrow \infty} v_i(k) = v_0 = (0.4, -0.4)^T$  for all  $i$ ; Fig. 4 depicts the control inputs of all followers in phase plane, which shows that the control inputs of each agent are constrained in the nonconvex set all the time.



**Fig. 2.** The position states of all agents.



**Fig. 3.** The velocities change of all agents.



**Fig. 4.** Control inputs of all followers in phase plane.

## 5 Conclusions

In this paper, we studied the constrained leader-following consensus problem of second-order multi-agent networks, where the leader was assumed to move with a constant velocity and the input of each agent was assumed to constrained to lie in some non-convex constraint set. A new distributed algorithm with estimator of the velocity of the leader was proposed firstly. Two coordination transformations were adopted to change the consensus problem into the stability problem of a new system. Then, by constructing

some auxiliary functions, we proved that the constrained leader-following consensus can be reached by our proposed algorithm under some mild conditions.

**Acknowledgments.** This work is supported by National Natural Science Foundation (NNSF) of China (Grant Nos. 61973329 and 61772063) and the Beijing Natural Science Foundation (Grant Nos. Z180005 and 9192008).

## References

1. Vicsek, T., Czirók, A., Ben-Jacob, C.E.I.I., Shochet, O.: Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.* **75**(6), 1226–1229 (1995)
2. Moreau, L.: Stability of multiagent systems with time-dependent communication links. *IEEE Trans. Autom. Control* **50**(2), 169–182 (2005)
3. Ren, W., Beard, R.W.: Consensus seeking in multi-agent systems under dynamically changing interaction topologies. *IEEE Trans. Autom. Control* **50**(5), 655–661 (2005)
4. Hong, Y., Gao, L., Cheng, D., Hu, J.: Lyapunov-based approach to multiagent systems with switching jointly connected interconnection. *IEEE Trans. Autom. Control* **45**(9), 943–948 (2007)
5. Xiao, F., Wang, L.: State consensus for multi-agent systems with switching topologies and time-varying delays. *Int. J. Control.* **79**(10), 1277–1284 (2006)
6. Lin, P., Jia, Y.: Consensus of second-order discrete-time multi-agent systems with nonuniform time-delays and dynamically changing topologies. *Automatica* **52**(5), 2154–2158 (2009)
7. Nedić, A., Ozdaglar, A., Parrilo, P.A.: Constrained consensus and optimization in multi-agent networks. *IEEE Trans. Autom. Control* **55**(4), 922–938 (2010)
8. Liu, Z., Chen, Z.: Discarded consensus of network of agents with state constraint. *IEEE Trans. Autom. Control* **57**(11), 2869–2874 (2012)
9. Lin, P., Ren, W., Gao, H.: Distributed velocity-constrained consensus of discrete-time multi-agent systems with nonconvex constraints, switching topologies, and delays. *IEEE Trans. Autom. Control* **62**(11), 5788–5794 (2017)
10. Lin, P., Ren, W., Yang, C., Gui, W.: Distributed consensus of second-order multi-agent systems with nonconvex velocity and control input constraints. *IEEE Trans. Autom. Control* **63**(4), 1171–1176 (2018)
11. Mo, L., Lin, P.: Distributed consensus of second-order multiagent systems with nonconvex input constraints. *Int. J. Robust Nonlin.* **28**(11), 3657–3664 (2018)
12. Jiang, R., Wu, Q., Zhu, Z.: Full velocity difference model for a car-following theory. *Phys. Rev. E* **64**, 017101–1 (2001)
13. Hong, Y., Hu, J., Gao, L.: Tracking control for multi-agent consensus with an active leader and variable topology. *Automatica* **42**, 1177–1182 (2006)
14. Hu, J., Hong, Y.: Leader-following coordination of multi-agent systems with coupling time delays. *Phys. A* **374**, 853–863 (2007)
15. Mo, L., Niu, G., Pan, T.: Consensus of heterogeneous multi-agent systems with switching jointly-connected interconnection. *Phys. A* **427**, 132–140 (2015)
16. Yang, C., Duan, M., Lin, P., Ren, W., Gui, W.: Distributed containment control of continuous-time multi-agent systems with nonconvex control input constraints. *IEEE Trans. Ind. Electron.* **66**(10), 7927–7934 (2019)
17. Huang, Y., Duan, M., Mo, L.: Multiagent containment control with nonconvex states constraints, nonuniform time delays, and switching directed networks. *IEEE Trans. Neural Network Learn. Syst.* (2019). <https://doi.org/10.1109/TNNLS.2019.2955678>

18. Guo, L.: Time-varying stochastic systems. Jilin Science and Technology Press, Jilin, China (1993)
19. Wang, Q., Gao, H., Yu, C.: Global leader-following consensus of discrete-time linear multi-agent systems subject to actuator saturation. In: Proceedings of the 2013 Australian Control Conference, pp. 360–363 ,Perth, Australia (2013)



# A Novel Deep Learning Ensemble Model with Secondary Decomposition for Short-Term Electricity Price Forecasting

Na Chen<sup>1</sup>, Xueqing Yang<sup>2(✉)</sup>, Yiming Gan<sup>1,3(✉)</sup>, Wuneng Zhou<sup>1,4</sup>, and Hangyang Cheng<sup>1</sup>

<sup>1</sup> College of Information Sciences and Technology, Donghua University, Shanghai 201620, China

<sup>2</sup> Educational Technology Center, Donghua University, Shanghai 200051, China  
etdaqing@163.com

<sup>3</sup> Guangdong Polytechnic College, Foshan 528041, China  
ygan@elivtex.com

<sup>4</sup> Engineering Research Center of Digitized Textile and Fashion Technology, Donghua University, Shanghai, China

**Abstract.** Accurate electricity price forecasting plays a crucial role in the operation and development of the electricity market. In this paper, a novel hybrid model based on hybrid mode decomposition (HMD), convolutional long short term memory network (CNNLSTM), Elman neural network, and Bayesian optimization (BO) is proposed to forecast the electricity price. HMD is used to deeply decompose data into several subsequences, which consists of complete ensemble empirical mode decomposition with adaptive noise, sample entropy and empirical wavelet transform. CNNLSTM and Elman are adopted to forecast the subsequences. Besides, BO is introduced to optimize parameters. Finally, two case studies are taken to justify the effectiveness of the proposed forecasting model. The results show that the proposed model can possess significantly superior forecasting performance.

**Keywords:** Electricity price forecasting · HMD · CNNLSTM · Elman · BO

## 1 Introduction

With the continuous innovation of trading mode in the global power market, high-precision electricity price forecasting has attracted worldwide attention [1]. In this field, scholars proposed many electricity price forecasting models as divided into the following three fundamental models: physical models, statistical models and artificial intelligence models [2].

The physical models are simple methods that approximately simulate the changes in electricity price by using basic physical information [3,4]. However,

these models require a large amount of computational cost to establish appropriate physical equations.

The statistical models established by mathematical statistical methods are based on historical electricity price data to further speculate on the future development trend of the market [5,6]. But, the statistical models are mainly used for linear sequence forecasting, there will be a great deviation in the forecasting for nonlinear sequence.

Under the trend of the rapid development of computer information technology, artificial intelligence algorithms have been successfully applied in the field of time series forecasting. Common artificial intelligence algorithms include artificial neural networks (ANN), convolutional neural networks (CNN), and recurrent neural networks (RNN). RNN can learn the nonlinear features of sequences with high efficiency [7,8].

According to the results of time series forecasting experiments, the hybrid algorithms usually have better prediction performance than single algorithms [9]. The algorithms commonly used for time series preprocessing include empirical wavelet transform (EWT) [10], complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) [11].

In the field of time series prediction, parameter optimization has a great influence on the performance of the forecasting model. Especially, Bayesian optimization (BO) algorithm is widely used in the industry as an excellent super-parameter tuning method for machine learning models [12].

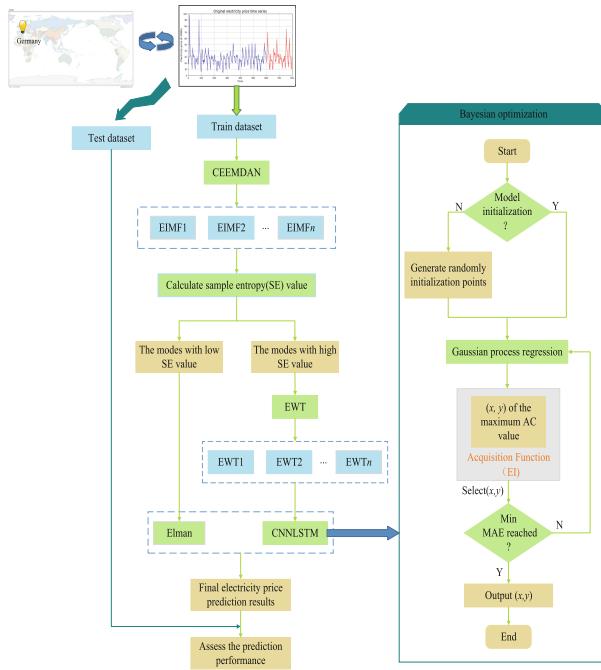
The main contributions of this paper are as follows: (1) A novel multi-step decomposition algorithm is proposed, which can effectively extract the wave characteristics and attributes of electricity price series. (2) The proposed CNNLSTM model has good feature extraction performance and generalization performance. (3) BO algorithm is introduced to optimize the parameters of CNNLSTM for improving forecasting performance.

## 2 The Proposed HMD-CNNLSTM-Elman Model

### 2.1 The Framework of the HMD-CNNLSTM-Elman Model

To obtain authentic and reliable forecasting results, this paper will introduce a hybrid model based on HMD, Elman, CNNLSTM and BO for electricity price forecasting. The overall framework of the HMD-CNNLSTM-Elman model is shown in Fig. 1. The whole process can be summarized as the following three steps:

- (1) CEEMDAN is used to decompose the original electricity price data, SE is employed to measure the complexity of decomposed subsequences, afterwards, the stable subsequences with small SE values are predicted by Elman. Section 2.2 introduces the HMD in detail, and Elman is in Sect. 2.4.
- (2) The complex subsequences need to be further decomposed by EWT, and then which are predicted by CNNLSTM. Section 2.3 describes the details of the CNNLSTM.



**Fig. 1.** The framework of the HMD-CNNLSTM-Elman model

- (3) To improve the prediction performance of CNNLSTM, this paper applies BO to optimize its parameters: learning rate and the number of hidden layer units. The detailed information of BO can be seen in Sect. 2.5.

## 2.2 The Hybrid Mode Decomposition Method

An excellent decomposition method can reduce the forecasting difficulty on the follow-up electricity price forecasting. In this study, the proposed HMD approach can possess higher decomposition ability. In which, the electricity price is first decomposed into several EIMFs by CEEMDAN. Then, the complexities of all EIMFs are calculated by SE, whose reconstruction dimension is 2, and threshold is set as 0.2. Finally, the EIMFs with the high SE values add up and are further decomposed into relatively steady subsequences by EWT.

## 2.3 Convolutional Long Short Term Memory Network

CNNLSTM consists of three convolutional layers, one fully connected layer. The convolutional layers are constructed by three one-dimensional convolution operators, and whose channels are 4, 16 and 32, respectively. The activation function is the Relu in each convolution layer. The dropout is put between convolution

layer 3 and fully connected layer, whose ratio is set to 0.6. Then we use the output of the fully connected layer as the input of LSTM and construct LSTM model. In addition, to enhance the generalization and robustness of CNNLSTM, Adam is selected as the optimization algorithm, and BO is used to optimize the parameters of CNNLSTM: learning rate and the number of hidden layer units.

## 2.4 Elman Neural Network

Elman adds a context layer to the hidden layer of the feedforward network, and whose dynamic memory and time-varying ability can directly reflect the characteristics of the dynamic process system. Therefore, Elman is expected to predict the low-frequency components in electricity price.

The calculation of the Elman neural network can be expressed as:

$$x_t = f(W_1 u_t + W_2 x_{t-1} + b_1) \quad (1)$$

$$y_t = g(W_3 x_t + b_2) \quad (2)$$

where  $u_t$  represents input data,  $x_t$  represents the intermediate layer node unit vector,  $y_t$  denotes the output data,  $W_1$ ,  $W_2$  and  $W_3$  denote the weight matrix,  $b_1$ ,  $b_2$  denote bias vectors,  $f$  and  $g$  denote activation functions.

## 2.5 Bayesian Optimization

The core of BO lies in constructing a probability model for each black box function. In this process, the Gaussian process is used to model the objective function in BO and obtain its posterior distribution. Then, the Expected Improvement as the Acquisition Function is employed to find the next  $x$  function for sample calculation, which can solve the expected value of the unknown point function value and the maximum target value. Therefore, the final result is given by:

$$EI(x) = \begin{cases} (\mu(x) - f(x^+))\Phi(Z) + \sigma(x)\phi(Z) & \sigma(x) > 0 \\ 0 & \sigma(x) = 0 \end{cases} \quad (3)$$

$$Z = \frac{\mu(x) - f(x^+)}{\sigma(x)} \quad (4)$$

where  $x$  is the observation point,  $\mu(\cdot)$  is the mean of all observation points,  $\sigma(\cdot)$  is the standard deviation of all observation points,  $\Phi(\cdot)$  represents the normal cumulative distribution function,  $\phi(\cdot)$  represents the normal probability density function,  $f(x^+)$  represents the existing maximum value.

## 2.6 Forecasting Performance Indices

To evaluate the predictive performance of each model fairly, this study used some predictive performance indices. Three commonly used statistical criteria for analysis results are mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE).

### 3 Case Study

This study contains two cases to explore the forecasting performance of the HMD-CNNLSTM-Elman model. Besides, all algorithms and models are based on Matlab R2014b and Python 3.6 environments, and the forecasting models involved in this study are for 2-step forecasting.

#### 3.1 Data Description

In this study, the original electricity price data were the hourly Day-Ahead prices for the German bidding zone, which gathered from EPEX SPOT energy exchange, where the prices are given in €/MWh. In the experiments, we used the two sets of 1-h electricity price time series including 800 samples, one set covering 34 days from July 20, 2007, to August 22, 2007, another is September 1, 2014, to October 4, 2014. Besides, the first 600 samples were selected as the training dataset, and the rest 200 samples were used as the testing dataset. The main descriptive statistics are shown in Table 1.

**Table 1.** The statistical information of electricity price data from 2007 and 2014.

Time	Dataset	Max	Median	Min	Mean	St.d	Var
2007	Entire dataset	98.43	26.08	4.28	27.73	11.02	121.53
	Training dataset	98.43	26.08	4.28	27.27	10.87	118.06
	Testing dataset	75.04	26.16	9.18	29.11	11.38	129.41
2014	Entire dataset	69.45	34.80	9.85	35.36	10.14	102.90
	Training dataset	69.30	34.64	9.85	34.58	9.36	87.69
	Testing dataset	69.45	35.04	17.41	37.73	11.88	141.07

#### 3.2 Case Study 1: Comparison with Other Hybrid Models

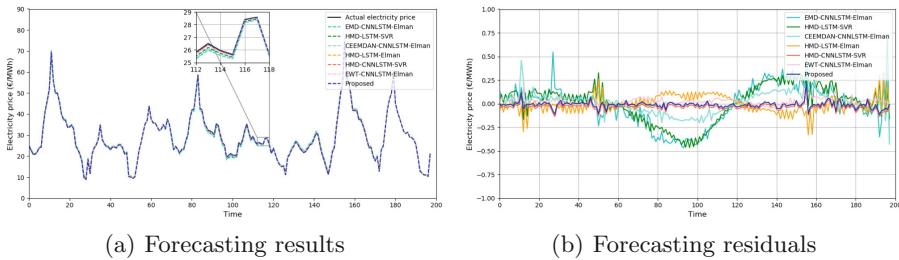
In this section, the 2007 electricity price data are applied for the forecast object. The SE values of all EIMFs obtained by CEEMDAN decomposition dataset are shown in Table 2, and Table 2 shows that the EIMF1-EIMF4 need to be further decomposed by EWT with large SE values. Table 3 shows the forecasting performance indices of different models. The electricity price forecasting results and forecasting residuals of different models are shown in Fig. 2.

**Table 2.** The SE values of EIMFs decomposed by CEEMDAN in the dataset in case study 1.

Mode	EIMF1	EIMF2	EIMF3	EIMF4	EIMF5	EIMF6
SE value	1.4013	0.7796	0.5931	0.2932	0.1392	0.0125

**Table 3.** Forecasting performance indices for different models in case study 1.

Model	MAE (€/MWh)	RMSE (€/MWh)	MAPE (%)
EMD-CNNLSTM-Elman	0.1676	0.2149	0.7051
HMD-LSTM-SVR	0.1585	0.1957	0.6377
CEEMDAN-CNNLSTM-Elman	0.0914	0.1184	0.3707
HMD-LSTM-Elman	0.0689	0.0929	0.2928
HMD-CNNLSTM-SVR	0.0391	0.0479	0.1687
EWT-CNNLSTM-Elman	0.0331	0.0425	0.1263
HMD-CNNLSTM-Elman	0.0242	0.0361	0.1072

**Fig. 2.** Electricity price forecasting results and forecasting residuals of different models in case study 1

Based on Table 3, Fig. 2, the following conclusions can be obtained:

- (1) Under different decomposition methods, compared with the other three models, the MAE, RMSE and MAPE of the HMD-CNNLSTM-Elman model are reduced by an average of 61.99%, 55.92% and 57.00%, respectively. It confirms that a reasonable decomposition method can reduce the nonlinear and non-stationary characteristics of electricity price, thus reducing the forecasting difficulty. The experimental results also display that HMD is an excellent secondary decomposition algorithm with the best forecasting performance.
- (2) The EMD algorithm has the worst forecasting performance due to the problem of pattern blending. As an improved algorithm of EMD, CEEMDAN and EWT have relatively better forecasting performance.
- (3) Under the same decomposition method, the MAE, RMSE and MAPE of the proposed model are reduced by an average of 62.67%, 55.77% and 61.01%, respectively. We can conclude that under the same decomposition method, the proposed hybrid model possesses both stronger robust stability and higher accuracy and it can give full play to the advantages of two single models and achieve higher forecasting precision.
- (4) In addition, the experiment shows that CNNLSTM has better forecasting performance than LSTM with more complex subsequences; when with low complexities, Elman performs better than SVR.

### 3.3 Case Study 2: Comparison with Other Single Models

In this section, the proposed model is compared with five single models, including the K-Nearest Neighbors (KNN), Elman, SVR, Gradient Boosting Decision Tree (GBDT) and CNNLSTM for predicting the 2014 electricity price data. Table 4 shows the sample entropy values of all EIMFs obtained by CEEMDAN decomposition dataset, which is seen that the EIMF1–EIMF4 need to be further decomposed by EWT. Table 5 shows the forecasting performance indices of different models. The electricity price forecasting results and forecasting residuals of different models are shown in Fig. 3. Besides, the parameters of every single model can be seted for following. KNN: the number of neighbors was set as 10, the leaf size was set as 28; Elman: The number of neurons and learning rate were set as 15 and 0.02; GBDT: The n-estimators, learning rate, max-depth and min-samples-split were set as 800, 0.006, 10 and 80, respectively; SVR:  $C$  and  $\sigma$ (BO); LSTM and CNNLSTM: learning rate and the number of hidden layer units(BO).

**Table 4.** The SE values of EIMFs decomposed by CEEMDAN in the dataset in case study 2.

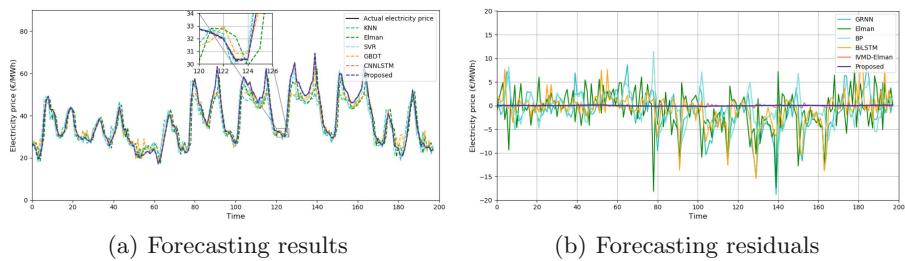
Mode	EIMF1	EIMF2	EIMF3	EIMF4	EIMF5	EIMF6	EIMF7
SE value	1.3470	0.6738	0.5535	0.2995	0.1276	0.0316	0.0030

**Table 5.** Forecasting performance indices for different models in case study 2.

Model	MAE (€/MWh)	RMSE (€/MWh)	MAPE (%)
KNN	3.3022	4.4919	8.5206
Elman	3.0809	4.2613	8.4383
SVR	2.4688	3.1983	6.6506
GBDT	2.3962	3.6486	6.1953
CNNLSTM	0.1175	0.1615	0.3167
HMD-CNNLSTM-Elman	0.0618	0.0765	0.1782

The forecasting results of case study 2 are shown in Table 5, Fig. 3, as described below:

- (1) The proposed HMD-CNNLSTM-Elman model performance is more accurate and stable than the individual models which have been widely used, the minimum MAE is 0.0618, the RMSE is 0.0765, the MAPE is 0.1782%, and the electricity price forecasting residual is in  $[-1, 1]$ .
- (2) In all the comparison models, compared with the worst forecasting performance KNN, the proposed HMD-CNNLSTM-Elman model has a decrease of 98.13% in MAE, 98.30% in RMSE, and 97.91% in MAPE. Therefore, the appropriate hybrid model can combine the advantages of every single model, thereby achieving an outstanding forecasting model.



**Fig. 3.** Electricity price forecasting results and forecasting residuals of different models in case study 2

- (3) Besides, we also need to consider the training speed of the models. There is no doubt that the hybrid model has better forecasting performance, but the training time is relatively longer. Therefore, a single model can be considered when encountering a job with a low forecasting performance requirement but requires a fast Training speed.

## 4 Conclusion and Future Work

In this paper, a short-term hybrid model based on HMD, CNNLSTM and Elman has been proposed to forecast electricity price. In our model, HMD has been used to deeply decompose the original electricity price data into a series of subsequences. CNNLSTM and Elman have been used to forecast the electricity price subsequence. Besides, BO has been introduced to optimize parameters for improving the forecasting performance of CNNLSTM. Finally, two different case studies have been taken to verify the forecasting performance of the proposed model. This paper studies the electricity price forecast based on univariate time series forecasting. In future work, we will consider the electricity price forecast with multiple influencing factors to further improve the forecasting performance. In addition, we will consider how to simplify the model to achieve faster forecasting speed while ensuring the accuracy of the model forecasting.

**Acknowledgements.** This work is partially supported by the National Natural Science Foundation of China (No. 61573095). This work is supported by the Natural Science Foundation of Shanghai under grant no. 20ZR1402800.

## References

1. Jia, Y.: Robust control with decoupling performance for steering and traction of 4WS vehicles under velocity-varying motion. *IEEE Trans. Control Syst. Technol.* **8**(3), 554–569 (2000). <https://doi.org/10.1109/87.845885>
  2. Jia, Y.: Alternative proofs for improved LMI representations for the analysis and the design of continuous-time systems with polytopic uncertainty: a predictive approach. *IEEE Trans. Autom. Control* **48**(8), 1413–1416 (2003). <https://doi.org/10.1109/TAC.2003.815033>

3. Weron, R.: Electricity price forecasting: a review of the state-of-the-art with a look into the future. *Int. J. Forecast.* **30**(4), 1030–1081 (2014). <https://doi.org/10.1016/j.ijforecast.2014.08.008>
4. Lago, J., Ridder, F., Vrancx, P., Schutter, B.: Forecasting day-ahead electricity price in Europe: the importance of considering market integration. *Appl. Energy* **211**, 890–903 (2018)
5. Girish, G.P.: Spot electricity price forecasting in Indian electricity market using autoregressive-GARCH models. *Energy Strategy Rev.* **11–2**, 52–7 (2016). <https://doi.org/10.1016/j.apenergy.2017.11.098>
6. Grossi, L., Nan, F.: Robust forecasting of electricity price: simulations, models and the impact of renewable sources. *Technol. Forecast. Soc. Chang.* **141**, 305–318 (2019). <https://doi.org/10.1016/j.techfore.2019.01.006>
7. Liu, H., Tian, H., Liang, X., Li, Y.: Wind speed forecasting approach using secondary decomposition algorithm and Elman neural networks. *Appl. Energy* **157**, 183–94 (2015). <https://doi.org/10.1016/j.apenergy.2015.08.014>
8. Chen, J., Zeng, G., Zhou, W., Du, W., Lu, K.: Wind speed forecasting using nonlinear-learning ensemble of deep learning time series forecasting and extremal optimization. *Energy Convers. Manag.* **165**, 681–695 (2018). <https://doi.org/10.1016/j.enconman.2018.03.098>
9. Chen, G., Yi, X., Zhang, Z., Wang, H.: Applications of multi-objective dimension-based firefly algorithm to optimize the power losses, emission, and cost in power systems. *Appl. Soft Comput.* **68**, 322–342 (2018). <https://doi.org/10.1016/j.asoc.2018.04.006>
10. Liu, H., Mi, X., Li, Y.: Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network. *Energy Convers. Manag.* **156**, 498–514 (2018). <https://doi.org/10.1016/j.enconman.2017.11.053>
11. Zhang, W., Qu, Z., Zhang, K., Mao, W., Ma, Y., Fan, X.: A combined model based on CEEMDAN and modified flower pollination algorithm for wind speed forecasting. *Energy Convers. Manag.* **136**, 439–451 (2017). <https://doi.org/10.1016/j.enconman.2017.01.022>
12. Cheng, H., Ding, X., Zhou, W., Ding, R.: A hybrid electricity price forecasting model with Bayesian optimization for German energy exchange. *Electr. Power Energy Syst.* **110**, 653–666 (2019). <https://doi.org/10.1016/j.ijepes.2019.03.056>



# Disturbance Observer-Based Finite-Time Control for Systems with Nonlinearity and Disturbance

Xinqing Li<sup>1</sup> and Xinjiang Wei<sup>2(✉)</sup>

<sup>1</sup> School of Mathematics and Statistics Science, Ludong University,  
Yantai 264000, China

Lixinqing0705@163.com

<sup>2</sup> School of Mathematics and Statistics Science, Ludong University,  
Yantai 264000, China  
weixinjiang@163.com

**Abstract.** In order to achieve the goal of high control accuracy and fast convergence, a disturbance observer-based finite-time control (DOBFTC) strategy is presented for system with nonlinearity and disturbance with partially known information in this note. The disturbance with partially known information is estimated and rejected with the help of disturbance observer (DO). Combining with the DO and the finite-time theory, a DOBFTC strategy is put forward, which ensures the globally finite-time stable of the composite system. It not only speeds up the convergence rate but also realizes the high control precision of the system. Finally, a simulation example is provided to test the capability of the developed DOBFTC approach.

**Keywords:** Nonlinearity and disturbance · Implicit Lyapunov function · Disturbance observer-based finite-time control · Globally finite-time stable

## 1 Introduction

Disturbances are ubiquitous in complex environments, which seriously affect the performance of the system. For example, the actual operation of the wind turbine system is a highly coupled nonlinear system [1], the impact of wind, rain, lightning and other environment results in system performance degradation. Therefore, the research on anti-disturbance control plays an vital role in the control field.

Disturbance observer-based control (DOBC) is a major disturbance suppression method, which was proposed at the end of 1980s. The disturbance observer was designed based on the known information of the disturbance to realize the estimation and compensation of the disturbance. Since then, linear DOBC and nonlinear DOBC have been developed. Compared with other control methodologies, DOBC has some significant advantages, such as its simple structure,

easy combination with other control strategies and so on. Recently, a composite hierarchical anti-disturbance control (CHADC) structure combining DOBC and other control strategies has been proposed for multiple disturbances systems [2–6]. CHADC method can make full use of disturbance information and analyze the features of multi-source disturbances, then reject and compensate for multi-source disturbances, which has the advantages of high precision. [2] introduced CHADC for the first time and compared it with other control methodologies, which proved that CHADC had better advantages in rejecting multiple disturbances. For uncertain MIMO systems, [3] proposed an adaptive neural network control strategy, whose control goal was to track the expected trajectory in the presence of both input saturation, external disturbances and system uncertainty. The antidisturbance control problem of stochastic discrete-time systems subject to multi-source disturbances and nonlinearity was investigated in [4]. For the known and unknown nonlinear cases, combined with DOBC and  $H_\infty$  method, the elegant anti-disturbance control scheme was developed to suppress and attenuate the disturbance. [5] proposed a new composite anti-disturbance control strategy for stochastic systems with input saturation and multi-source heterogeneous disturbances. This scheme was a generalization of CHADC structure under input saturation. An adaptive disturbance observer based control was constructed by [6], which could be suppressed and rejected the disturbance of unknown amplitude and unknown frequency.

Although the above literature results can achieve high control performance of system, they are all based on infinite time to achieve. In the actual industrial production, in order to obtain more benefits or achieve the expected system performance within the target time, higher requirements are often put forward for the system convergence time, such as the emergency braking of vehicles, the formation system of aircraft to eliminate enemy aircraft and so on. With the advancement and improvement of advanced control approaches, the finite-time control scheme guided by modern control theory can realize the system reaches to the region of convergence at a faster speed, thereby improving the robustness and anti-disturbance performance of the system, which has attracted widespread attention of many researchers [7–9]. In the light of the design method of dynamic gain control and the finite-time Lyapunov stability theorem, [10] put forward a state feedback finite time stable controller with on-line gain adjustment, which had lower control gain compared with the backstepping method. In [11], a nonlinear control law was constructed by applying the homogeneous theorem and implicit Lyapunov function approach, and the integrator chain finite-time stability control design problem was solved. The finite-time tracking problem for stochastic switched uncertain nonlinear systems was considered by [12]. Combining the adaptive fuzzy control technology with the Lyapunov function approach, a novel general control scheme was proposed, which made the tracking error achieve the expected goal in a finite-time. And in the last few decades, the finite-time control theory has been successfully applied to permanent magnet synchronous motor [13], biological system [14], wheeled mobile robots [15], multi-agents [16] and other fields.

On this basis, this article intends to develop a disturbance observer-based finite-time control (DOBFTC) scheme for systems subject to nonlinearity and disturbance with partially known information. There are the following highlights:

- 1) The research of disturbance observer-based control (DOBC) is extended to the finite-time control, which improves the convergence speed and guarantees that the system state reaches the equilibrium point within finite time.
- 2) The DOBFTC structure is constructed to realize the rejection and compensation of disturbances, which renders the system to achieve global finite-time stability.

The layout of the article is arranged as follows. Preliminary results are shown in Sect. 2. The main results are demonstrated by Sect. 3. The simulation tests the expected efficiency of the system as displayed in Sect 4. In Sect. 5, the conclusion remark is given.

## 2 Preliminaries

The system with disturbance and nonlinearity is considered:

$$\dot{x}(t) = A_0x(t) + B_0[u(t) + D(t)] + Hh(x, t), \quad (1)$$

where  $x(t) \in R^n$  and  $u(t) \in R^m$  are the state variable and the control input, respectively.  $A_0 \in R^{n \times n}$ ,  $B_0 \in R^{n \times m}$  and  $H \in R^{n \times p}$  are the system matrices.  $h(x, t)$  is a known bound nonlinear function.  $D(t) \in R^m$  represents exogenous disturbance, depicted by

$$\begin{aligned} D(t) &= M\eta(t) \\ \dot{\eta}(t) &= N\eta(t) \end{aligned} \quad (2)$$

where  $\eta(t) \in R^t$  is the exogenous disturbance state, the matrices  $M$  and  $N$  are known.

**Assumption 1.** *The pairs  $(A_0, B_0)$  and  $(N, M)$  are controllable and observable, respectively.*

## 3 Main Results

The partially information known disturbance  $D(t)$  is online estimated by a DO, and a DOBFTC scheme is presented.

### 3.1 Disturbance Observer (DO)

A DO is designed as

$$\begin{cases} \dot{\theta}(t) = (N - VB_0M)\hat{\eta}(t) - V[A_0x(t) + B_0u(t) + Hh(x, t)], \\ \hat{\eta}(t) = \theta(t) + Vx(t), \\ \hat{D}(t) = M\hat{\eta}(t), \end{cases} \quad (3)$$

where  $\theta(t)$  as the observer state and  $\hat{\eta}(t)$  is an estimation of  $\eta(t)$ . The observation gain is  $V$ . Define  $e_\eta(t) = \eta(t) - \hat{\eta}(t)$ , according to (1), (2) and (3), we have

$$\dot{e}_\eta(t) = (N - VB_0M)e_\eta(t). \quad (4)$$

Since  $(N, M)$  is controllable, the expected performance of the DO can be acquired by adjusting  $V$  through the pole assignment.

In the next step, the DOBFT controller is structured as follows

$$u(t) = -\hat{D}(t) + \Lambda^{1-\alpha}FD_{r_1}(\Lambda^{-1})x(t), \quad (5)$$

where  $D_r(\lambda) = \text{diag}\{D_{r_1}(\lambda), D_{r_2}(\lambda)\} = \text{diag}\{\lambda^{1+\alpha}I_n, \lambda I_t\}$  and  $\Lambda$  is a solution derived from the implicit Lyapunov function;  $\lambda \in R_+$  and  $0 < \alpha \leq 1$ .

Bringing (5) into (1), one has

$$\dot{x}(t) = (A_0 + \Lambda^{1-\alpha}B_0FD_{r_1}(\Lambda^{-1}))x(t) + B_0Me_\eta(t) + Hh(x, t). \quad (6)$$

Composite (4) and (6), the composite system can be gotten

$$\dot{s}(t) = (\bar{A}_1 + \bar{A}_2)s(t) + \bar{H}\bar{h}(x, t), \quad (7)$$

where

$$s(t) = \begin{bmatrix} x(t) \\ e_\eta(t) \end{bmatrix}, \quad \bar{A}_1 = \begin{bmatrix} A_0 + \Lambda^{1-\alpha}B_0FD_{r_1}(\Lambda^{-1}) & 0 \\ 0 & N - VB_0M \end{bmatrix},$$

$$\bar{A}_2 = \begin{bmatrix} 0 & B_0M \\ 0 & 0 \end{bmatrix}, \quad \bar{H} = \begin{bmatrix} H \\ 0 \end{bmatrix}, \quad \bar{h}(x, t) = h(x, t). \quad (8)$$

### 3.2 Disturbance Observer-Based Finite-Time Control (DOBFTC)

The DOBFTC scheme is constructed to guarantee global finite-time stability of system state in this part.

**Theorem 1.** *For system (1) with the partially information known disturbance (2), suppose there are  $P > 0, Q > 0$ , constants  $0 < \alpha < 1, 0 < \varsigma < 1$  and  $\gamma > 0$ , satisfy*

$$\Psi = \begin{bmatrix} \Upsilon_1 & \Lambda^{-\alpha}B_0M & H \\ * & \Upsilon_2 & 0 \\ * & * & -\Lambda^\alpha I \end{bmatrix} < 0, \quad (9)$$

where

$$\begin{aligned}\Upsilon_1 &= A_0 X + \Lambda^{1-\alpha} B_0 R_1 + X A_0^T + \Lambda^{1-\alpha} R_1^T B_0^T + \Lambda^{-\alpha} X, \\ \Upsilon_2 &= Q(N - V B_0 M) + (N - V B_0 M)^T Q + \Lambda^{-\alpha} Q,\end{aligned}$$

then the composite system (7) under the gain of observer  $V$  and controller (5) with  $F = R_1 X^{-1} D_{r_1}^{-1}(\Lambda^{-1})$  realizes global finite-time stability with  $T(x_0) < \frac{\gamma A_0^\alpha}{(1-\varsigma)\alpha}$ .

*Proof.* Introducing implicit Lyapunov function candidate

$$G(\Lambda, s) = s^T(t) D_r(\Lambda^{-1}) \Phi D_r(\Lambda^{-1}) s - 1. \quad (10)$$

Setting

$$\Phi = \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix} = \begin{bmatrix} X^{-1} & 0 \\ 0 & Q \end{bmatrix} > 0. \quad (11)$$

The following inequalities

$$\frac{\lambda_{\min}(\Phi) \|s\|^2}{\max\{\Lambda^{2+2(n-1)\alpha}, \Lambda^2\}} \leq G(\Lambda, s) + 1 \leq \frac{\lambda_{\max}(\Phi) \|s\|^2}{\min\{\Lambda^{2+2(n-1)\alpha}, \Lambda^2\}}. \quad (12)$$

is true for all  $s \in R^n \setminus \{0\}$  and  $\Lambda \in R_+$ , then the formulation (10) is radially unbounded and positive definite. Besides

$$-\gamma \Lambda^{-1} < \frac{\partial G}{\partial \Lambda} = \Lambda^{-1} s^T(t) D_r(\Lambda^{-1}) (\Gamma_\alpha \Phi + \Phi \Gamma_\alpha) D_r(\Lambda^{-1}) s(t) < 0, \quad (13)$$

where  $\gamma > 0$ ,  $\Gamma_\alpha = -\text{diag}\{(1+\alpha)I_n, I_t\}$ . Then, we can obtain

$$\begin{aligned}\frac{\partial G}{\partial s} &((\bar{A}_1 + \bar{A}_2)s(t)) + \bar{H}\bar{h}(x, t) \\ &= 2s^T(t) D_r(\Lambda^{-1}) \Phi D_r(\Lambda^{-1}) ((\bar{A}_1 + \bar{A}_2)s(t) + \bar{H}\bar{h}(x, t)).\end{aligned} \quad (14)$$

Considering the  $D_r(\Lambda^{-1})\bar{A}_2 = \Lambda^{-\alpha}\bar{A}_2 D_r(\Lambda^{-1})$ , yields

$$\begin{aligned}\frac{\partial G}{\partial s} &((\bar{A}_1 + \bar{A}_2)s(t)) + \bar{H}\bar{h}(x, t) \\ &= s^T(t) D_r(\Lambda^{-1}) (\bar{A}_1^T \Phi + \Phi \bar{A}_1 + \Lambda^{-\alpha} (\bar{A}_2^T \Phi + \Phi \bar{A}_2)) D_r(\Lambda^{-1}) s(t) \\ &\quad + 2s^T(t) D_r(\Lambda^{-1}) \Phi \bar{H} D_r(\Lambda^{-1}) \bar{h}(x, t) \\ &\leq \Theta^T(t) \Psi_1 \Theta(t) - \Lambda^{-\alpha} s^T(t) D_r(\Lambda^{-1}) \Phi D_r(\Lambda^{-1}) s(t) \\ &\quad + \Lambda^\alpha \bar{h}^T(x, t) D_r(\Lambda^{-1}) D_r(\Lambda^{-1}) \bar{h}(x, t)\end{aligned} \quad (15)$$

where

$$\Theta(t) = \begin{bmatrix} D_r(\Lambda^{-1}) s(t) \\ D_r(\Lambda^{-1}) \bar{h}(x, t) \end{bmatrix}, \Psi_1 = \begin{bmatrix} \bar{A}_1^T \Phi + \Phi \bar{A}_1 + \Lambda^{-\alpha} (\bar{A}_2^T \Phi + \Phi \bar{A}_2 + \Phi) & \Phi \bar{H} \\ * & -\Lambda^\alpha I \end{bmatrix}$$

As  $\bar{h}^T(x, t)D_r(\Lambda^{-1})D_r(\Lambda^{-1})\bar{h}(x, t) < \varsigma\Lambda^{-2\alpha}$ , if  $\Psi_1 < 0$ , then

$$\frac{\partial G}{\partial s}((\bar{A}_1 + \bar{A}_2)s(t) + \bar{H}\bar{h}(x, t)) \leq \frac{1-\varsigma}{\gamma}\Lambda^{1-\alpha}\frac{\partial G}{\partial \Lambda}, \quad (16)$$

for  $\Lambda(s, t) : s^T(t)D_r(\Lambda^{-1})\bar{\Phi}D_r(\Lambda^{-1})s(t) = 1$ . Based on the proof of [8, Theorem15], the composite system (7) achieves global finite-time stability with  $T(x_0) < \frac{\gamma A_0^\alpha}{(1-\varsigma)\alpha}$ .

Next, we will prove that  $\Psi < 0 \Leftrightarrow \Psi_1 < 0$ .

(1)  $\Psi_1 < 0 \Leftrightarrow \Psi_2 < 0$ . According to (8), (10) and Schur complement, yields

$$\Psi_2 = \begin{bmatrix} \Sigma_1 & \Lambda^{-\alpha}PB_0M & PH \\ * & \Sigma_2 & 0 \\ * & * & -\Lambda^\alpha I \end{bmatrix} < 0, \quad (17)$$

with

$$\begin{aligned} \Sigma_1 &= (A_0 + \Lambda^{1-\alpha}B_0FD_{r_1}(\Lambda^{-1}))^T P + P(A_0 + \Lambda^{1-\alpha}B_0FD_{r_1}(\Lambda^{-1})) + \Lambda^{-\alpha}P, \\ \Sigma_2 &= Q(N - VB_0M) + (N - VB_0M)^TQ + \Lambda^{-\alpha}Q. \end{aligned}$$

(2)  $\Psi_2 < 0 \Leftrightarrow \Psi_3 < 0$ . Multiply  $\Psi_2$  by diag  $\{X, I, I\}$  on both sides, results in

$$\Psi_3 = \begin{bmatrix} \Omega_1 & \Lambda^{-\alpha}B_0M & H \\ * & \Omega_2 & 0 \\ * & * & -\Lambda^\alpha I \end{bmatrix} < 0, \quad (18)$$

with

$$\begin{aligned} \Omega_1 &= X(A_0 + \Lambda^{1-\alpha}B_0FD_{r_1}(\Lambda^{-1}))^T + (A_0 + \Lambda^{1-\alpha}B_0FD_{r_1}(\Lambda^{-1}))X + \Lambda^{-\alpha}X, \\ \Omega_2 &= Q(N - VB_0M) + (N - VB_0M)^TQ + \Lambda^{-\alpha}Q. \end{aligned}$$

(3)  $\Psi_3 < 0 \Leftrightarrow \Psi < 0$ . Setting  $F = R_1X^{-1}D_{r_1}^{-1}(\Lambda^{-1})$  in (9) can prove  $\Psi_3 < 0 \Leftrightarrow \Psi < 0$ .

Based on the above process, the composite system (7) achieves global finite-time stability with  $T(x_0) < \frac{\gamma A_0^\alpha}{(1-\varsigma)\alpha}$ .

## 4 Simulation Example

The system matrices are shown below to test the capability of the presented method

$$A_0 = \begin{bmatrix} -0.2 & -5.97 \\ 1.98 & -2.72 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0.01 \\ 0.01 \end{bmatrix}, \quad H = \begin{bmatrix} 0.76 \\ -1.05 \end{bmatrix}.$$

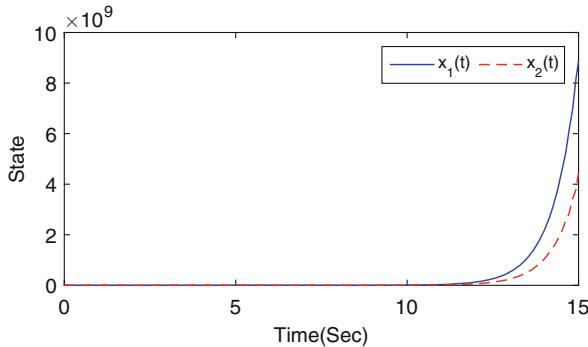
The parameter matrices of disturbance  $D(t)$  are given by

$$N = \begin{bmatrix} 0 & 1.2 \\ -1.2 & 0 \end{bmatrix}, \quad M = \begin{bmatrix} 1.2 & -0.29 \end{bmatrix}.$$

Suppose the nonlinear function is represented as  $h(x, t) = \cos(2\pi * 6t)x_1(t)$ , the  $x(0)$  is selected as  $[1, -2]^T$ , the poles are placed in the  $[-4, -6]$ , we have

$$V = \begin{bmatrix} 572.5346 & 572.5346 \\ 644.9708 & 644.9708 \end{bmatrix}.$$

On the basis of Theorem 1, it could be obtained

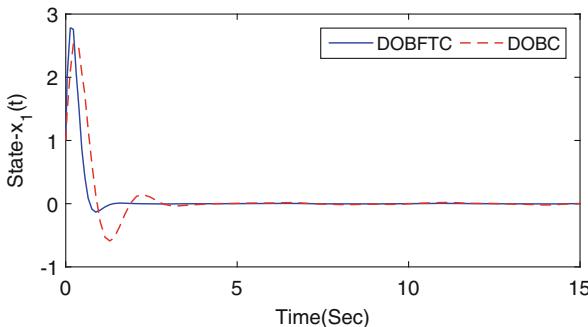


**Fig. 1.** Trajectory of the system states without control.

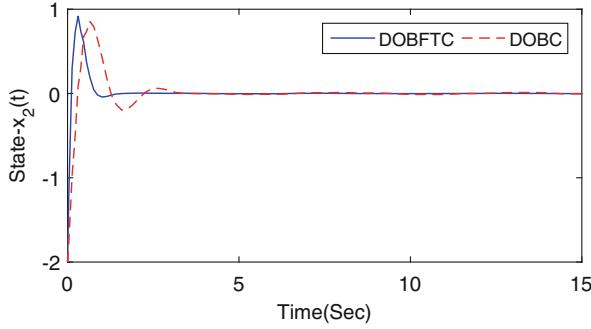
$$X = \begin{bmatrix} 1.5889 & 0.4559 \\ 0.4559 & 0.3163 \end{bmatrix}, \quad Q = \begin{bmatrix} 0.2245 & 0.4765 \\ 0.4765 & 1.5347 \end{bmatrix},$$

$$F = [24.8127 \quad -85.9318], \quad R_1 = [1.1991 \quad -75.0483].$$

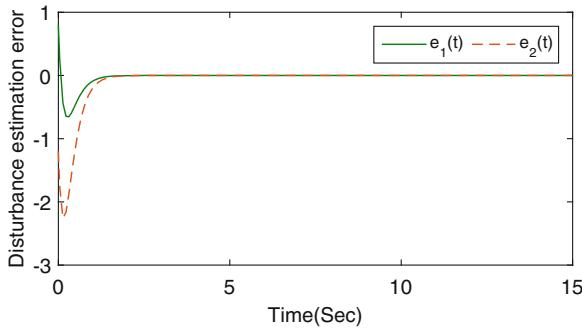
The responses curve of composite system are demonstrated by the above Figures. Fig. 1 shows that system states are divergent in the presence of uncontrolled conditions. According to the proposed DOBFTC strategy, the system



**Fig. 2.** Response of state  $x_1$  with DOBFTC and DOBC.



**Fig. 3.** Response of state  $x_2$  with DOBFTC and DOBC.



**Fig. 4.** Curve of disturbance estimation error.

state can achieve global finite-time stability, and has faster convergence speed than DOBC, which can be shown in Fig. 2 and Fig. 3. Fig. 4 is the disturbance estimation error curve, which reveals that the design of DO is successful. The simulation figures show that the system can achieve ideal performance in the presence of both disturbance with partially known information and nonlinearity, which prove the feasibility of the DOBFTC scheme.

## 5 Conclusion

In view of a class systems with nonlinearity and disturbance with partially known information, a DOBFTC strategy is developed, which renders the composite system to achieve global finite-time stability. It not only assures the high control precision of the system, but also speeds up the convergence rate. Most of the existing finite-time control methods can merely handle the system with single type of equivalent disturbance, which was integrated by various uncertainties.

However, the impact of multi-source disturbances on the system and the analysis of their properties have not been paid enough attention by scholars. Hence, the study of finite-time control for systems with different characteristics and sources of disturbance is one of the new challenges.

## References

1. Corradini, M.L., Ippoliti, G., Orlando, G.: Robust control of variable-speed wind turbines based on an aerodynamic torque observer. *IEEE Trans. Control Syst. Technol.* **21**(4), 1199–1206 (2013). <https://doi.org/10.1109/TCST.2013.2257777>
2. Guo, L., Cao, S.Y.: Anti-disturbance control theory for systems with multiple disturbance: a survey. *ISA Trans.* **53**, 846–849 (2014). <https://doi.org/10.1016/j.isatra.2013.10.005>
3. Chen, M., Shao, S.Y., Jiang, B.: Adaptive neural control of uncertain nonlinear systems using disturbance observer. *IEEE Trans. Cybern.* **47**(10), 3110–3123 (2017). <https://doi.org/10.1109/TCYB.2017.2667680>
4. Wei, X.J., Sun, S.X.: Elegant anti-disturbance control for discrete-time stochastic systems with nonlinearity and multiple disturbances. *Int. J. Control* **91**(3), 706–714 (2018). <https://doi.org/10.1080/00207179.2017.1291996>
5. Wei, X.J., Dong, L.W., Zhang, H.F., Han, J., Hu, X.: Composite anti-disturbance control for stochastic systems with multiple heterogeneous disturbances and input saturation. *ISA Trans.* **100**, 436–445 (2019). <https://doi.org/10.1016/j.isatra.2019.12.006>
6. Wei, X.J., Dong, L.W., Zhang, H.F., Hu, X., Han, J.: Adaptive disturbance observer-based control for stochastic systems with multiple heterogeneous disturbances. *Int. J. Robust Nonlinear Control* **29**, 5533–5549 (2019). <https://doi.org/10.1002/rnc.4683>
7. Ding, S.H., Li, S.H.: Stabilization of the attitude of a rigid spacecraft with external disturbances using finite-time control techniques. *Aerospace Sci. Technol.* **13**(4–5), 256–265 (2009). <https://doi.org/10.1016/j.ast.2009.05.001>
8. Polyakov, A., Efimov, D., Perruquetti, W.: Robust stabilization of MIMO systems in finite/fixed time. *Int. J. Robust Nonlinear Control* **26**, 69–90 (2016). <https://doi.org/10.1002/rnc.3297>
9. Chen, X., Zhang, X.F.: Output-feedback control strategies of lower-triangular nonlinear nonholonomic systems in any prescribed finite time. *Int. J. Robust Nonlinear Control* **29**(4), 904–918 (2018). <https://doi.org/10.1002/rnc.4413>
10. Zhang, X.F., Feng, G., Sun, Y.H.: Finite-time stabilization by state feedback control for a class of time-varying nonlinear systems. *Automatica* **48**(3), 499–504 (2012). <https://doi.org/10.1016/j.automatica.2011.07.014>
11. Zimenko, K., Polyakov, A., Efimov, D.: On finite-time robust stabilization via nonlinear state feedback. *Int. J. Robust Nonlinear Control* **28**, 4951–4965 (2018). <https://doi.org/10.1002/rnc.4292>
12. Wang, F., Chen, B., Sun, Y., Lin, C.: Finite time control of switched stochastic nonlinear systems. *Fuzzy Sets Syst.* **365**, 140–152 (2018). <https://doi.org/10.1016/j.fss.2018.04.016>
13. Li, S.H., Liu, H.X., Ding, S.H.: A speed control for a PMSM using finite-time feedback control and disturbance compensation. *Trans. Inst. Meas. Control* **32**(2), 170–187 (2009). <https://doi.org/10.1177/0142331209339860>

14. Xing, S.H., Zhang, Q.L., Zhang, Y.: Finite-time stability analysis and control for a class of stochastic singular biological economic systems based on T-S fuzzy model. *Abstract Appl. Anal.* **2013**, 233–255 (2013). <https://doi.org/10.1155/2013/946491>
15. He, X.D., Geng, Z.Y.: Arbitrary point-to-point stabilization control in specified finite time for wheeled mobile robots based on dynamic model. *Nonlinear Dyn.* **97**(2), 937–954 (2019). <https://doi.org/10.1007/s11071-019-05019-0>
16. Sharghi, A., Baradarannia, M., Hashemzadeh, F.: Finite-time-estimation-based surrounding control for a class of unknown nonlinear multi-agent systems. *Nonlinear Dyn.* **96**(3), 1–10 (2019). <https://doi.org/10.1007/s11071-019-04884-z>



# Disturbance Observer-Based Disturbance Attenuation Control for Dynamic Positioning System of Ships

Lihong You and Xinjiang Wei<sup>(✉)</sup>

School of Mathematics and Statistics Science, Ludong University,  
Yantai 264000, China  
[youlihong1006@126.com](mailto:youlihong1006@126.com), [weixinjiang@163.com](mailto:weixinjiang@163.com)

**Abstract.** The anti-disturbance control problem is addressed for dynamic positioning (DP) system of ships affected by the ocean disturbances. The ocean disturbances consist of slowly varying environmental disturbances and a class of long peak wave disturbances in mature period. The stochastic disturbance observers are constructed to estimate them online, respectively. On this basis, a novel disturbance observer-based disturbance attenuation control (DOBDAC) strategy is proposed to ensure that all signals of DP composite system are asymptotically bounded in mean square. Finally, the feasibility of the control strategy is verified by a simulation research on the supply ship.

**Keywords:** Dynamic positioning system · Stochastic disturbance observers · Slowly varying environmental disturbances · Long peak wave disturbances in mature period · Disturbance observer-based disturbance attenuation control

## 1 Introduction

Dynamic positioning (DP) means that ships rely on the power generated by their own thrust system to against the disturbance of the marine environment caused by wind, waves, currents, which makes ships locate at a certain target position on the sea surface or sail along a preset trajectory [1, 2].

The disturbance of wind, wave, current and other marine environment will have a negative impact on the operating ship, which will make the measurement of the ship's position and handing inaccurate [3, 4]. In fact, the wave on the sea is extremely complicated, it is an irregular random wave. Irregular wave drift forces can be seen as the superposition of regular wave drift forces of various frequencies [4]. They are wave forces in the mature period, and the wave crest and trough lines are parallel to each other and perpendicular to the direction of advance. However, the slowly varying environmental disturbances, including wind, current, second-order wave drift force, will significantly affect the position and handing of the ship, and the ship will slowly drift the original position,

resulting in long-period and large amplitude periodic motion [5]. Therefore, the anti-disturbance control of the ship's DP system is particularly important. The control solution of DP system has experienced advanced control methods such as PID regulator, linear model-based controller, nonlinear model-based controller and sliding mode controller [6–8]. Because the environmental disturbances in the ocean are unpredictable, the above control method is not easy to achieve high control accuracy and strong DP system stability when dealing with the control problem of DP ship under real sea conditions. As a result, control technology with strong control performance and better anti-disturbance performance are explored by people.

As a robust control method, a good deal of anti-disturbance problems can be solved and the system stability can be effectively analyzed by disturbance observer based control (DOBC) [9, 10]. Moreover, DOBC can be combined with different control laws according to different requirements of control performance, which is easy to adjust online, so it has the characteristics of high precision and simple structure [11, 12]. Using the DOBC method to deal with dynamic positioning ships can maintain an ideal posture when running on the sea surface, which realizes the stability of the ship and simplifies the complexity of the model. A DOBDAC strategy is proposed for DP system of ships under the multiple heterogeneous disturbances. As a result, the main innovations are as follows

- (1) With the utilization of DOBC method and stochastic theory, a DOBDAC strategy for DP system of ships was proposed to address slowly varying environmental disturbances and long peak wave disturbances in mature period.
- (2) Compared with the PID control methods of DP system, ship's position and heading can achieve the desired target values while guaranteeing that all signals in DP composite control system are asymptotically bounded in mean square.

## 2 Mathematical Modeling of Ships

### 2.1 Kinematics Model of Dynamic Positioning Ship

According to the position information  $(x, y)$  and yaw angle information  $\psi$  of the DP ship are considered as  $\eta = [x, y, \psi]^T$  in the earth-fixed frame, and the surge velocity  $u$ , sway velocity  $v$  and yaw angular velocity  $r$  are expressed as  $v = [u, v, r]^T$  in body-fixed frame. The kinematics model of DP ship is established as follows:

$$\begin{cases} \dot{\eta} = R(\psi)v, \\ R(\psi) = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}. \end{cases} \quad (1)$$

where  $R(\psi)$  represents the rotation matrix transformed from the body-fixed frame to the earth-fixed frame.

## 2.2 Dynamics Model of Dynamic Positioning Ship

Dynamics model of DP ship is generated as

$$M\dot{v}(t) = -Dv(t) + \tau(t) + D_0(t) + D_1(t), \quad (2)$$

where  $\tau = [\tau_1, \tau_2, \tau_3]$  is a three-dimensional column vector composed of forces and moments, representing the surge force  $\tau_1$ , sway  $\tau_2$  and sway moment  $\tau_3$ , respectively.  $D_0(t)$  and  $D_1(t)$  represent slowly varying environmental disturbances and long peak wave disturbances in mature period respectively.  $M$  is a inertia matrix, which is composed of hydrodynamic added inertia:

$$M = \begin{bmatrix} m - X_{\dot{u}} & 0 & 0 \\ 0 & m - Y_{\dot{v}} & mx_G - Y_{\dot{r}} \\ 0 & mx_G - N_{\dot{v}} & I_Z - N_{\dot{r}} \end{bmatrix}, \quad (3)$$

where  $m$  is the mass of the ship. The additional mass of surge, sway and yaw under the acceleration of the ship along the corresponding axis is defined as  $X_{\dot{u}} < 0, Y_{\dot{v}} < 0, N_{\dot{r}} < 0$ . Because the ship will be affected by wave drift and laminar friction when it is running on the sea surface, it will produce a linear damping matrix, which is strictly positive. The damping matrix  $D$  as follows:

$$D = \begin{bmatrix} -X_u & 0 & 0 \\ 0 & -Y_v & -Y_r \\ 0 & -N_v & -N_r \end{bmatrix}. \quad (4)$$

## 2.3 State Space Model of Dynamic Positioning Ship

Due to the small deflection angle of the ship, it is considered here

$$R(\psi) \cong I, \quad (5)$$

let  $U = \tau$ , state space model of DP ship can be developed as

$$\dot{X}(t) = AX(t) + BU(t) + BD_0(t) + BD_1(t), \quad (6)$$

$$X = [\eta^T, v^T], A = \begin{bmatrix} 0 & I \\ 0 & -M^{-1}D \end{bmatrix}, B = \begin{bmatrix} 0 \\ M^{-1} \end{bmatrix},$$

where  $A \in \mathbb{R}^{n \times n}$  is the system matrix,  $X(t) \in \mathbb{R}^n$  is the state vector,  $B \in \mathbb{R}^{n \times m}$  is the disturbance coefficient matrix,  $U(t) \in \mathbb{R}^m$  is the vector of control inputs.  $D_0(t)$  represents the slowly varying disturbances generated by wind, ocean currents, second-order wave drift and unmodeled dynamics, which were included in the bias term. The first order Markov process is used to simulate the bias model, as follows:

$$\begin{cases} D_0(t) = R^{-1}(\psi)b(t), \\ \dot{b}(t) = -T^{-1}b(t) + \Psi\xi_1(t), \end{cases} \quad (7)$$

where  $T \in \mathbb{R}^{n \times n}$  is a time constant matrix,  $b$  is the vector of bias forces and moments,  $\Psi \in \mathbb{R}^{n \times n}$  is a positive definite diagonal matrix of bounded zero-mean Gauss white noise  $\xi_1 \in \mathbb{R}^n$ , with  $\|\xi_1(t)\| \leq d^*(t)$ ,  $d^*(t)$  is a positive constant. Disturbance  $D_1(t)$  can be developed by the following stochastic exogenous system

$$\begin{cases} D_1(t) = V\omega(t), \\ \dot{\omega}(t) = W\omega(t) + B_1X(t)\xi(t), \end{cases} \quad (8)$$

where  $W \in \mathbb{R}^{r \times r}$ ,  $B_1 \in \mathbb{R}^{r \times \varsigma}$ , and  $V \in \mathbb{R}^{m \times r}$  are the known matrices.  $\xi(t)$  is white noise with finite bandwidth. Its amplitude is random at any time and satisfies the Gauss distribution function as a whole.

*Notes and Comments.* The slowly varying environmental disturbances  $D_0(t)$  will significantly affect the heading and position of the ship, and make the ship drift slowly from its original position, resulting in long-period and large amplitude periodic motion. In other words, the low-frequency motion of the ship is described here. In fact, the waves on the sea are extremely complicated. The disturbances  $D_1(t)$  represent a class of long peak wave disturbances in mature period. They can be seen as the superposition of regular wave drift forces of various frequencies.

**Assumption 1.**  $(A, B)$  is controllable,  $(T^{-1}, BR^{-1}(\psi))$  and  $(W, BV)$  are observable.

**Lemma 1** [13]: Assume exist  $V \in C^2(R^n \times R_+)$ ,  $\kappa \in K_\nu \subset K_\infty$  and positive number  $\rho, \beta, \lambda$ , such that for all  $(x, t) \in R^n \times R_+$ ,

$$\begin{aligned} \kappa(|x|)^\rho &\leq V(x, t) \text{ and,} \\ LV(X, t) &= \frac{\partial V}{\partial X}f(X, t) + \frac{1}{2}Tr\{g(X, t)^T \frac{\partial^2 V}{\partial X^2}g(X, t)\} \\ &\leq -\lambda V(x, t) + \beta. \end{aligned} \quad (9)$$

Then, the equilibrium  $X = 0$  is asymptotically bounded in  $p$ th moment.

### 3 Main Results

#### 3.1 Stochastic Disturbance Observers (SDO)

The following stochastic disturbance observer is designed to estimate slowly varying environmental disturbances

$$\begin{cases} dg(t) = (-T^{-1} - L_1BR^{-1}(\psi))\hat{b}(t)dt - L_1[AX(t) + BU(t) + B\hat{D}_1(t)]dt, \\ \hat{D}_0(t) = R^{-1}(\psi)\hat{b}(t), \hat{b}(t) = g(t) + L_1X(t), \end{cases} \quad (10)$$

where  $D_0(t)$  is the disturbance estimation,  $L_1$  is the observed gain, and  $g(t)$  is the auxiliary vector. Based on (6), (7) and (10), we have

$$de_b(t) = (-T^{-1} - L_1BR^{-1}(\psi))e_b(t)dt + \Psi\xi_1(t)dt - L_1BV e_\omega(t), \quad (11)$$

Since  $(T^{-1}, BR^{-1}(\psi))$  is observable, the pole of the error system (11) can be placed to the desired position through the pole placement [14,15].

In order to estimate the long peak wave disturbances in mature period, the stochastic disturbance observer is described by

$$\begin{cases} dq(t) = (W - L_2 BV)\hat{\omega}(t)dt - L_2[AX(t) + BU(t) + B\hat{D}_0(t)]dt, \\ \hat{D}_1(t) = V\hat{\omega}(t), \quad \hat{\omega}(t) = q(t) + L_2X(t), \end{cases} \quad (12)$$

where  $\hat{D}_1(t)$  is the estimation of  $D_1(t)$ ,  $L_2$  is the observed gain, and  $q(t)$  is the auxiliary vector. Based on (6), (8) and (12), then

$$de_\omega(t) = (W - L_2 BV)e_\omega(t)dt + B_1X(t)\xi(t) - L_2BR^{-1}(\psi)e_b(t), \quad (13)$$

The desirable system performance can be achieved by adjusting  $L_2$ . Since  $(W, BV)$  is observable, the pole of the error system (13) can be placed to the expected position through the pole placement [14,15].

Next, the DOBDAC controller is designed as

$$U(t) = -\hat{D}_0(t) - \hat{D}_1(t) + KX(t), \quad (14)$$

where  $K$  is the gain of controller. Draging (14) into (6), yield

$$dX(t) = (A + BK)X(t)dt + BR^{-1}(\psi)e_b(t)dt + BVe_\omega(t)dt. \quad (15)$$

Combining (15), (11) and (13), the composite system can be obtained

$$d(\bar{X}(t)) = \bar{A}\bar{X}(t)dt + \bar{\Psi}dW_1(t) + \bar{B}\bar{X}(t)dW(t), \quad (16)$$

where

$$\begin{aligned} \bar{A} &= \begin{bmatrix} A + BK & BR^{-1}(\psi) & BV \\ 0 & -T^{-1} - L_1BR^{-1}(\psi) & -L_1BV \\ 0 & -L_2BR^{-1}(\psi) & W - L_2BV \end{bmatrix}, \\ \bar{X}(t) &= \begin{bmatrix} X(t) \\ e_b(t) \\ e_\omega(t) \end{bmatrix}, \bar{\Psi} = \begin{bmatrix} 0 \\ \Psi \\ 0 \end{bmatrix}, \bar{B} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ B_1 & 0 & 0 \end{bmatrix}. \end{aligned}$$

### 3.2 Disturbance Observer-Based Disturbance Attenuation Control (DOBDAC)

**Theorem 1.** For DP system of ships (2) with slowly varying environmental disturbances (7) and long peak wave disturbances in mature period (8) under Assumption 1, there exist matrices  $Q_1 = P_1^{-1} > 0$ ,  $Q_2 = P_2^{-1} > 0$ ,  $Q_3 = P_3^{-1} > 0$  and  $R_1$  satisfying

$$\Upsilon = \begin{bmatrix} A_1 & BQ_2 & BVQ_3 & 0 & 0 & {Q_1B_1}^T \\ * & A_2 & -L_1BVQ_3 - Q_2^T B^T L_2^T & 0 & 0 & 0 \\ * & * & A_3 & 0 & 0 & 0 \\ * & * & * & -Q_1 & 0 & 0 \\ * & * & * & * & -Q_2 & 0 \\ * & * & * & * & * & -Q_3 \end{bmatrix} < 0, \quad (17)$$

where

$$\begin{aligned}\Lambda_1 &= AQ_1 + Q_1^T A^T + BR_1 + R_1^T B^T, \\ \Lambda_2 &= -T^{-1}Q_2 - Q_2^T(T^{-1})^T - L_1 B Q_2 - Q_2^T B^T L_1^T, \\ \Lambda_3 &= WQ_3 + Q_3^T W^T - L_2 B V Q_3 - Q_3^T V^T B^T L_2^T.\end{aligned}$$

Then, by designing stochastic disturbance observer (10) with gain  $L_1$ , stochastic disturbance observer (12) with gain  $L_2$  and controller (14) with gain  $K = R_1 Q_1^{-1}$ , the composite system (16) is asymptotically bounded in mean square.

*Proof.* Chose the following Lyapunov function

$$V(\bar{X}(t), t) = \bar{X}^T(t)P\bar{X}(t). \quad (18)$$

Setting

$$P = \begin{bmatrix} P_1 & 0 & 0 \\ 0 & P_2 & 0 \\ 0 & 0 & P_3 \end{bmatrix} = \begin{bmatrix} Q_1^{-1} & 0 & 0 \\ 0 & Q_2^{-1} & 0 \\ 0 & 0 & Q_3^{-1} \end{bmatrix} > 0. \quad (19)$$

Differentiating (17) along with (15), yields

$$\begin{aligned}LV(\bar{X}(t), t) &= \frac{\partial \nu}{\partial \bar{X}}[\bar{A}\bar{X}^T(t)]dt + Tr\{\bar{X}^T(t)\bar{B}^TP\bar{B}\bar{X}^T(t)\} + Tr\{\bar{\Psi}^TP\bar{\Psi}\} \\ &\leq \bar{X}^T(t)(P\bar{A} + \bar{A}^TP)\bar{X}(t) + \bar{X}^T(t)(\bar{B}^TP\bar{B})\bar{X}(t) + Tr\{\bar{\Psi}^TP\bar{\Psi}\} \\ &= \bar{X}^T(t)(P\bar{A} + \bar{A}^TP + \bar{B}^TP\bar{B})\bar{X}(t) + Tr\{\bar{\Psi}^TP\bar{\Psi}\} \\ &= \bar{X}^T(t)\Upsilon_1\bar{X}(t) + \gamma(t),\end{aligned}$$

where

$$\Upsilon_1 = \begin{bmatrix} P\bar{A} + \bar{A}^TP & \bar{B}^T \\ * & -P^{-1} \end{bmatrix}, \gamma(t) = Tr\{\bar{\Psi}^TP\bar{\Psi}\}. \quad (20)$$

If there exists a constant  $\beta \geq 0$ , such that  $0 \leq \gamma(t) \leq \beta$ . Then, following inequality can be obtained

$$LV(\bar{X}(t), t) \leq \bar{X}^T(t)\Upsilon_1\bar{X}(t) + \gamma(t) \leq \bar{X}^T(t)\Upsilon_1\bar{X}(t) + \beta. \quad (21)$$

Next, we will prove that  $\Upsilon < 0 \Leftrightarrow \Upsilon_1 < 0$ .

1):  $\Upsilon_1 < 0 \Leftrightarrow \Upsilon_2 < 0$ , yields

$$\Upsilon_2 = \begin{bmatrix} \Xi_1 & P_1 B & P_1 B V & 0 & 0 & B_1^T \\ * & \Xi_2 & -P_2 L_1 B V - B^T L_2^T P_3 & 0 & 0 & 0 \\ * & * & \Xi_3 & 0 & 0 & 0 \\ * & * & * & -P_1^{-1} & 0 & 0 \\ * & * & * & * & -P_2^{-1} & 0 \\ * & * & * & * & * & -P_3^{-1} \end{bmatrix} < 0, \quad (22)$$

where

$$\begin{aligned}\Xi_1 &= P_1 A + A^T P_1 + P_1 B K + K^T B^T P_1, \\ \Xi_2 &= -P_2 T^{-1} - (T^{-1})^T P_2 - P_2 L_1 B - B^T L_1^T P_2, \\ \Xi_3 &= P_3 W + W^T P_3 - P_3 L_2 B V - V^T B^T L_2^T P_3.\end{aligned}$$

2):  $\Upsilon_2 < 0 \Leftrightarrow \Upsilon_3 < 0$ .  $\Upsilon_2 < 0$  is pre-multiplied and post-multiplied simultaneously by  $\text{diag}\{Q_1, Q_2, Q_3, I, I, I\}$ , yields

$$\Upsilon_3 = \begin{bmatrix} \Lambda_1 & BQ_2 & BVQ_3 & 0 & 0 & {Q_1 B_1}^T \\ * & A_2 & -L_1 BVQ_3 - Q_2^T B^T L_2^T & 0 & 0 & 0 \\ * & * & \Lambda_3 & 0 & 0 & 0 \\ * & * & * & -Q_1 & 0 & 0 \\ * & * & * & * & -Q_2 & 0 \\ * & * & * & * & * & -Q_3 \end{bmatrix} < 0, \quad (23)$$

where

$$\begin{aligned}\Lambda_1 &= AQ_1 + Q_1^T A^T + BR_1 + R_1^T B^T, \\ \Lambda_2 &= -T^{-1}Q_2 - Q_2^T (T^{-1})^T - L_1 B Q_2 - Q_2^T B^T L_1^T, \\ \Lambda_3 &= WQ_3 + Q_3^T W^T - L_2 B V Q_3 - Q_3^T V^T B^T L_2^T.\end{aligned}$$

3):  $\Upsilon_3 < 0 \Leftrightarrow \Upsilon < 0$ . It can be shown that  $\Upsilon_3 < 0 \Leftrightarrow \Upsilon < 0$  based on  $K = R_1 Q_1^{-1}$  in (23).

Based on the above discussion and Lemma 1, we can prove that all state signals of composite system are asymptotically bounded in mean square.

## 4 Simulation Example

The effectiveness of the method is verified by simulation on a supply ship. Referring to [6–8], the length and mass of ship are  $7.620 \times 10^1$  m and  $4.591 \times 10^6$  kg respectively. The dynamic parameters of the DP ship system (2) are given by:

$$\begin{aligned}M &= \begin{bmatrix} 5.3122 \times 10^6 & 0 & 0 \\ 0 & 8.2831 \times 10^6 & 0 \\ 0 & 0 & 3.7454 \times 10^9 \end{bmatrix}, \\ D &= \begin{bmatrix} 5.0242 \times 10^4 & 0 & 0 \\ 0 & 2.7229 \times 10^5 & -4.3933 \times 10^6 \\ 0 & -4.3933 \times 10^6 & 4.1894 \times 10^8 \end{bmatrix}.\end{aligned}$$

Take  $X(0) = [20, 20, 10, 0, 0, 0]^T$  as the initial state, where  $v_0 = [0m/s, 0m/s, 0rad/s]^T$  and  $\eta_0 = [20m, 20m, 10^\circ]^T$ . And the parameters of the slowly varying

environmental disturbances (7) and the long peak wave disturbance in mature period (8) are as follows

$$T = \begin{bmatrix} 1.000 \times 10^3 & 0 & 0 \\ 0 & 1.000 \times 10^3 & 0 \\ 0 & 0 & 1.000 \times 10^3 \end{bmatrix}, \Psi = \begin{bmatrix} 130 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 80 \end{bmatrix},$$

$$V = \begin{bmatrix} 6.2300 \times 10^3 & 5.0050 \times 10^4 & 8.0000 \times 10^3 \\ 5.0000 \times 10^3 & -9.0000 \times 10^3 & 5.0800 \times 10^4 \\ -9.4620 \times 10^6 & 6.0800 \times 10^3 & -5.5000 \times 10^3 \end{bmatrix},$$

$$W = \begin{bmatrix} 3.0000 \times 10^{-1} & 0 & 0 \\ 0 & 0 & -3.0000 \times 10^{-1} \\ 0 & 3.0000 \times 10^{-1} & 0 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} 1.0000 \times 10^{-3} & 1.0000 \times 10^{-3} & 1.0000 \times 10^{-3} \\ 2.3000 \times 10^{-3} & -3.0000 \times 10^{-3} & -4.0000 \times 10^{-3} \\ 9.0000 \times 10^{-3} & 2.0000 \times 10^{-3} & 0 \\ 2.0000 \times 10^{-3} & 1.0010 \times 10^{-3} & 1.5000 \times 10^{-3} \\ 2.0000 \times 10^{-3} & 2.1000 \times 10^{-2} & 1.8000 \times 10^{-2} \\ 1.0000 \times 10^{-3} & 1.0000 \times 10^{-3} & 1.0000 \times 10^{-3} \end{bmatrix}^T.$$

By placing the poles  $J_1$  at  $[-0.15 - 0.16 - 0.14]$  in (12) and  $J_2$  at  $[-30 - 30 - 30]$  in (15), we can obtain

$$L_1 = \begin{bmatrix} 0 & 0 & 0 & 2.8400 \times 10^{-2} & -5.5300 \times 10^{-2} & 2.0188 \times 10^{-2} \\ 0 & 0 & 0 & -6.1743 \times 10^{-1} & 1.6786 \times 10^{-1} & -2.4652 \times 10^{-1} \\ 0 & 0 & 0 & -8.3439 \times 10^0 & -6.3177 \times 10^{-1} & 1.8969 \times 10^{-1} \end{bmatrix},$$

$$L_2 = \begin{bmatrix} 0 & 0 & 0 & 1.6000 \times 10^8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.5000 \times 10^8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.1236 \times 10^{11} \end{bmatrix}.$$

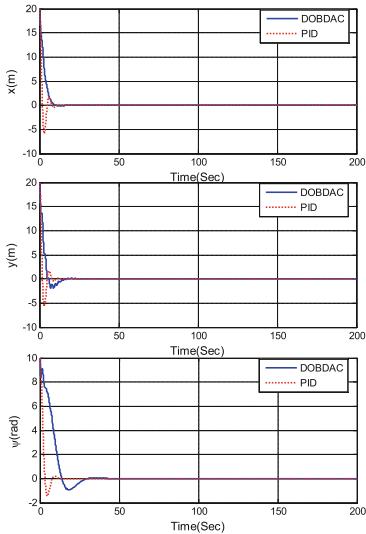
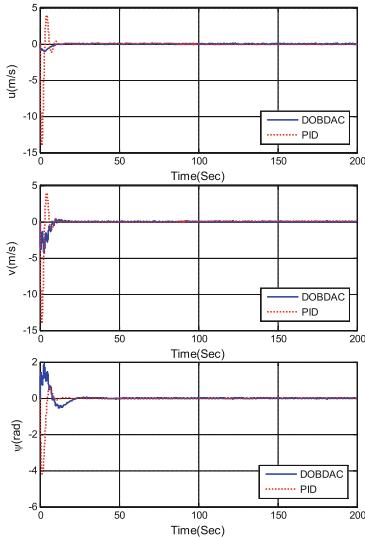
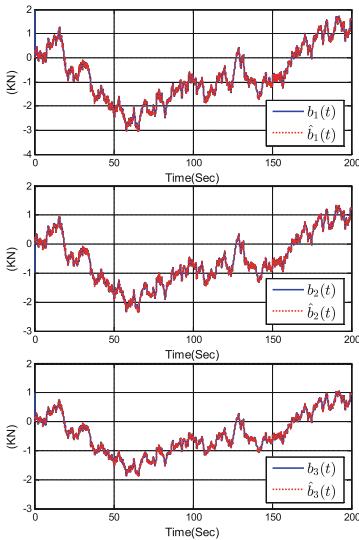
According to Theorem 1, yield

$$K = \begin{bmatrix} -9.3000 \times 10^1 & 0 & 0 \\ 0 & -6.0000 \times 10^1 & -2.2600 \times 10^2 \\ 0 & -1.0000 \times 10^0 & -3.0000 \times 10^0 \\ -1.2052 \times 10^4 & 0 & 0 \\ 0 & -2.9490 \times 10^3 & -1.1399 \times 10^4 \\ 0 & -1.4257 \times 10^4 & -5.5817 \times 10^4 \end{bmatrix}^T.$$

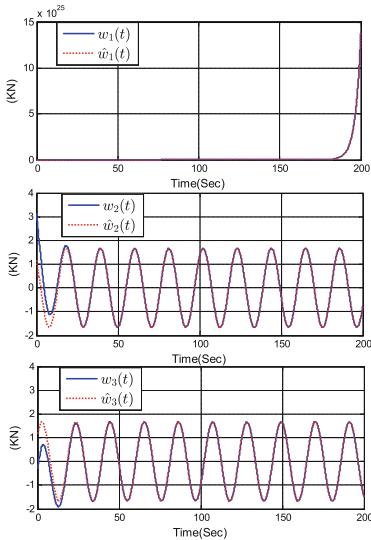
To illustrate the effectiveness of the proposed DOBDAC control scheme, the following PID control law (24) was used for simulation.

$$\tau_{PID} = K_p \eta(t) + K_I \int_0^t \eta(\tau_t) d\tau_t + K_D \dot{\eta}(t). \quad (24)$$

where  $K_p = diag(1 \times 10^4, 2 \times 10^4, 1.5 \times 10^5)$ ,  $K_I = diag(10, 10, 150)$  and  $K_D = diag(1.4 \times 10^6, 2.8 \times 10^6, 1 \times 10^9)$ .

(a) The response of ship position information ( $x, y, \psi$ ).(b) The response of ship velocity vector ( $u, v, r$ ,  $\dot{\psi}$ ).

(c) Trajectory of slowly varying environmental disturbance estimation.



(d) Trajectory of long peak wave disturbance in mature period estimation.

**Fig. 1.** (a) The response of ship position information ( $x, y, \psi$ ), (b) The response of ship velocity vector ( $u, v, r$ ), (c) Trajectory of slowly varying environmental disturbance estimation, (d) Trajectory of long peak wave disturbance in mature period estimation.

Figure 1(a) display that the designed DOBDAC scheme can regulate the ships position and heading at desired position compared with PID control. Figure 1(b) illustrate that the more significant advantages can be achieved on the ship velocities ( $u, v, r$ ) by DOBDAC scheme. Figure 1(c) demonstrate the designed stochastic disturbance observer (10) is successful to estimate the slowly varying environmental disturbances; Figure 1(d) show that the designed stochastic disturbance observer (12) has good disturbance estimation performances for long peak wave disturbances in mature period. The simulation figures show that through the designed DOBDAC scheme, the system of DP ship has achieved satisfactory results.

## 5 Conclusions

The anti-disturbance problem is discussed for DP system of ships with multiple heterogeneous disturbances in this paper. The slowly varying environmental disturbances and the long peak wave disturbance in mature period are estimated through the designed stochastic disturbance observers, respectively. Then, a DOBDAC scheme is developed to ensure the asymptotically bounded in mean square of all signals in the DP composite system.

**Acknowledgments.** This work was supported by the National Natural Science Foundation of China 61973149.

## References

1. Deng, F., Yang, H.L., Wang, L.J.: Adaptive unscented Kalman filter based estimation and filtering for dynamic positioning with model uncertainties. *Int. J. Control Autom. Syst.* **17**(3), 667–678 (2019). <https://doi.org/10.1007/s12555-018-9503-4>
2. Sørensen, J.A.: A survey of dynamic positioning control systems. *Ann. Rev. Control* **35**(1), 123–136 (2011). <https://doi.org/10.1016/j.arcontrol.2011.03.008>
3. Fossen, T.I., Grøvlen, A.: Nonlinear output feedback control of dynamically positioned ships using vectorial observer backstepping. *IEEE Trans. Control Syst. Technol.* **6**(1), 121–128 (1998). <https://doi.org/10.1109/87.668046>
4. Fossen, T.I., Strand, J.P.: Passive nonlinear observer design for ships using Lyapunov methods: full-scale experiments with a supply vessel. *Automatica* **35**(1), 3–16 (1999). [https://doi.org/10.1016/S0005-1098\(98\)00121-6](https://doi.org/10.1016/S0005-1098(98)00121-6)
5. Balchen, J.G., Jenssen, N.A., Eldar, M., Saelid, S.: A dynamic positioning system based on Kalman filtering and optimal control. *Model. Identif. Control* **1**(3), 135–163 (1980). <https://doi.org/10.4173/mic.1980.3.1>
6. Du, J., Hu, X., Krsti, M., Sun, Y.: Robust dynamic positioning of ships with disturbances under input saturation. *Automatica* **73**(3), 207–214 (2016). <https://doi.org/10.1016/j.automatica.2016.06.020>
7. Du, J., Hu, X., Krsti, M., Sun, Y.: Dynamic positioning of ships with unknown parameters and disturbances. *Control Eng. Pract.* **76**(3), 22–30 (2018). <https://doi.org/10.1016/j.conengprac.2018.03.015>

8. Hu, X., Du, J., Zhu, G., Sun, Y.: Robust adaptive NN control of dynamically positioned vessels under input constraints. *Neurocomputing* **318**(27), 201–212 (2018). <https://doi.org/10.1016/j.neucom.2018.08.056>
9. Wei, X., Guo, L.: Composite disturbance-observer-based control and  $H_\infty$  control for complex continuous models. *Int. J. Robust Nonlinear Control* **20**(1), 106–118 (2010). <https://doi.org/10.1002/rnc.1425>
10. Guo, L., Cao, S.Y.: Anti-disturbance control theory for systems with multiple disturbance: a survey. *ISA Trans.* **53**(4), 846–849 (2014). <https://doi.org/10.1016/j.isatra.2013.10.005>
11. Wei, X.J., Wu, Z.J., Hamid, R.K.: Disturbance observer-based disturbance attenuation control for a class of stochastic systems. *Automatica* **63**(162), 21–25 (2016). <https://doi.org/10.1016/j.automatica.2015.10.019>
12. Wei, X.J., Sun, S.X.: Elegant anti-disturbance control for discrete-time stochastic systems with nonlinearity and multiple disturbances. *Int. J. Control* **91**(3), 706–714 (2018). <https://doi.org/10.1080/00207179.2017.1291996>
13. Deng, H., Krstic, M., Williams, R.: Stabilization of stochastic nonlinear system driven by noise of unknown covariance. *IEEE Trans. Autom. Control* **46**(8), 1237–1253 (2001). <https://doi.org/10.1109/ACC.1998.694673>
14. Willems, J.L., Willems, J.C.: Feedback stabilizability for stochastic systems with state and control dependent noise. *Automatica* **12**, 277–283 (1976). [https://doi.org/10.1016/0005-1098\(76\)90029-7](https://doi.org/10.1016/0005-1098(76)90029-7)
15. Zhang, W.H., Chen, B.: On stabilizability and exact observability of stochastic systems with their applications. *Automatica* **40**(1), 87–94 (2004). <https://doi.org/10.1016/j.automatica.2003.07.002>



# Multiple Change Points Detection Method Based on TSTKS and CPI Sliding Window Strategy

Jialun Liu, Jinpeng Qi<sup>(✉)</sup>, Junchen Zou, and Houjie Zhu

School of Information Science and Technology, Donghua University,  
Shanghai 200051, China  
[Riva\\_Liu@163.com](mailto:Riva_Liu@163.com), [qipengkai@dhu.edu.cn](mailto:qipengkai@dhu.edu.cn)

**Abstract.** This paper takes time series data as the research object, proposes a TSTKS (Trigeminal search tree and KS Statistic) change point detection method based on CPI (Change Point Interval) sliding window strategy. First of all, this paper constructs a TSTKS change point detection method, which improves the accuracy of change point detection. Secondly, for the characteristics of multiple change points for time series data. Based on the fixed sliding window model, a CPI sliding window strategy is proposed, conditionally change the window size of the sliding window. It improves the accuracy of change detection and shortens the detection time. Finally, in order to verify the effect of the TSTKS change point detection method based on CPI sliding window strategy, simulation experiments were performed and used to detect the change points of epilepsy brain wave data. The results show the effectiveness of the proposed method and has practical application value.

**Keywords:** Change point detection · CPI sliding window strategy · TSTKS anomaly detection method · Time series

## 1 Introduction

With the advent of the information age, the output of data is growing at an order of magnitude. These data include industrial production data, commercial transaction data, financial stock data, meteorological data, user behavior information data, etc. At the same time, more and more enterprises pay attention to these data, which is regarded as the valuable wealth of enterprises [1–3]. In order to promote the development of companies and government agencies, all walks of life began to mine and analyze a large number of data, in-depth understanding of the relevance and differences between the data. According to the change point detection method is an important research branch in the field of modern data mining technology.

Inspired by the theory of random complexity, Jimmy Baikovicus et al. proposed a method based on the ARMA (autoregressive moving average) model [4].

The central idea of this method is to apply the minimum description length in the form of predicting random complexity, which provides an idea for selecting the best model from a given set of models. Kawahara Youshinobu et al. proposed a non-parametric detection method based on direct density ratio [5]. This method avoids the problem of non-parametric density estimation by estimating the ratio of probability density. Takehisa Yairi et al. proposed a series of change point detection methods based on subspace recognition [6]. This method considers that the subspace spanned by the columns of the observability matrix is approximately equal to the subspace spanned by the subsequences of the time series data. Mehdi Sharifzadeh et al. used the characteristics of footprint transformation data to generate non-zero coefficients only at changing points, and proposed a method for detecting change points based on wavelet footprints [7]. Wenhua Chen et al. introduced wavelet transform technology into the change detection framework, and proposed a change detection method based on local energy feature changes [8]. Shows the obvious advantages of the wavelet transform detection method. Huang Jing et al. proposed a change point detection method based on multi-scale straight line fitting in 2015 [9]. This method divides the time series into several segments according to the initial method, and uses the straight line fitted by the least square method to replace several sub-segments. Select the two line segments that satisfy the maximum slope difference between two adjacent straight lines as the range of the next fitting and reduce the fitting scale. Repeat the above steps until the fitting scale converges to 1.

This paper takes time series data as the research object, and proposes a TSTKS change point detection method based on CPI sliding window strategy. In order to increase the speed of change detection, a TSTKS change detection method is proposed based on the HWKS (Harr Wavelet and KS Statistic) method. This method adds a virtual branch between the mean binary tree and the difference binary tree to improve the accuracy of change point detection. According to the single time dimension of time series data, a CPI sliding window strategy is proposed. The method changes the window size of the sliding window by conditions, which improves the accuracy of the detection of the change point and the speed of the detection of the abnormal point. Simulation analysis shows that the TSTKS change point detection method based on the CPI sliding window strategy proposed in this paper has better change point detection effect.

## 2 Theoretical Principles

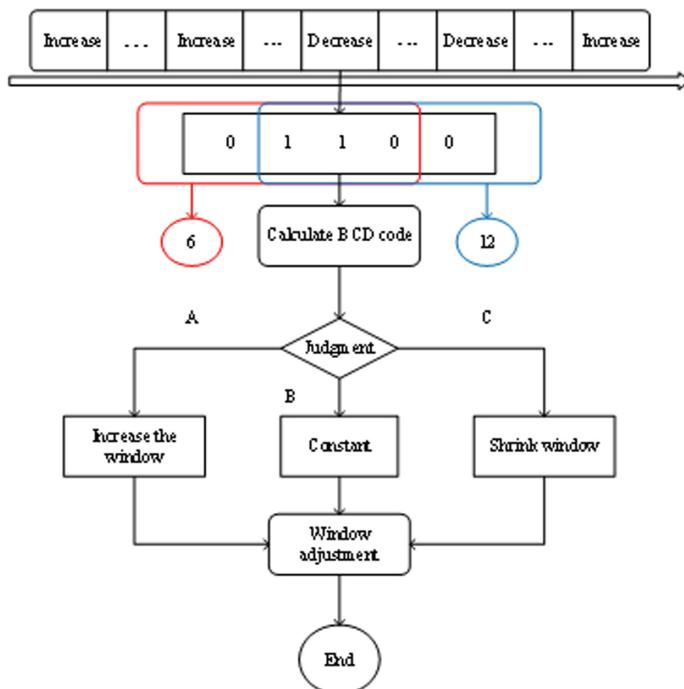
### 2.1 CPI Sliding Window Strategy

In view of the feature that large-scale time series data will increase rapidly at a certain time, a buffer area needs to be established to store data that has not been processed. After the cache area is established, the detection method processes the data in the cache area. However, the cache area is generally large, and the detection method needs to process the data in the cache area in batches. Through batch processing, the concept of sliding windows is introduced [10, 11].

The size of the window and the method of updating the window are two important parameters of the sliding window. According to the different window settings, it is divided into sliding windows based on time intervals and sliding windows based on the number of elements. The size of the sliding window determines the amount of data observed each time, and the size of the observed data directly affects the detection effect of the change point. Therefore, determining the size of the sliding window plays an important role in detecting abnormal points [12].

The CPI sliding window strategy not only considers the overall distribution characteristics of the data, but also the local characteristics of the data during the selection of the sliding window size, and dynamically adjusts the size of the sliding window, thereby improving the effect of change point detection.

The CPI sliding window strategy is based on the knowledge of the BCD code, and maintains the latest change point interval characteristics of the four windows during the change point detection process, arranged in a queue from right to left, first in first out. Among them, the characteristic of the change point interval is the size of the interval between this change point and the interval between the previous change point. If the interval becomes larger, the change point interval feature is recorded as 0, otherwise it is recorded as 1, so as to obtain a 0–1 queue, and convert it into a decimal value, adjust the size of the sliding window by the size of the value. The specific principle of the CPI sliding window strategy is shown in Fig. 1:

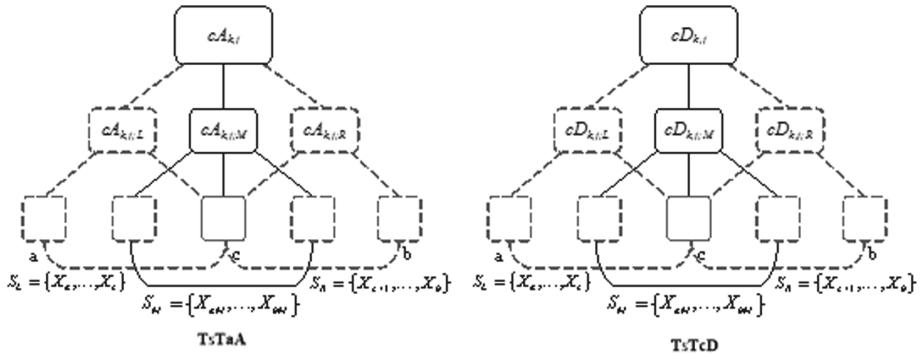


**Fig. 1.** The CPI sliding window strategy

In several consecutive windows in the figure, 5 windows have detected change points, and the change of the interval relative to the previous change point is increase, increase, decrease, decrease and increase respectively. Make an observation every four 0–1 sequences. The 0–1 distribution in the red box on the left in the figure is 0110, and the corresponding code value is 6, which is located between [5, 10], and the window remains unchanged. The 0–1 distribution in the blue box on the right is 1100, and the corresponding code value is 12, which is located between [11]. If the code value is at [0, 4], the window should be increased appropriately.

## 2.2 TSTKS Detection Method

The TSTKS detection method is an improvement of the HWKS method and belongs to a multi-path change detection method. The TSTKS detection method is to build a trigeminal tree on the basis of a binary tree, and add an intermediate branch between the left and right subtrees of the binary tree, thereby dividing the data into three segments. By constructing a trigeminal tree, the detection effect of the change point appearing at the middle position in the process of abnormal point detection is improved. The TSTKS method first performs multilevel haar wavelet decomposition on the data, and constructs a mean trigeminal tree (TsTaA) and a difference trigeminal tree (TsTcD), as shown in Fig. 2.



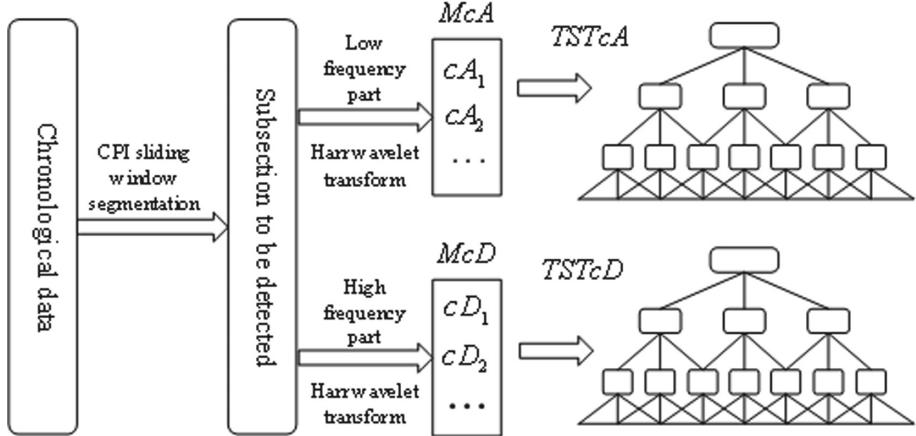
**Fig. 2.** TsTaA and TsTcD

It is an improvement to the HWKS method by constructing a trigeminal tree to improve the detection effect of the change point that appears in the middle position during the abnormal point detection process. Among them, the expressions of the mean parameter node  $cA_{k,j;M}$  and the difference parameter node  $cD_{k,j;M}$  corresponding to the added middle branch are:

$$cA_{k,j;M} = \frac{cA_{k-2,4j-2} + cA_{k-2,4j-1}}{\sqrt{2}} \quad (1)$$

$$cD_{k,j;M} = \frac{cD_{k-2,4j-2} + cD_{k-2,4j-1}}{\sqrt{2}} \quad (2)$$

### 3 CPI-TSTKS Change Detection Method



**Fig. 3.** The flow chart of CPI-TSTKS change point detection method

The TSTKS change point detection method based on CPI sliding window strategy proposed in this paper is based on the TSTKS method and sliding window principle. At the same time, it combines the CPI sliding window adjustment strategy, which acts on the detection of sudden changes in time series data. The specific flowchart is shown in Fig. 3.

The TSTKS method mainly involves three change point search steps. The first two steps are change point search based on detail fluctuation and statistical fluctuation, which is consistent with the search step in the HWKS method to determine the best search path for the change point. The third search step further determines the precise position information of the change point on this basis. According to the two-sample KS statistical theory. Calculate the statistical fluctuation between the distribution function to be detected  $F_m(x)$  and the known standard distribution function  $G_n(x)$ . And select the leaf node with the largest statistical fluctuation among the bottom leaf nodes as the change point.

## 4 Simulation Analysis

### 4.1 Detection and Evaluation Index

In order to objectively evaluate the detection performance of the TTKKS change point detection method based on the CPI sliding window strategy proposed in this paper, three detection evaluation indicators are used in the simulation analysis:

- (1) Hit rate, which represents the proportion of correctly identified change points to the number of all change points, expressed as:

$$Hit = \frac{TP}{TP + FN} = \frac{TP}{n} \quad (3)$$

The meanings of TP, FP, FN and TN are shown in Table 1:

**Table 1.** Parameter meaning

	Number of change points detected	Number of non-change points detected
The actual number of change point	TP	FN
Actually the number of non-abrupt points	FP	TN

- (2) Time consumption refers to the time it takes for the change detection algorithm to start and finish:

$$Tim = TimeEnd - TimeStart \quad (4)$$

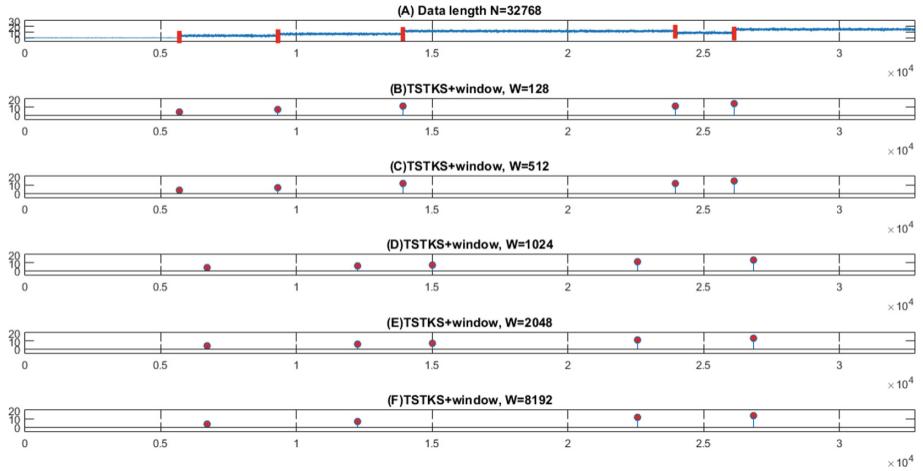
- (3) The mean absolute error represents the ratio of the sum of the distances between each change point detected by the algorithm and the actual change point to the number of change points. Expressed as:

$$MaE = \frac{\sum_{j=1}^c \sum_{i=1}^n |Test(j) - Actual(i)|}{n} \quad (5)$$

Where  $i, j$  are positive integers,  $i \in [1, n], j \in [1, c]$ . Test represents the set of detected change points, the set size is  $c$ . Actual represents the collection of actual change points, the set size is  $n$ .

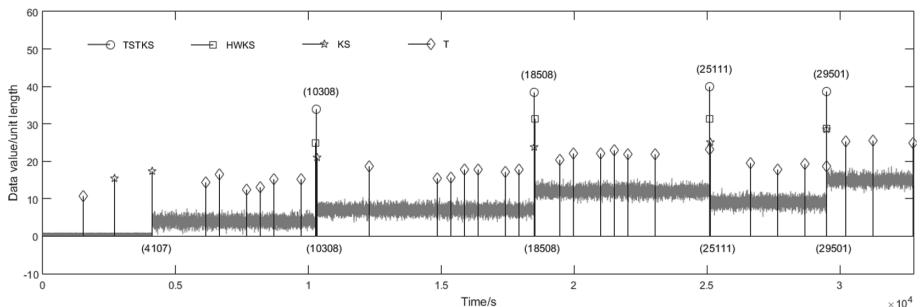
## 4.2 Detection and Evaluation Index

In order to verify the detection effect of the multiple change points detection method based on the CPI sliding window model, the experimental part first determines the optimal sliding window size, performs the detection of multiple change points. Then, introduce the CPI sliding window strategy into the simulation comparison experiment, performs the detection of multiple change points, and verifies the effectiveness of the CPI sliding window strategy. Finally, select actual data to verify the importance of the CPI sliding window strategy in practical applications.



**Fig. 4.** The effect of fixed sliding window size on the detection result

**Multiple Change Detection Based on Fixed Slide Window.** In the multiple change points detection experiment, multiple change points are randomly set as the detected target. The sliding window is used in the random multiple change detection experiment. Therefore, the influence of the sliding window size on the detection result must be considered. Therefore, the most suitable sliding window size must be selected first, as shown in Fig. 4. It is a random 5 change point detection experiment, and finally determined the optimal sliding window size is 512. In Fig. 5, the detection results of the four change detection algorithms TSTKS, HWKS, KS and T are represented by different graphs. At the same time, Table 3 shows the average time consumption and average absolute error obtained after 10,000 simulations of these four algorithms and the statistical results of the average hit rate (Table 2).

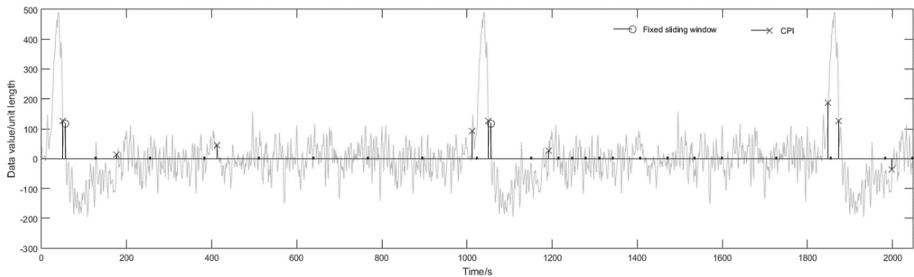


**Fig. 5.** Multiple change points detection results based on fixed sliding windows

**Table 2.** The average test result of 10000 multiple change detection based on random slide window model

	TSTKS	HWKS	KS	T
Time consumption(s)	0.1486	0.0671	0.7421	1.8720
Mean absolute error	3.9526	4.5512	3.6284	54.3817
Hit rate (%)	86.09	80.21	85.22	27.43

By observing the simulation experiment results of random multiple change points, it has been verified that the TSTKS change detection method has higher detection accuracy than the HWKS, KS and T change detection methods in random multiple change detection.



**Fig. 6.** Comparison of multiple change point detection results between CPI strategy model and fixed window model

**Multiple Change Detection Experiment Based on CPI Sliding Window Strategy.** The above simulation experiment proves the effectiveness of TSTKS change point detection method. In this experiment, CPI sliding window strategy is applied to it, and the random multiple change points detection experiment is carried out in combination with the simulation data. As shown in Fig. 6, the CPI sliding window strategy is added to the TSTKS change point detection method, and an experimental comparison is made with the fixed sliding window strategy. Table 4 gives the statistical results of the average time consumption, mean absolute error, and average hit rate obtained after 10,000 simulations.

The experiments compare the detection effect of the TSTKS method based on the fixed sliding window strategy and the TSTKS method based on the CPI sliding window strategy in the detection of change points, proving that the CPI sliding window strategy has higher detection accuracy and better adaptability.

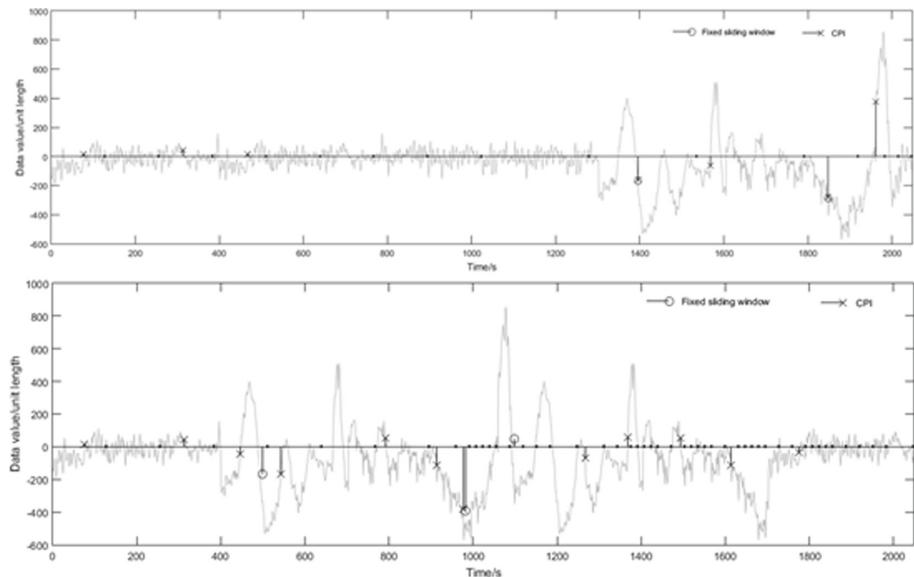
**Comparative Experiment of Real EEG Data Detection.** In order to further prove the detection effect of the TSTKS change points detection method

**Table 3.** Statistical table of random multiple change detection results of CPI strategy model and fixed window model

	Time consumption (s)	Mean absolute error	Hit rate (%)
Fixed sliding window model	0.5658	4.1128	37.52
CPI strategy model	0.1503	3.0178	72.45

based on CPI sliding window strategy proposed in this paper in actual data. In this experiment, two sets of real EEG data with a length of 2048 were selected. The experimental data comes from the PhysioBank shared database, which shares data about patients with sudden cardiac death, epilepsy, gait disorders, etc. In this experiment, part of normal and abnormal epilepsy brain wave data was extracted as experimental sample data.

In view of the difference in brain wave data at the time of onset, the experiment used brain wave data from a normal state to the onset state and the process of an epilepsy patient. A fixed window model and a CPI strategy model were used to conduct a comparative test of multiple change point detection. The comparison results of the change point detection of the two models are shown in Fig. 7.

**Fig. 7.** Comparison results of the detection effect between the fixed window model and the CPI strategy model

For real EEG data, the Hit and the MaE cannot be used to measure the algorithm's change detection performance. Therefore, SF (statistical fluctuation)

and DF (detail fluctuation) are used to measure the accuracy of the algorithm. Among them, SF represents the variance fluctuation of the two distributions before and after the change point, and DF represents the fluctuation of the standard deviation of the distribution at both ends before and after the change point. Since there are multiple change points and the number of change points detected by the two algorithms is different, the cumulative sum of SF and DF is averaged. The results are shown in Table 4. According to the meanings of SF and DF, the larger the value of these two data, the more accurate the selected change point position and the more prominent the detection result.

**Table 4.** Statistical table of EEG data detection results of CPI strategy model and fixed window model

Model	Time consumption (s)	DF	SF
Fixed sliding window model	0.0761	0.1713	7.3208
CPI strategy model	0.0080	0.3002	40.5551
Fixed sliding window model	0.1120	0.2273	14.1242
CPI strategy model	0.0062	0.2494	33.0323

Analysis of Fig. 7 and Table 4 shows that the value of the SF and the DF of the fixed window model are smaller than the CPI sliding window strategy, and the time consumption is large. In the early stage of the onset of epilepsy, the data distribution is relatively smooth, and during the onset, the distribution of the EEG data is more intense and complex, and the number of change points is greater. It can be seen that in these two practical application scenarios, the detection effect of the CPI sliding window strategy is significantly better than the detection effect of the fixed window model, and the detection result is more accurate. Fully show the practicality and great guiding significance of CPI sliding window in the actual detection of epilepsy brain waves.

Through the above simulation experiment, Verified the effectiveness of TSTKS detection method in detecting random single change points and random multiple change points. Also validated the effectiveness of the CPI sliding window strategy in practical applications. Furthermore, it proves that the TSTKS change point detection method based on CPI sliding window strategy proposed in this paper is effective and has application value in the detection of time series data change points.

## 5 Conclusion

This paper takes time series data as the research object, and proposes a TSTKS change point detection method based on CPI sliding window strategy. In order to improve the accuracy of change point detection, improvements have been made on the basis of the HWKS method, and then constructed a TSTKS change point

detection method. According to the multiple change points in time series data, a CPI sliding window strategy is proposed. By changing the window size of the sliding window under certain conditions, the accuracy of change point detection is improved, and the speed of abnormal point detection is also improved. After simulation experiments and analysis, using the method to carry out mutation detection on epilepsy brain wave data, it proves that the TSTKS change point detection method based on the CPI sliding window strategy proposed in this paper has a better change point detection effect.

## References

1. Bifet, A., Holmes, G., Kirkby, R., et al.: MOA: massive online analysis. *J. Mach. Learn. Res.* **11**(2), 1601–1604 (2010)
2. Chen, H.L., Chen, M.S., Lin, S.C.: Catching the trend: a framework for clustering concept-drifting categorical data. *IEEE Trans. Knowl. Data Eng.* **21**(5), 652–665 (2009)
3. Qi, J.P., Zhang, Q., Zhu, Y., Qi, J.: A novel method for fast change-point detection on simulated time series and electrocardiogram data. *PLoS ONE* **9**(4), 1–15 (2014)
4. Baikovicius, J., Gerencser, L.: Change point detection in a stochastic complexity framework. In: 1990 29th IEEE Conference on Decision and Control, Honolulu, USA, pp. 3554–3555 (1990)
5. Yoshinobu, K., Masashi, S.: Change-point detection in time-series data by direct density-ratio estimation. In: 2009 SIAM International Conference on Data Mining, pp. 389–400 (2009)
6. Kawahara, Y., Yairi, T., Machida, K.: Change-point detection in time-series data based on subspace identification. In: 2007 Seventh IEEE International Conference on Data Mining, Omaha, pp. 559–564 (2007)
7. Sharifzadeh, M., Azmoekeh, F., Shahabi, C.: Change detection in time series data using wavelet footprints. In: Lecture Notes in Computer Science. Springer, Heidelberg, pp. 127–144 (2005)
8. Wenhua, C., Jay, K.C.: Change-point detection using wavelets. In: Digital Signal Processing Technology (1996)
9. Huang, J., Li, C.C., Yan, H., et al.: Application of multi-scale line fitting method in change point detection in time series. *ACTA ARMAMENTAR* **36**(6), 1110–1116 (2015)
10. Song, Q.H., Qi, J.P., Zhang, Y.: A method for fast abrupt-point detection based on Haar Wavelet and KS statistic. *Comput. Eng.* **44**(5), 14–18 (2018)
11. Pang, J.Y.: Asaotive anomaly detection for data stream of sequence-based sliding windows model. In: Harbin Institute of Technology (2013)
12. Song, Q.H.: Research and implementation of a fast detection algorithm for abrupt-point change of data stream. In: Donghua University (2018)



# Steam Pressure Control Based on PLC 200 Smart

Jing Lv<sup>(✉)</sup> and Zhongsuo Shi

School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China  
603991040@qq.com, szs919@163.com

**Abstract.** This paper presents a PLC control system based on the principle of expert system, which is used to adjust the steam pressure. According to the reconstruction requirements of a property heat exchange station project in Beijing, the existing problems of the steam pressure regulating system are found out and the solutions are given. This paper first introduces the meaning of expert system, and then designs a steam pressure regulating system based on expert system, including the hardware structure and software design. Siemens S7-200 smart PLC is used to complete the transformation of the original heat exchange system. The field operation results show that the system designed in this paper runs stably and the error meets the requirements.

**Keywords:** PLC 200 smart · Steam pressure · Expert system

## 1 Introduction

City heating is an important cause, which is related to people's livelihood and social harmony and stability. It is an important matter related to social stability and sustainable economic development to do a good job of city heat supply. In the 21st century, China will enter a well-off society. People's living and working environment must meet the comfortable standard. In summer, there must be air conditioning, and in winter, there must be heating. In order to enter a well-off society, urban heating system is an essential project [1].

The heat exchange station is the base station connecting the heat source plant and the heat users, and it is the key link of the whole heating system. The work flow of the heat exchange station is as follows: the high-temperature steam provided by the heat source is delivered to the heat exchange station by the primary pipe network. In the heat exchange station, through the heat exchanger, the high-temperature steam on the primary side exchanges heat with the water on the secondary side and the heat energy is delivered to each heat user through the secondary pipe network. After the temperature of the hot water decreases, it returns to the return pipe of the secondary network and finally returns to the heat source [2].

The project is a steam pressure renovation project of a property heat exchange station in Beijing, which provides the power of summer air conditioning and winter heating for the building. The system uses steam as the heat source, and the heat exchange station aims to control the steam pressure on the secondary side at about 0.36 MPa. When the pressure feedback value on the secondary side is more than 0.43 MPa or less than 0.20 MPa, an alarm will be generated.

## 2 Scheme Design

The traditional PID control is used to regulate the steam pressure in the existing heat exchange station, which can be controlled accurately. But because of PID control, the parameters change quickly and the motor starts frequently, which leads to the motor burnout. Through the above analysis of the current situation of the system, it can be concluded that the frequent operation of the system causes the motor to burn out. In order to solve the above problems, other control methods can be used, such as adaptive control, fuzzy control, predictive control, neural network control, expert intelligent control, etc. In this paper, the rule-based expert system is used to control the steam pressure. Expert control is an important branch of intelligent control. It introduces the ideas and methods of expert system into control system and its engineering application. Expert system is an important way to make the research of artificial intelligence move from the theoretical research oriented to the basic technology and methods to the specific research to solve the practical problems [3]. As far as its essence is concerned, expert control is the sum of all kinds of knowledge based on the control object and control law, which should be used in an intelligent way to optimize and apply the system as practical as possible. It reflects many important features and functions of intelligent control [4]. The expert control system based on PLC realizes the real-time monitoring and controlling of steam pressure, which can meet the functional requirements.

## 3 Hardware Design

PLC expert control system is mainly composed of PLC, pressure sensor and touch screen. Its control core is Siemens S7-200 smart. The hardware is mainly composed of power supply, CPU SR30, EM AE04 four-way analog input module, Siemens Smart Line700 IE touch screen, communication module, etc.

Siemens S7-200 smart PLC is a small PLC, but its function is very powerful. It integrates microprocessor, integrated power supply, digital I/O, communication interface and other components in a compact package. Because of its strong function, low price, high reliability and convenient use, it is widely used in all walks of life [5].

CJBP series universal pressure sensors and transmitters are made of high precision and high stability pressure sensor components of famous international companies, which ensure the technical indicators of the sensors. The series of

products have a variety of shapes, interface forms and lead ways, which can meet the actual needs to the maximum extent. Besides, they are suitable for the matching use of various measurement and control equipment [6]. The pressure sensor collects the primary and secondary side pressure data and transmits them to PLC through analog input module.

The touch screen is widely used in the control system which requires strong human-computer interface function. The operator can set relevant system parameters and control parameters through it, monitor various state information of the system, send control commands with it, and realize data and information interaction with PLC [7]. Considering the functional requirements, display effect, convenient downloading procedure and economic cost, the 7-in. touch screen Smart Line700 IE with 65536 colors and resolution of 800 \* 480 is selected.

The control system mainly completes the functions of analog signal acquisition, touch screen communication, expert system control logic analysis, digital signal output, etc. The control system collects and transmits the steam pressure of the primary side and the secondary side to the PLC in real time, controls the valve opening degree through the control data generated by the expert system processing, and simultaneously displays the real-time pressure data on the touch screen. The control system is provided with manual control mode and automatic control mode, which can be selected through the touch screen. In the manual mode, the pressure can be controlled to rise or decrease on the touch screen.

## 4 Software Design

The software program of the control system is mainly composed of PLC program and touch screen program of upper computer.

### 4.1 PLC Program

This system uses the Siemens programming software STEP7 to program the software. The system software design mainly has two modules, including the steam pressure control module and the alarm module. The core function of the control system is to obtain the control parameters to control the motor and further control the opening of the control valve.

The data processing is based on the theory of model-based expert system. The control causal model is constructed and the heuristic knowledge base is established according to the expert experience. The heuristic rules and causal model are combined and the decision when to use these meta-rules for decision control is determined through rule reasoning. The characteristic of model-based expert system is to use heuristic rules and causal model to express heuristic knowledge in the form of rules, control knowledge in the form of meta rules, and decide when to use which rules for rule reasoning [8].

The system sets the corresponding expected value and compares the observed value with the expected value. In this way, rules can be established. The production rule of expert system uses IF P THEN Q structure to express the causal

relationship between various things and knowledge, which means that if the premise P is satisfied, then Q can be derived (or corresponding actions can be performed). The structure is very similar to the normal thinking of human beings and the efficiency of empirical correlation processing is high [9]. Rules for building knowledge base by heuristic knowledge are as follows:

IF secondary side pressure is more than 0.43 MPa THEN the pressure will give a high alarm

IF secondary side pressure is less than 0.20 MPa THEN the pressure will give a low alarm

IF system is in automatic state and pressure difference is big THEN control valve opening changes fast

IF system is in automatic state and pressure difference is small THEN control valve opening changes slowly

IF system is in automatic state and pressure difference is within the dead band THEN control valve remains unchanged

Some PLC programs are shown in Fig. 1.

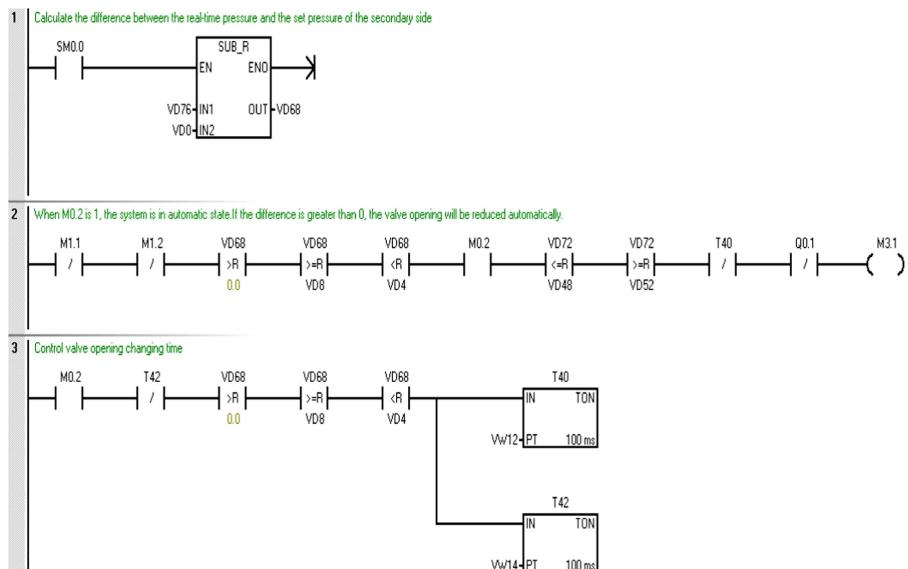


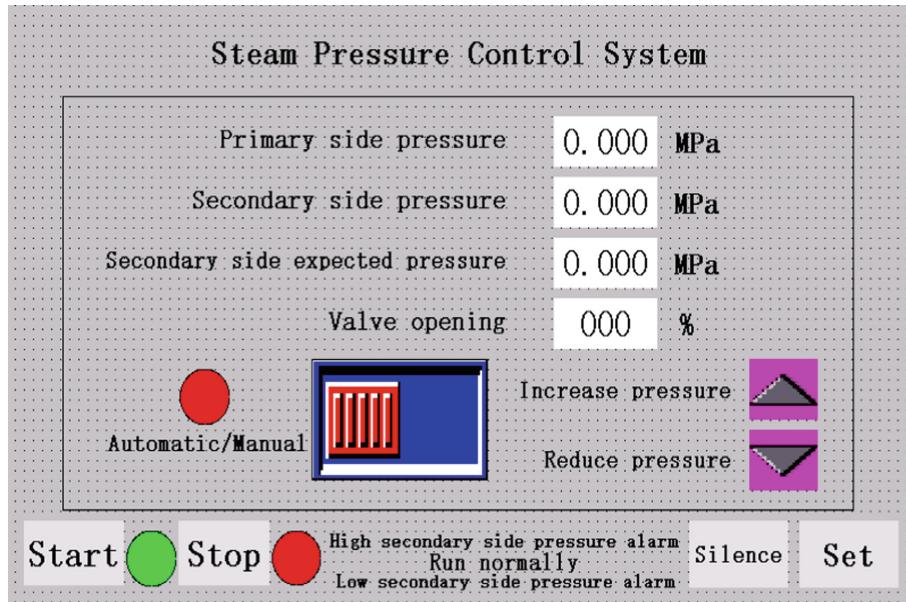
Fig. 1. Partial program diagram of PLC

Matching the obtained steam pressure difference value with the difference value given by expert experience, determining the stage of the difference value, and adjusting according to the control decision of valve opening in this stage. According to the difference of pressure, the control strategy is determined by the reasoning of steam pressure decision base: fast rise/drop strategy, slow rise/drop strategy and dead zone uncontrolled strategy. After the control strategy is determined, the motor is controlled to adjust the valve opening.

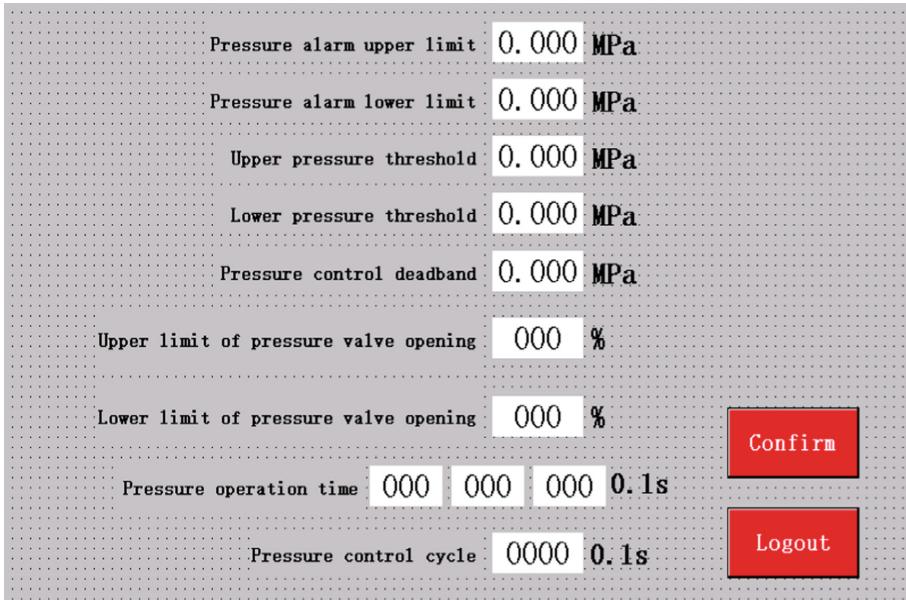
## 4.2 Interface Design of Upper Computer

Siemens human-machine interface can work normally only when it is configured. It is realized by WinCC flexible smart software. The software provides a lot of picture libraries and objects for user configuration programming. Using the relevant functions of libraries and objects reasonably can quickly get many different beautiful pictures [10].

The touch screen of this system mainly has two interfaces, the main interface and the setting interface. The main interface is as shown in Fig. 2, which shows the pressure value of primary side and secondary side, the feedback value of valve opening, the expected pressure value of secondary side, the selection of manual and automatic state, the buttons for manually raising and lowering pressure, the buttons for starting and stopping system (after operation, the corresponding display light is green, and it is red when not operated), alarm information, buzzer silencing button, and setting button. The setting interface is as shown in Fig. 3, mainly used to set system parameters, including upper and lower limit of pressure alarm, upper and lower limit of pressure valve opening, upper and lower limit of pressure threshold (threshold is the difference between the real-time pressure value of secondary side and the set target value), control dead zone, operation time and control cycle.



**Fig. 2.** Main interface of steam pressure regulating system



**Fig. 3.** Parameter setting interface

## 5 Conclusion

The steam pressure regulating system designed in this paper is a control system based on S7-200 smart PLC, which is based on expert system theory. The field operation results show that the system can adjust the pressure in time, meet the control requirements, solve the problem that the motor will be burned in the traditional PID control and achieves good results in practical application. The system is still improving to meet the changing needs of many factors.

## References

1. Kouvakas, N.D., Kouboulis, F.N., Paraskevopoulos, P.N.: Modeling and control of a neutral time delay test case central heating system (2007)
2. Zhang, H., Shang, S.Q., Yang, R.B.: Design and implementation of heat exchange station control system, vol. 170, pp. 2666–2669 (2012)
3. Du, W.L., Tian, X.L.: An automatic evaluation platform for feature matching algorithms based on an orbital optical pushbroom stereo imaging system. *IEEE Geosci. Remote Sens. Lett.* **15**(99), 1–5 (2018)
4. Vásquez, R.P., Alfonso, A.L.A., López-Segura, M.V., et al.: Expert system based on a fuzzy logic model for the analysis of the sustainable livestock production dynamic system. *Comput. Electron. Agri.* **161**, 104–120 (2018). S0168169918300280
5. Gelen, G., Uzam, M.: The synthesis and PLC implementation of hybrid modular supervisors for real time control of an experimental manufacturing system. *J. Manuf. Syst.* **33**(4), 8–15 (2014)

6. Matthews, C., Pennecch, F., Elster, C., et al.: Mathematical modelling to support traceable dynamic calibration of pressure sensors. *Metrologia* **51**(3), 326–338 (2014)
7. Lee, C., Kim, D.Y., Park, J.K., et al.: A characterization method for projected capacitive touch screen panel using 3-port impedance measurement technique. In: *IEEE Sensors*, pp. 1–3 (2015)
8. Luger, G.E.: Industry intelligence and the design of expert system. *Redwood* **1**, 135 (2014)
9. Richer, M.H.: An evaluation of expert system development tools. *Expert Syst.* **3**(3), 166–183 (2012)
10. Zhou, G., Zhu, Z., Chen, G., et al.: Technique of WinCC long-distance accessing exterior SQL server database. In: *First International Workshop on Education Technology and Computer Science*, pp. 153–155. IEEE Computer Society (2009)



# Event-Triggered Anti-disturbance Tracking Control for Systems with Exogenous Disturbances

Xiaoli Zhang, Xiang Gu, Yang Yi<sup>(✉)</sup>, and Tianping Zhang

College of Information Engineering, Yangzhou University, Yangzhou, China  
yiyangcontrol@163.com

**Abstract.** A novel event-triggered anti-disturbance PI control strategy for nonlinear systems is studied in this essay. Firstly, the general method of event-triggered control is extended, and the event-triggered mechanism is leaded into the control system. Secondly, the designed observer is used to estimate the external disturbances. Then, by combining the state feedback with disturbance estimation, a PI-type dynamic tracking controller is introduced to ensure that the output tracking error approaches zero. What's more, the Lyapunov function model is established, and the relevant theorem is proved. In addition, by solving the lower bound of the minimum trigger time interval to be greater than zero, it is guaranteed that the Zeno phenomenon in the event-triggered mechanism can be avoided. Finally, the A4D model is simulated, which proves the algorithm is effective.

**Keywords:** Event-triggered mechanism · Anti-disturbance control · Disturbance Observer (DO) · Tracking control

## 1 Introduction

The periodic sampling mechanism is adopted by the most control systems. In this strategy, the controller is updated at every sampling instant, which causes waste of resources [1]. Subsequently, an alternative method of saving resources emerged: the event-triggered mechanism. In the late 1990s, Kopetz and Astom first proposed the event-triggered mechanism [2,3]. In this mechanism, it sends the current object status only when the trigger condition is met, effectively reducing the data sending rate. Therefore, event-triggered policies can reduce unnecessary calculations and waste of resources. As a result, research on control systems based on event-triggered mechanism has attracted widespread attention. In [4] and [5], authors discussed event-triggered fault detection in networked control systems. The robot's event-triggered tracking control was studied in [6], and it was verified in wireless robot teleoperation that event-triggered control can achieve similar control performance to cycle-triggered control at a lower data transmission rate. Trigger control of a semi-Markovian jump nonlinear system

based on DO was proposed in [7], and an effective event trigger mechanism controlled the interaction of information.

So far, most event-triggered control methods, such as [4–7], consider the use of proportional controllers and focus on the stability of the event-triggered control loop. However, if the event-triggered control is to be applied in practice, it is important to track constant reference and set values of certain signals. In [8], Lehmann et al. analyzed the stability of a first-order system with actuator saturation and implemented the tracking of reference input for the method of applying PI controller combined with event-triggering. In [9], Sven Reimann et al. discussed event-triggered PI control based on a discrete-time model and implemented output tracking for constant reference. In addition, [10] mentioned that for the purpose of avoiding the occurrence of the Zeno phenomenon in the event-triggered mechanism, the time interval between two consecutive triggering moments needs to be calculated.

As we all know, unknown interference and uncertainties are common in various practical systems. For example, control systems with saturated inputs [11], non-Gaussian random distribution systems with mismatched disturbances [12], and control systems in robots [13]. In order to have better performance when encountering those unknown interferences, how to suppress or attenuate unknown interferences becomes a key issue. In these decades of development, there are also many excellent results, such as the famous  $H\infty$  or  $L_2/L\infty$  robust control methods [14,15] and output regulation theory [16]. Although these control methods have proven to be effective, they have all adopted feedback control methods. The solution based on disturbance observer control is to design an observer to estimate unknown disturbances, and use feed-forward compensator and conventional control law to suppress the disturbance. Nowadays, the disturbance-observer-based control (DOBC) method is successfully applied in classical nonlinear systems such as Markov transition systems [17], stochastic distributed systems [18], and multi-agent systems [19].

Motivated by the above observations, a novel event-triggered PI control strategy for nonlinear systems based on disturbance observer is proposed. Firstly, an extended event-triggered mechanism is leaded into the control system to avoid unnecessary waste of resources. Secondly, an disturbance observer is designed to estimate interference. A PI-type dynamic tracking controller is designed to track the output by combining the interference estimates and the state feedback under the event-triggered mechanism. Finally, the simulation's results show the rationality of the proposed event-triggered controller. In this article, the common event-triggered control is extended to event-triggered PI control, which can effectively avoid resource waste and achieve output tracking. At the same time, design an interference observer to effectively suppress the influence of external interference on the system.

Nomenclature. If not stated, all involved vectors or matrices are supposed to have suited dimensions. For any matrices  $Y$ , define the mark  $\text{sym}$  as  $\text{sym}(Y) = Y + Y^T$ .

## 2 System Description

The following continuous-time nonlinear system with exogenous disturbance is considered:

$$\begin{cases} \dot{\bar{f}}(t) = Af(t) + B[u(t) + r(t)] \\ z(t) = Cf(t) \end{cases} \quad (1)$$

where  $A$ ,  $B$  and  $C$  are system matrices with appropriate dimension.  $f(t) \in R^n$  is the state vector,  $u(t) \in R^m$  represents the controlled input vector,  $r(t) \in R^m$  stands for the external disturbance, and  $z(t) \in R^p$  represents the output vector.

Then the unknown disturbance  $r(t)$  is designed as

$$\begin{cases} \dot{h}(t) = Hh(t) \\ r(t) = Vh(t) \end{cases} \quad (2)$$

*Assumption 1:* The pair  $(A, B)$  is controllable, and the pair  $(H, BV)$  is observable.

In order to achieve good dynamic tracking performance, we extend the system state as

$$\bar{f}(t) := \left[ f^T(t) \int_0^t e^T(\tau) d\tau \right]^T \quad (3)$$

where the tracking error  $e(t)$  is defined as  $e(t) := z(t) - z_d$ ,  $z_d$  is the designed output. Based on (1) and (2), the tracking augmented scheme is deduced as

$$\begin{cases} \dot{\bar{f}}(t) = \bar{A}\bar{f}(t) + \bar{B}[u(t) + r(t)] + \bar{F}z_d \\ z(t) = \bar{C}\bar{f}(t) \end{cases} \quad (4)$$

$$\text{where } \bar{A} = \begin{bmatrix} A & 0 \\ C & 0 \end{bmatrix}, \bar{B} = \begin{bmatrix} B \\ 0 \end{bmatrix}, \bar{C} = \begin{bmatrix} C^T \\ 0 \end{bmatrix}^T, \bar{F} = \begin{bmatrix} 0 \\ -I \end{bmatrix}.$$

## 3 Event-Triggered Scheme

In this section, the event detector monitors whether an event occurs. Based on threshold conditions, the event detector can decide whether to update the controller time. Threshold condition is given as:

$$\bar{e}_k^T(t)\bar{\Phi}\bar{e}_k(t) \leq \bar{\delta}^2 \bar{f}^T(t)\bar{\Phi}\bar{f}(t) \quad (5)$$

where

$$\bar{e}_k(t) = \begin{bmatrix} e_{1k}(t) \\ e_{2k}(t) \end{bmatrix} = \begin{bmatrix} f(t) - f(t_k) \\ \int_{t_k}^t e(\tau) d\tau \end{bmatrix} = \bar{f}(t) - \bar{f}(t_k), \bar{\Phi} = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix}, \bar{\delta} = \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix}.$$

Then we can have

$$\dot{\bar{e}}_k(t) = \dot{\bar{f}}(t) \quad (6)$$

where  $\delta_i \in [0, 1], i = 1, 2$  are given vectors in order to set the designed threshold.  $\Phi_i > 0, i = 1, 2$  are designed event-triggered matrices.

It can be seen that we define the state of the trigger controller and its integral as “Event 1” and “Event 2”, respectively. As long as one of the events does not meet the threshold condition, the detector will be triggered. Then, the detector will send the updated data to the controller.  $\bar{f}(t)$  is the current sampled state, and  $\bar{f}(t_k)$  is the latest data which have been sent to the controller. Then the next trigger moment can be represented by the following formula:

$$t_{k+1} = \inf\{t > t_k | \|\bar{\Phi}^{1/2}[\bar{f}(t) - \bar{f}(t_k)]\|_2 > \|\bar{\delta}^{1/2}\bar{\Phi}^{1/2}\bar{f}(t)\|_2\}, \quad (7)$$

where  $t \in [t_k, t_{k+1})$ .

## 4 Anti-disturbance Event-Triggered Composite Controller Design Method

For purpose of estimating unknown disturbances, the disturbance observer is found as:

$$\begin{cases} \hat{r}(t) = V\hat{h}(t) \\ \hat{h}(t) = \beta(t) - L\bar{f}(t) \\ \dot{\beta}(t) = (H + L\bar{B}V)(\beta(t) - L\bar{f}(t)) - L(-\bar{A}\bar{f}(t) - \bar{B}u(t) - \bar{F}z_d) \end{cases} \quad (8)$$

Under the event-triggered scheme, the controlled input is found as:

$$u(t) = -\hat{r}(t) + K\bar{f}(t_k), K = [K_P \ K_I] \quad (9)$$

where  $\hat{r}(t)$  is the estimate of  $r(t)$ ,  $K$  and  $L$  are controller gain and observer gain, respectively.  $\nu(t)$  is designed as a auxiliary variable. In the event-triggered scheme, the controller (9) is calculated only at the instant of sampling.

Based on (4) and (9), the augmented system is deduced as

$$\dot{\bar{f}}(t) = (\bar{A} + \bar{B}K)\bar{f}(t) - \bar{B}K\bar{e}_k(t) - \bar{B}Ve_h(t) + \bar{F}z_d \quad (10)$$

By defining  $e_h(t) := \hat{h}(t) - h(t)$ , and from (2), (4) and (9), we can obtain

$$\dot{e}_h(t) = (H + L\bar{B}V)e_h(t) \quad (11)$$

Combining with (6), (10) and (11), the composite system can be obtained:

$$\dot{\Lambda}(t) = \begin{bmatrix} \bar{A} + \bar{B}K & -\bar{B}V & -\bar{B}K \\ 0 & H + L\bar{B}V & 0 \\ \bar{A} + \bar{B}K & -\bar{B}V & -\bar{B}K \end{bmatrix} \Lambda(t) + \begin{bmatrix} \bar{F} \\ 0 \\ \bar{F} \end{bmatrix} z_d \quad (12)$$

$$\text{where } \Lambda(t) = \begin{bmatrix} \bar{f}(t) \\ e_h(t) \\ e_k(t) \end{bmatrix}.$$

## 5 Related Theorem Proof

Theorem 1 solves the observer gain and the controller gain by designing an algorithm. In order to avoid the Zeno phenomenon, that is, an event that cannot occur an infinite number of times is triggered in a limited time, Theorem 2 is given, it indicates that the minimum trigger time is a number greater than 0 and not sufficiently small.

**Theorem 1.** *Based on the system (4) and the disturbance estimation error model (11), if there are matrices  $Q = P_1^{-1} > 0$  and  $X, Y, P_2 > 0$  satisfying the following inequality*

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \bar{F} \\ * & \sigma_{22} & 0 & 0 \\ * & * & \sigma_{33} & 0 \\ * & * & * & -\mu^2 I \end{pmatrix} < 0 \quad (13)$$

with

$$\begin{cases} \sigma_{11} = \text{sym}\{\bar{A}Q + \bar{B}Y\} + \delta^2\tilde{\Phi} + Q \\ \sigma_{12} = -\bar{B}V \\ \sigma_{13} = -\bar{B}Y \\ \sigma_{22} = \text{sym}\{P_2W + X\bar{B}V\} \\ \sigma_{33} = -\tilde{\Phi} \end{cases}$$

where  $\mu > 0$  is a known parameter, and  $\tilde{\Phi} = Q\bar{\Phi}Q$ , then the enhanced system (12) under the compound controller (9) is stochastically stable. What's more,  $\lim_{t \rightarrow \infty} z(t) = z_d$  holds. The gains can be given by  $K = YQ^{-1}$ ,  $L = P_2^{-1}X$ .

*Proof.* Lyapunov functions are designed as

$$V_1 = \bar{f}^T(t)P_1\bar{f}(t) \quad (14)$$

and

$$V_2 = e_h^T(t)P_2e_h(t) \quad (15)$$

According to (10) and (11), it can be deduced as

$$\begin{aligned} \dot{V}_1 + \dot{V}_2 &\leq \bar{f}^T(t)(P_1\bar{A} + \bar{A}^TP_1)\bar{f}(t) + \text{sym}\{\bar{f}^T(t)P_1\bar{B}K(\bar{f}(t) - \bar{e}_k(t)) \\ &\quad - \bar{f}^T(t)P_1\bar{B}Ve_h(t) + \bar{F}^Tz_dP_1\bar{f}(t)\} + e_h^T(t)\text{sym}\{P_2H \\ &\quad + P_2L\bar{B}V\}e_h(t) + \bar{\delta}^2\bar{f}^T\bar{\Phi}\bar{f} - \bar{e}_k^T\bar{\Phi}\bar{e}_k^T \\ &\leq \Lambda^T(t) \begin{bmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ * & \Omega_{22} & 0 \\ * & * & \Omega_{33} \end{bmatrix} \Lambda(t) + \mu^2z_d^2 \end{aligned} \quad (16)$$

where

$$\begin{cases} \Omega_{11} = \text{sym}\{P_1\bar{A} + P_1\bar{B}K\} + \bar{\delta}^2\Phi + \mu^{-2}P_1\bar{F}\bar{F}^TP_1 \\ \Omega_{12} = -P_1\bar{B}V \\ \Omega_{13} = -P_1\bar{B}K \\ \Omega_{22} = \text{sym}\{P_2(W + L\bar{B}V)\} \\ \Omega_{33} = -\bar{\Phi} \end{cases}$$

Based on Schur complement formula, we can have

$$\dot{V}_1 + \dot{V}_2 \leq -\varsigma\|\Lambda(t)\|^2 + \mu^2 z_d^2 \quad (17)$$

where  $\varsigma > 0$  is a appropriate constant. We can get  $\dot{V}_1 + \dot{V}_2 < 0$  when  $\|\Lambda(t)\|^2 > \varsigma^{-1}\mu^2 z_d^2$ . Aware of that for any  $\bar{f}(t)$ ,  $e_h(t)$  and  $\bar{e}_k(t)$ , we have

$$\Lambda^T(t)\Lambda(t) \leq \max\{\Lambda^T(0)\Lambda(0), \varsigma^{-1}\mu^2 z_d^2\} \quad (18)$$

where  $\Lambda(0)$  is the initial value of  $\Lambda(t)$ . It means that the state  $\Lambda(t)$  can be guaranteed to converge into the designed compact set  $\Omega_{\Lambda(t)}$ .

Aware of  $\int_0^t e(\tau)d\tau$  is one of variables in  $\Lambda(t)$ , we can claim that  $\int_0^t e(\tau)d\tau$  must also fall into the set  $\Omega_{\Lambda(t)}$  when  $t \rightarrow \infty$ . As long as the limitation of  $\int_0^t e(\tau)d\tau$  exists, the dynamical tracking performance  $\lim_{t \rightarrow \infty} z(t) = z_d$  is easy to be derived by using Babalat lemma. On the contrary, if the limit of  $\int_0^t e(\tau)d\tau$  does not exist, it is easy to judge from (18) that the tracking error  $e(t)$  will chatter at near zero. In general, the favourable tracking performance can be pledged.

**Theorem 2.** For system (4), under the event-trigger condition (5), the lower bound of the minimum trigger time interval is:

$$\tilde{T} = \min\{T_k\} = \frac{1}{a} \ln\left(\frac{a}{b}\Psi(t) + 1\right) > 0 \quad (19)$$

where  $a = |\lambda_{\max}(\bar{A})|$ ,  $b = (a + \|\bar{B}\|\|K\|)\|\bar{f}(t_k)\| + \|\bar{B}\|\|V\|\|e_h(t)\| + \|\bar{F}\|\|z_d\|$ ,  $\Psi = \|\bar{\delta}\bar{f}(t)\|$ .

*Proof.* Based on  $\bar{e}_k(t) = \bar{f}(t) - \bar{f}(t_k)$ , we can have  $\dot{\bar{e}}_k(t) = \bar{A}\bar{f}(t) + \bar{B}(u(t) + r(t)) + \bar{F}z_d$ , then when  $t \in [t_k, t_{k+1})$ ,

$$\begin{aligned} \frac{d}{dt}\|\bar{e}_k(t)\| &\leq \|\dot{\bar{e}}_k(t)\| \\ &\leq |\lambda_{\max}(\bar{A})|\|\bar{f}(t)\| + \|\bar{B}\|\|u(t) + r(t)\| + \|\bar{F}\|\|z_d\| \\ &\leq |\lambda_{\max}(\bar{A})|\|\bar{f}(t)\| + \|\bar{B}\|\|K\|\|\bar{f}(t_k)\| + \|\bar{B}\|\|V\|\|e_h(t)\| + \|\bar{F}\|\|z_d\| \\ &\leq |\lambda_{\max}(\bar{A})|\|\bar{e}_k(t)\| + (|\lambda_{\max}(\bar{A})| + \|\bar{B}\|\|K\|)\|\bar{f}(t_k)\| \\ &\quad + \|\bar{B}\|\|V\|\|e_h(t)\| + \|\bar{F}\|\|z_d\| \end{aligned} \quad (20)$$

By solving  $\frac{d}{dt}\|\bar{e}_k(t)\| = a\|\bar{e}_k(t)\| + b$ , we can get the upper bound of  $\|\bar{e}_k(t)\|$ , where  $a = |\lambda_{\max}(\bar{A})|$ ,  $b = (a + \|\bar{B}\|\|K\|)\|\bar{f}(t_k)\| + \|\bar{B}\|\|V\|\|e_h(t)\| + \|\bar{F}\|\|z_d\|$ .

It can be known from the event-triggered scheme that  $\|\bar{e}_k(t)\| = 0$ , when  $\bar{f}(t) = \bar{f}(t_k)$ . Then the solution of the above equation can be found as  $\|\bar{e}_k(t)\| = \frac{a}{b}(e^{a(t-t_k)} - 1)$ .

Define  $\Psi(t) = \|\bar{\delta}\bar{x}(t)\|$ ,  $\Psi(t) > 0$  can be known when  $\|\bar{\delta}\bar{x}(t)\| \neq 0$ . The next trigger time will occur after  $\|\bar{e}_k(t)\| = \Psi(t)$ .

So the minimum trigger interval is  $\tilde{T} = \frac{1}{a} \ln(\frac{a}{b}\Psi(t) + 1) > 0$ . In conclusion, the event-triggered system can avoid Zeno phenomenon.

## 6 Numerical Illustrations

To confirm the above algorithm, the longitudinal model of A4D aircraft with 15000 ft altitude and 0.9 Mach can be given by

$$\begin{cases} \dot{f}(t) = Af(t) + B[u(t) + r(t)] \\ z(t) = Cf(t) \end{cases}$$

Related parameters are given as

$$A = \begin{bmatrix} -0.0605 & 32.37 & 0 & 42 \\ -0.00014 & -1.475 & 1 & 0 \\ -0.0111 & -34.72 & -2.793 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ -0.1064 \\ -33.8 \\ 0 \end{bmatrix}, C = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}^T.$$

Next, in DO, we design the following parameters:

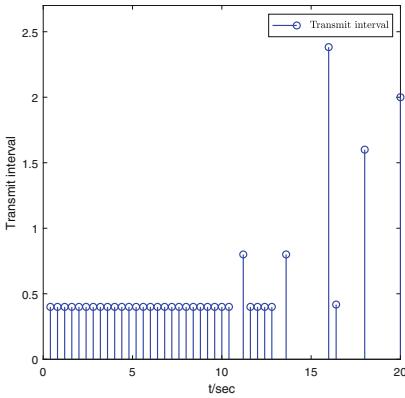
$$H = \begin{bmatrix} 0 & 6 \\ -6 & -0.4 \end{bmatrix}, V = [40 \ 0].$$

What's more, define  $\mu = 0.8$  and solve LMI (13), then the control gains  $K_P$ ,  $K_I$  observer gain  $L$  are solved as

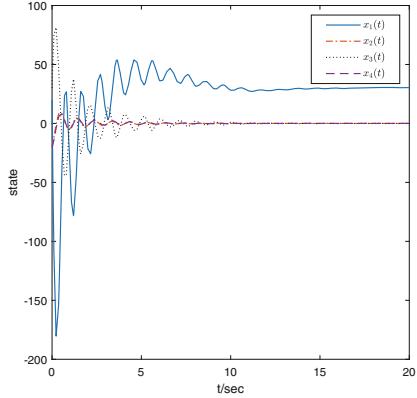
$$K_P = [0.0076 \ -0.4015 \ 0.0319 \ 0.5043], K_I = 0.0032,$$

$$L = 10^{-3} \begin{bmatrix} 0 & 0.0009 & 0.4058 & 0 & 0 \\ 0 & 0.0000 & 0.0018 & 0 & 0 \end{bmatrix}.$$

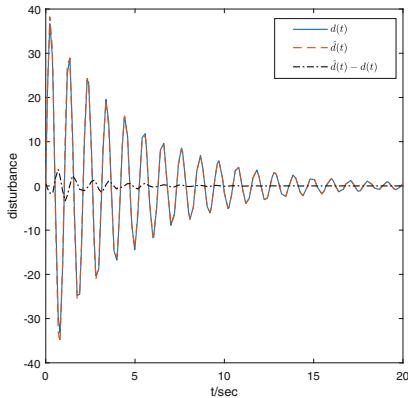
Assume that the initial states are  $f(0) = [20 \ -20 \ 30 \ -20]^T$ . In the event-triggered scheme, supposed that  $\Phi_1$ ,  $\Phi_2$  are identity matrix, and  $\delta_1^2 = \delta_2^2 = 0.0001$ . Besides, design the tracking objective is  $z_d = 30$ . In Fig. 1, it displays the trigger times and intervals. Besides, it can be found out that the system is stochastically stable in Fig. 2. Then  $d(t)$ ,  $\hat{d}(t)$  and  $\hat{d}(t) - d(t)$  are all displayed in Fig. 3. Obviously, the disturbance observer is effective. Figure 4 is the dynamics of system output. It reflects good dynamic tracking performance.



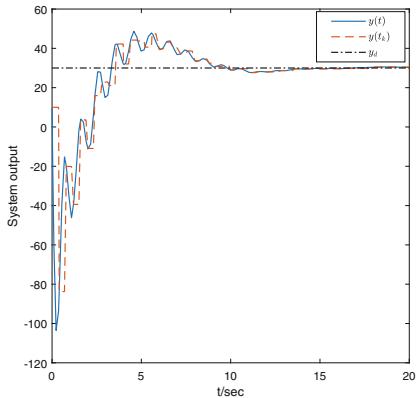
**Fig. 1.** The event-triggered release times and intervals.



**Fig. 2.** Trajectory of the system states.



**Fig. 3.** Disturbance estimation error.



**Fig. 4.** Trajectory of the system output.

## 7 Conclusion

In this essay, we have discussed a novel event-triggered PI control strategy for nonlinear systems based on disturbance observers. Controller calculation is saved by efficacious event-triggered scheme. The introduction of the PI-type dynamic tracking controller ensures that the output tracking error tends to zero. Through a series of proofs, the proposed event-triggered PI control strategy is clearly depicted. Finally, the rationality of the algorithm is verified by simulation.

**Acknowledgment.** This work was supported in part by NSFC under Grants 61803331 and 61973266, the Nature Science Foundation of Jiangsu Province under Grant BK20170515.

## References

1. Pawłowski, A., Cervin, A., Guzman, J.L., Berenguel, M.: Generalized predictive control with actuator Deadband for event-based approaches. *IEEE Trans. Ind. Inf.* **10**(1), 523–537 (2014). <https://doi.org/10.1109/TII.2013.2270570>
2. Kopetz, H.: Should responsive systems be event-triggered or time-triggered? *IEICE Trans. Inf. Syst.* **76**(11), 1325–1332 (1993)
3. Åström, K., Bernhardsson, B.: Comparison of periodic and event based sampling for first-order stochastic systems. In: Proceedings of the 14th IFAC World Congress, vol. 11, pp. 301–306 (1999)
4. Wang, Y.L., Shi, P., Lim, C.C., Liu, Y.: Event-triggered fault detection filter design for a continuous-time networked control system. *IEEE Trans. Cybern.* **46**(12), 3414–3426 (2016). <https://doi.org/10.1109/TCYB.2015.2507177>
5. Jin, Z.W., Hu, Y.Y., Li, C., Sun, C.G.: Event-triggered fault detection and diagnosis for networked systems with sensor and actuator faults. *IEEE Access* **7**, 95857–95866 (2019). <https://doi.org/10.1109/ACCESS.2019.2928473>
6. Postoyan, R., Bragagnolo, M.C., Galbrun, E., Daafouz, J., Nešić, D., Castelan, E.B.: Event-triggered tracking control of unicycle mobile robots. *Automatica* **52**, 302–308 (2015). <https://doi.org/10.1016/j.automatica.2014.12.009>
7. Yao, X.M., Lian, Y., Park, J.H.: Disturbance-observer-based event-triggered control for semi-Markovian jump nonlinear systems. *Appl. Math. Comput.* **363** (2019). <https://doi.org/10.1016/j.amc.2019.124597>
8. Lehmann, D., Johansson, K.H.: Event-triggered PI control subject to actuator saturation. In: Proceedings of the 18th IFAC World Congress, pp. 3262–3267 (2011)
9. Reimann, S., Van, D.H., Al-Areqi, S., Liu, S.: Stability analysis and PI control synthesis under event-triggered communication. In: Proceedings of 2015 European Control Conference (2015). <https://doi.org/10.1109/ECC.2015.7330862>
10. Ma, T.S., Cao, W.J., Lin, Y.J., Zhang, J.H.: Observer-based event-triggered control with disturbance compensation. In: Proceedings of the 33rd Chinese Control Conference (2014). <https://doi.org/10.1109/ChiCC.2014.6895897>
11. Shao, L.R., Yi, Y., Liu, B., Zheng, W.X.: Neural network modeling-based anti-disturbance tracking control for complex systems with input saturation. In: Proceedings of the 37th Chinese Control Conference (2018). <https://doi.org/10.23919/ChiCC.2018.8484060>
12. Yi, Y., Guo, L., Wang, H.: Anti-disturbance iterative learning tracking control for general non-Gaussian stochastic systems. In: Proceedings of the 11th World Congress on Intelligent Control and Automation (2014). <https://doi.org/10.1109/WCICA.2014.7052735>
13. Chen, W.H., Ballanceand, D.J., Gawthrop, P.J.: A nonlinear disturbance observer for robotic manipulators. *IEEE Trans. Ind. Electron.* **47**(4), 932–938 (2000). <https://doi.org/10.1109/41.857974>
14. Marino, R., Tomei, P.: Nonlinear Control Design: Geometric, Adaptive and Robust. Prentice Hall, Upper Saddle River (1996)
15. Li, Y.K., Sun, H.B., Zong, G.D.: Disturbance-observer-based-control and  $L_2/L_\infty$  resilient control for Markovian jump nonlinear systems with multiple disturbances and its application to single robot arm system. *IET Control Theory Appl.* **10**(2), 226–233 (2016). <https://doi.org/10.1049/iet-cta.2015.0430>
16. Xiong, S., Xie, H., Song, K., Zhang, G.H.: A speed tracking method for autonomous driving via ADRC with extended state observer. *Appl. Sci. Basel.* **9**(16) (2019). <https://doi.org/10.3390/app9163339>

17. Yao, X.M., Wu, L.G., Guo, L.: Disturbance-observer-based fault tolerant control of high-speed trains: a Markovian jump system model approach. *IEEE Trans. Syst. Man Cybern. Syst.* **50**(4), 1476–1485 (2020). <https://doi.org/10.1109/TSMC.2018.2866618>
18. Yi, Y., Zheng, W.X., Sun, C.Y., Guo, L.: DOB fuzzy controller design for non-Gaussian stochastic distribution systems using two-step fuzzy identification. *IEEE Trans. Fuzzy Syst.* **24**(2), 401–418 (2016). <https://doi.org/10.1109/TFUZZ.2015.2459755>
19. Wang, X.Y., Li, S.H., Yu, X.H., Yang, J.: Distributed active anti-disturbance consensus for leader-follower higher-order multi-agent systems with mismatched disturbances. *IEEE Trans. Autom. Control* **62**(11), 5795–5801 (2017). <https://doi.org/10.1109/TAC.2016.2638966>



# Parallel Label Consistent KSVD-Stacked Autoencoder for Industrial Process Fault Diagnosis

Hongpeng Yin<sup>(✉)</sup>, Jiaxin Guo, Guobo Liao, and Yi Chai

Chongqing University, Chongqing 400044, China  
yinhongpeng@gmail.com

**Abstract.** Both linear and nonlinear relationships are two typical characteristics among industrial process variables, and diagnosing a process with such complicated correlations among variables is indispensable. However, individual dictionary learning or autoencoder based method is hard to extract these complicated correlations well. The parallel label consistent KSVD-stacked autoencoder (P-LCKSVD-SAE) model is proposed to integrate the linear and nonlinear feature extraction for effective industrial process fault diagnosis. First, LCKSVD and SAE methods are applied to extract the linear and nonlinear features from industrial process. Second, a matrix fusion algorithm is proposed for combining these intrinsic fault features. Then, the SVM classifier is utilized to verify the effectiveness of the proposed model for fault classification. Finally, several case studies on Tennessee Eastman process demonstrate that the proposed P-LCKSVD-SAE fault diagnosis scheme is better than the conventional LC-KSVD as well as the SAE methods at performing industrial process fault diagnosis.

**Keywords:** Nonlinear industrial process · Fault diagnosis · Label consistent KSVD · Stacked autoencoder

## 1 Introduction

Industrial process monitoring technologies that focus on fault diagnosis play an important role in guaranteeing industrial plant safety, reflecting the product quality timely and decreasing the production cost. Inspired by the powerful ability to extract the underlying intrinsic information from the process data, the data-driven methods have attracted intensively attentions in the field of fault diagnosis over the past two decades. Representative methods include Principal Component Analysis (PCA), Independent Component Analysis (ICA), Slow Feature Analysis methods, Bayesian Analysis methods [1] and so on. Although the numerous successful applications have been reported, those methods perform poorly in dealing with nonlinear processes because they characterize only the linear correlation variables and does not explore the nonlinear relationships [2]. Since the complicated chemical reactions of industrial system, most modern

industrial processes always exist nonlinearity in raw data. For settling this problem, some nonlinear methods have been proposed. Adding the kernel function to the traditional linear models is the typical measures for mapping the nonlinear data into a high-dimensional space, such as Kernel Principle Component Analysis (KPCA), Kernel Independent Component Analysis (KICA) and so on. However, those methods rely heavily on a Gaussian assumption and need to pre-determine the weight value of kernel function, which may bring the uncertainty influence to the final consequences. Additionally, due to the single-layer of feature extraction, the shallow/just-one-layer Kernel-based models are insufficient to extract the intrinsic information of process data. Hence, it is desirable that a multi-layer feature extraction model is utilized to extract the intrinsic features sufficiently. The linear-nonlinear model have been used for overcoming the linear and nonlinear problems simultaneously in the past years like [2,3].

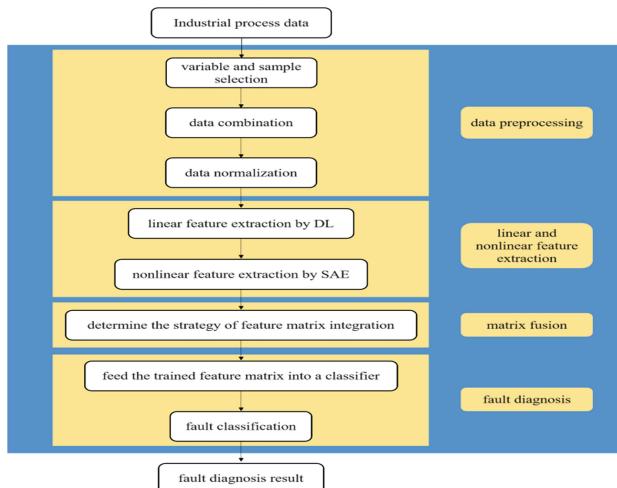
Although several researches have paid attentions to extract the linear and nonlinear features simultaneously, there is still the space for improving the performance. How to extract the intrinsic information from the industrial process data is still an open issue until now.

Motivated by the hybrid structure, the parallel label consistent KSVD-stacked autoencoder (P-LCKSVD-SAE) method is proposed to deal with the industrial data that contain linearly related and nonlinearly correlated features. The traditional multivariate statistical methods detect the fault based on the distance between the testing and training data that would ignore the structural features hidden in the raw process data. The main assumption of dictionary learning (DL) methods is that the data is sparse and can be represented by a series of basis, which is in accordance with the industrial process data [4]. Naturally, due to the powerful ability of extracting the underlying structure of the process data and the interpretability of representing the data, the DL based methods have attracted more attentions. In this paper, a discriminative DL method is applied for training the dictionary matrix with the label information. To develop a method for dealing with the nonlinear problem in industrial process, a neural network-based method is proposed, which can extract the nonlinear information from the raw process data. However, most existing supervised deep learning methods are not suitable since the most data is unlabeled in the industrial process, which restrict the development of supervised deep learning methods in industrial process monitoring. Inspired by the unsupervised deep learning methods [5], Autoencoder (AE) based method is applied in this paper. Finally, the two learned feature matrixes from Dictionary learning step and deep learning step are concentrated via a matrix fusion method and further be fed into a classifier.

The remainder of this paper is structured as follow: Sect. 2 details the proposed parallel label consistent KSVD-stacked autoencoder (P-LCKSVD-SAE) fault diagnosis model, meanwhile, linear and nonlinear formulations also illustrated. Section 3 demonstrates the example on the Tennessee Eastman(TE) industrial process to show performance comparisons with the state-of-art algorithms. Section 4 comes up with the conclusions.

## 2 P-LCKSVD-SAE for Industrial Process Fault Diagnosis

The proposed P-LCKSVD-SAE model and diagnosis scheme mainly consists of three steps, namely, feature extraction, matrix fusion as well as fault classification, as illustrated in Fig. 1. In the feature extraction step, dictionary learning (DL) and stacked autoencoder (SAE) undertake the extraction of linear and nonlinear features hidden in the raw industrial process; in the matrix fusion step, learned linear and nonlinear feature matrixes will be integrated into a single feature representation in a horizontal manner; in the fault classification step, a mature classifier model will be utilized for diagnosing the category to which the samples belong.



**Fig. 1.** The diagram of proposed P-LCKSVD-SAE fault diagnosis scheme

The training data can be preprocessed based on different targets by variable and sample selection, data combination and data normalization. In this paper, the input data is a mixture of linear, nonlinear, structural and high-dimensional information. How to extract these hidden features is a basic but significant problem for fault diagnosis. For the linear and structural feature, Label Consistent KSVD (LCKSVD) method tend to be applied. LCKSVD method combines the label consistency constraint on the process of DL and classifier training, which attempt to optimize the learned dictionary discriminatively. In this case, the objective function for learning  $D$ ,  $W$  (classifier matrix),  $A$  (linear transformation matrix) and sparse representation  $X$  jointly can be defined as:

$$\begin{aligned} < D, W, A, X > = \arg \min_{D, W, A, X} & ||Y - DX||_2^2 + \alpha ||Q - AX||_2^2 \\ & + \beta ||H - WX||_2^2 \quad s.t. \forall i, ||x_i||_0 \leq T \end{aligned} \quad (1)$$

The specific meanings of each matrix can be referred to [6].

Taking account into the complicated process of industrial production and chemical reactions, there are many structural and redundancy knowledge embedding in the raw industrial data. DL methods, as a kind of data-driven method, could extract structural features and corresponding label-consistency information hidden in the industrial process data efficiently. Recent works about feature extraction and fault diagnosis based on DL [1, 7] inspire us to exploring further. The application of Neural Networks in process fault diagnosis have received more and more attentions [8, 9]. Autoencoder model, based on the unsupervised learning running on neural networks, is used to extract features by the powerful ability of learning the nonlinear knowledges. The data flows from input layer to output layer should be similar as much as possible with the backpropagation (BP) training optimization. The squared error is usually calculated as the error signal to compare the output with target [8]. For the nonlinear and high-dimensional information, the stacked autoencoder, which get rid of the insufficient extraction of single autoencoder via connecting multiple autoencoders successively, tend to be utilized.

The main idea of stacked autoencoder (SAE) is to obtain the nonlinearly important features hidden in the raw industrial data, which trains multiple autoencoders continuously, and then connects the encoder layers of each autoencoder. The new networks should be fine-tuning via adding a classifier layer finally. The SAE with a classifier could implement the better representation of input data.

Given a SAE including  $s$  autoencoders, the encoding is formulated as:

$$h = f_{encoder_s}(f_{encoder_{s-1}}(\dots f_{encoder_1}(x))) \quad (2)$$

where  $f_{encoder_i}$  refers to an activation function of the  $i^{th}$  encoder.

The decoding is formulated as:

$$y = f_{decoder_s}(f_{decoder_{s-1}}(\dots f_{decoder_1}(h))) \quad (3)$$

where  $f_{decoder_i}$  refers to an activation function of the  $i^{th}$  decoder.

#### **Algorithm 1: Label Consistent K-SVD for Feature Extraction**

- (1) The fault type and magnitude are determined. The training data consist of sample data  $Y = [y_1 \dots y_N] \in R^{m \times N}$  and label matrix  $H = [h_1 \dots h_N] \in R^{n \times N}$ .
- (2) Initialize the parameters:  $D_0, A_0, W_0$  based on [6]. The  $D_0$  is initialized by several iterations of KSVD within each class, and then, fixed the  $D_0$ , compute the associated sparse codes  $X_0$  of training data  $Y$  via original KSVD algorithm.  $A_0$  and  $W_0$  initialized by Eq. (4) and Eq. (5).

$$A_0 = (X_0 X_0^T + \lambda_2 I)^{-1} X_0 Q \quad (4)$$

$$W_0 = (X_0 X_0^T + \lambda_1 I)^{-1} X_0 H^T \quad (5)$$

- (3) Jointly optimized the Eq. (1) to obtain the desired  $D$ , transfer parameters  $A$ , classifier parameter  $W$  and the representation matrix  $X$  of training data  $Y$ .
- (4) The faulty features matrix is  $X \in R^{K \times N}$ .

**Algorithm 2: Stacked Autoencoder for Feature Extraction**

- (1) The fault dimension and magnitude of hidden layer are determined.
- (2) The training data is  $Y = [y_1 \dots y_N] \in R^{m \times N}$ , raw dimension is  $m$ , the stacked autoencoder network defined as  $[m, F_1, F_2]$ .
- (3) Establish a two-layer stacked autoencoder network. The first autoencoder is trained by the raw dimension  $m$ , learned representation  $H_1 \in R^{N \times F_1}$  of the first autoencoder will be the input data of the second autoencoder.
- (4) Connect the encoder layers of two autoencoders and fine-tuning to obtain the appropriate parameters ( $\theta = (W_{encoder} b_{encoder} W_{decoder} b_{decoder})$ ) by adding a classifier layer.
- (5) The hidden layer of the second layer becomes the new feature matrix in nonlinear process, the faulty feature matrix is  $H_2 \in R^{N \times F_2}$ .

**Algorithm 3: Matrix Fusion for Combining the Linear and Nonlinear Features**

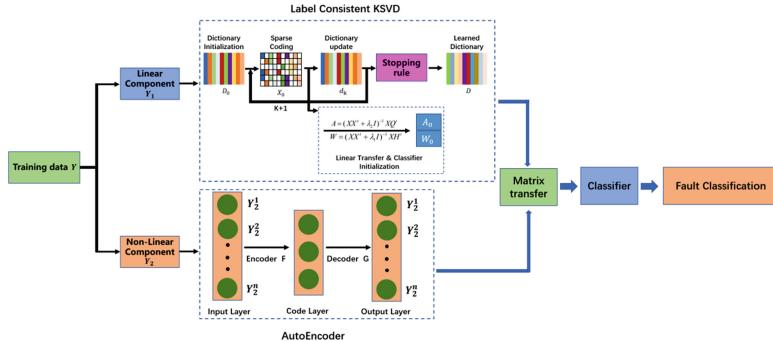
- (1) Get the trained feature matrix:  $X$  and  $H_2$ , the  $X$  contains the linear and structural information, the  $H_2$  contains the nonlinear information hidden in the industrial process.
- (2) Compute the fusion matrix  $M = [X^T, H_2]$ ,  $M \in R^{N \times (K+F_2)}$ .
- (3) Feed the upgraded feature matrix into a classifier, such as svm, softmax and so on.
- (4) Evaluate the performance of the model by the fault diagnosis rate (FDR).  $FDR_i$  represents the fault diagnosis rate of class  $i$ .  $p_i$  is the quantity that predicted labels are equal to actual labels within the class  $i$ .

$$FDR_i = \frac{p_i}{\text{total number of type } i \text{ samples}} \quad (6)$$

As shown in Fig. 2, the general P-LCKSVD-SAE fault diagnosis model is demonstrated. In the complicated industrial process, this model tend to apply the label consistent KSVD algorithm to extract the linear and structural features and utilize the stacked autoencoder to learn the nonlinear features. Then, a new feature matrix integrates linear and nonlinear features is fed into a classifier to implement the fault diagnosis.

### 3 Experiment

The Tennessee Eastman process (TEP), proposed by Downs and Vogel [10], has become a benchmark process for validating process control and fault diagnosis techniques [11]. The corresponding simulation data contains 52 monitoring variables, 22 continuous process variables, 19 composition measurement variables and 11 manipulated variables, a normal operation case and 21 pre-programmed process faults with different fault types, including step changes, random variations in the process variables, slow drift in reaction kinetics, valve sticking



**Fig. 2.** The structure of proposed P-LCKSVD-SAE fault diagnosis model for fault diagnosis

and some unknown faults [11], which can be downloaded from <http://web.mit.edu/braatzgroup/links.html>. In this paper, the label consistent KSVD, stacked autoencoder, and P-LCKSVD-SAE for nonlinear industrial fault diagnosis will be applied to compare and validate the high-precision as well as the effectiveness of the proposed industrial process fault diagnosis method.

Base on the characteristics of fault variables and fault type, three cases will be demonstrated to show the superiority of the proposed method respectively.

### Case 1: Fault 2, 4 and 11

Faults 4 and 11 are associated with the same fault variables but different fault types, and these faults can well represent the overlapping data, which is hard to extract by using conventional multivariate statistical methods like [12]. Table 1 lists the TE process faults, each fault contains 480 training samples and 800 validating samples. For the linear feature extraction step, the size of dictionary atoms is 1050, the sparsity threshold is 20, the weight parameters  $\sqrt{\alpha}$  and  $\sqrt{\beta}$  for label consistent term is 0.04 and 0.02 respectively. For the nonlinear feature extraction step, the SAE model includes 3 layers of neurons: 52, 40, 21. The learning rate is 1, and the activation function is sigmoid function, besides, the number of each batch is 160 for the sake of efficient. In the nonlinear section, after establishing the SAE structure, the features correspond with the **case 1** would be selected to feed into the SVM classifier to fine-tuning and obtain the suitable parameters. The hyperparameters settings in this case are same with the following cases.

Trained feature matrix from linear step and nonlinear step could be combined by the proposed fusion method (**Algorithm 3.(2)**). It is obvious that the newly integrated matrix contains linear and nonlinear information will implement the better performance than the simple linear or nonlinear feature matrix. Table 2 shows the comparison results of three method. The average fault diagnosis rate are 30.45%, 87.5% and 90.2% respectively. Therefore, compared with

the LCKSVD method and SAE algorithm, the proposed P-LCKSVD-SAE model has the better performance.

**Table 1.** List of TE process fault used in case 1

Fault label	Fault variable	Fault type
F2	B composition (Stream 4)	Step
F4	Reactor cooling water inlet temperature	Step
F11	Reactor cooling water inlet temperature	Random

**Table 2.** The comparison of diagnosis performance with three methods

FDR(%)	LCKSVD	SAE	P-LCKSVD-SAE
Fault02	19.5	95.3	96.6
Fault04	8.6	100	96.8
Fault11	63.25	67.3	77.1
Average	30.45	87.5	90.2

### Case 2: Fault 10, 11 and 12

In this section, three faults associated with the same fault type but different fault variable are considered, as detailed in Table 3. Compared to the case 1, the random faults are more difficult to diagnose, which lead to the lower fault diagnosis rate especially in fault 12. Table 4 shows the comparison results of three method, the proposed model reaches 62.4%, which still higher than the other two methods.

**Table 3.** List of TE process fault used in case 2

Fault label	Fault variable	Fault type
F10	C feed temperature (Stream 4)	Random
F11	Reactor cooling water inlet temperature	Random
F12	Condenser cooling water inlet temperature	Random

### Case 3: Fault 1–2, 4–7, 10

In this section, seven faults including different fault variables are considered, as shown in Table 5.

Compared with the above-mentioned cases, case 3 contains more complex and unmanageable training data, which greatly increase the difficulty of data processing. As shown in Table 6, there are many nonlinear information in fault

**Table 4.** The comparison of diagnosis performance with three methods

FDR(%)	LCKSVD	SAE	P-LCKSVD-SAE
Fault10	36.7	0.3	72.0
Fault11	63.2	67.3	77.1
Fault12	0.0	17.3	38.0
Average	33.3	28.3	62.4

**Table 5.** List of TE process fault used in case 5

Fault label	Fault variable	Fault type
F1	A/C feed ratio (Stream 4)	Step
F2	B composition (Stream 4)	Step
F4	Reactor cooling water inlet temperature	Step
F5	Condenser cooling water inlet temperature	Step
F6	A feed loss (Stream 1)	Step
F7	C header pressure loss-reduced availability	Random
F10	C feed temperature (Stream 4)	Random

5–6 that LC-KSVD is hard to extract, but the SAE is proved to be great in solving nonlinear data. Integrated with two methods, the fault diagnosis rate of proposed method reaches 79.2%, which shows the overwhelming superiority than the other two methods.

**Table 6.** The comparison of diagnosis performance with three methods

FDR(%)	LCKSVD	SAE	P-LCKSVD-SAE
Fault01	96.0	88.1	78.0
Fault02	93.1	96.7	96.6
Fault04	99.2	97.5	96.8
Fault05	0.0	94.7	63.2
Fault06	0.0	95.7	94.5
Fault07	99.0	9.6	53.1
Fault10	36.7	3.2	72.0
Average	61.0	69.4	79.2

The diagnosis results of LCKSVD, SAE and P-LCKSVD-SAE model for nonlinear process fault diagnosis are presented in Table 2, 4 and 6. The FDR of proposed parallel method shows the best performance in all five cases. These findings verify the effectiveness and superiority of proposed parallel LCKSVD-SAE model in the nonlinear industrial process.

## 4 Conclusion

In this study, the diagnosis scheme based on the parallel LCKSVD-SAE model is proposed to deal with a chemical process that includes variables with linear and nonlinear correlations. First, label consistent KSVD and stacked autoencoder methods are utilized to extract the linear and nonlinear features hidden in raw industrial process, which determine the performance of the subsequent fault classification problem. Second, fusion matrix contains the linear and nonlinear features is established by proposed feature matrix fusion algorithm. Third, feed the newly fusion matrix into a mature SVM classifier, the effectiveness and diagnosis performance of proposed parallel LCKSVD-SAE method are superior to those of label consistent KSVD and stacked autoencoder methods, which demonstrated by application example on Tennessee Eastman process. This work studies the feature extraction and diagnosis of nonlinear processes. However, it discusses only fault classification issues. Fault modeling and detection using the proposed parallel scheme are topics for further research.

## References

- Huang, K., Wen, H., Ji, H., et al.: Nonlinear process monitoring using kernel dictionary learning with application to aluminum electrolysis process. *J. Control Eng. Pract.* **89**, 94–102 (2019)
- Jiang, Q., Yan, X.: Parallel PCACKPCA for nonlinear process monitoring. *J. Control Eng. Pract.* **80**, 17–25 (2018)
- Yuan, X., Wang, Y., Yang, C., et al.: Weighted linear dynamic system for feature representation and soft sensor application in nonlinear dynamic industrial processes. *J. IEEE Trans. Ind. Electron.* **65**(2), 1508–1517 (2017)
- Guo, T., Zhou, D., Zhang, J., et al.: Fault detection based on robust characteristic dimensionality reduction. *J. Control Eng. Pract.* **84**, 125–138 (2019)
- Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. *Science* **313**(5786), 504–507 (2006)
- Jiang, Z., Lin, Z., Davis, L.S.: Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In: The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011. IEEE (2011)
- Peng, X., Tang, Y., Du, W., et al.: Multimode process monitoring and fault detection: a sparse modeling and dictionary learning method. *J. IEEE Trans. Ind. Electron.* **64**(6), 4866–4875 (2017)
- Zhang, Z., Zhao, J.: A deep belief network based fault diagnosis model for complex chemical processes. *J Comput. Chem. Eng.* **107**, 395–407 (2017)
- Shahnazari, H.: Fault diagnosis of nonlinear systems using recurrent neural networks. *J. Chem. Eng. Res. Design* **153**, 233–245 (2020)
- Downs, J.J., Vogel, E.F.: A plant-wide industrial control problem. *J. Comput. Chem. Eng.* **17**(3), 245–255 (1993)
- Deng, X., Tian, X., Chen, S., et al.: Nonlinear process fault diagnosis based on serial principal component analysis. *J. IEEE Trans. Neural Netw. Learn. Syst.* **29**(3), 560–572 (2018)
- He, X.B., Wang, W., Yang, Y.P., et al.: Variable-weighted Fisher discriminant analysis for process fault diagnosis. *J. Process Control* **19**(6), 923–931 (2009)



# Iterative Learning Formation Control for Multi-agent Systems with Randomly Varying Trial Lengths

Yimin Fan, Yang Liu<sup>(✉)</sup>, and Na Wang

The Seventh Research Division and School of Automation Science and Electrical Engineering, Beihang University (BUAA), Beijing 100191, China  
ylbuaa@163.com

**Abstract.** The formation control problem of iterative learning system with stochastic variable lengths is studied. In particular, we establish an iterative learning control protocol for multi-agent system with switching topology, in which a new formation state error is proposed to deal with different lengths. Using the redefined  $\lambda$ -norm and mathematical expectation, the convergence conditions are derived. The simulation results show that the method is effective.

**Keywords:** Multi-agent formation system · Control protocol · Iterative Learning Control (ILC) · Different lengths

## 1 Introduction

Multi-agent formation technology is widely used in spacecraft, mobile robot and aircraft control [1–4]. Specially, iterative learning control (ILC) has been proposed to multi-agent systems [5–8]. In classical iterative learning control, the trial length of each iteration should be unchanged. However, the trial length of a single agent can vary within a range [9]. In the recent years, some iterative systems have involved an agent with stochastic variable lengths [9–13]. However, the research results of formation control based on ILC under different trial lengths are few.

In this article, noting that a portion of formation information will be lost because of the randomness of the iterative lengths, we give a modified formation state error. The convergence analysis of formation system under the proposed ILC scheme is investigated. The results show that the desired formation is achieved in the whole motion process. Finally, the simulation results are given.

**Notations:** **1** and **0** denote the column vectors.  $I_c$  is the  $c \times c$  identity matrix.  $\otimes$  is the Kronecker product.  $\|\cdot\|$  is the Euclidean norm.  $\|\pi(t)\|_\lambda$  is the  $\lambda$ -norm of  $\|\pi(t)\|$  and  $\|\pi(t)\|_\lambda = \sup_{t \in \zeta} \nu^{-\lambda t} E \|\pi(t)\|$ ,  $\lambda > 0$ ,  $\nu > 1$ .

## 2 Problem Description

There is a system consisting  $n$  agents with the  $l$ th one such that

$$\dot{g}_{k,l}(t) = h(g_{k,l}(t)) + W(t)u_{k,l}(t). \quad (1)$$

The state and the input of the agent  $l$  are  $g_{k,l}(t) \in R^m$  and  $u_{k,l}(t) \in R^{m_1}$  for the  $k$ th iteration, respectively. The desired trial length is  $T_d$ ,  $t \in [0, T_d]$ .  $h(g_{k,l}(t)) \in R^m$  is a continuously nonlinear function.

The actual trial length is defined by  $T_k$ , which can vary from  $T_{min}$  to  $T_{max}$ . There are two conditions that  $T_k < T_d$  and  $T_k \geq T_d$  ( $T_d \in [T_{min}, T_{max}]$ ). The second case can be viewed as  $T_k = T_d$ . If  $T_k < T_d$ , the information between  $T_k$  and  $T_d$  is missing.

The system (1) is rewritten as

$$\dot{g}_k(t) = H(g_k(t)) + (I_n \otimes W(t))u_k(t), \quad (2)$$

in which  $g_k(t) \in R^{nm}$  and  $u_k(t) \in R^{nm_1}$  consist of  $g_{k,l}(t)$  and  $u_{k,l}(t)$ , respectively.  $H(g_k(t)) = [h^T(g_{k,1}(t)), \dots, h^T(g_{k,n}(t))]^T$ .

**Assumption 1.**  $H(g_{k,l}(t))$  is globally Lipschitz in  $g_{k,l}(t)$ ,

$$\|h(g_{k+1,l}(t)) - h(g_{k,l}(t))\| \leq k_f \|g_{k+1,l}(t) - g_{k,l}(t)\|,$$

in which  $k_f$  is a suitable Lipschitz constant, then

$$\|H(g_{k+1}(t)) - H(g_k(t))\| \leq k_f \|g_{k+1}(t) - g_k(t)\|. \quad (3)$$

**Assumption 2.**  $g_k(0) = d(0)$  and  $d(0)$  is a constant.

The desired formation is achieved when

$$e_l(t) = e_j(t), \quad \forall l, j \in N, \quad (4)$$

in which

$$e_l(t) = g_l(t) - d_l(t) \quad (5)$$

denotes the state error of the agent  $l$ , and  $d_l(t)$  is fixed relative state. A new variable is introduced by

$$y_l(t) = e_1(t) - e_{l+1}(t). \quad (6)$$

Setting  $y(t) = [y_1^T(t), \dots, y_{n-1}^T(t)]^T$  and  $e(t) = [e_1^T(t), \dots, e_n^T(t)]^T$ , we have

$$y(t) = (S \otimes I_m)e(t), \quad (7)$$

and

$$e(t) = \mathbf{1} \otimes e_1(t) + (M \otimes I_m)y(t), \quad (8)$$

where

$$S = [\mathbf{1} \quad -I_{n-1}], \quad M = \begin{bmatrix} \mathbf{0}^T \\ -I_{n-1} \end{bmatrix}. \quad (9)$$

It can be seen from the above that the consensus is reached on state errors when  $y(t) = \mathbf{0}$ .

### 3 ILC Design

Inserting the iteration index  $k$  into (5) and (6), and observing that the expected formation is fixed, we can derive

$$e_{k,l}(t) = g_{k,l}(t) - d_l(t), \quad l = 1, \dots, n, \quad (10)$$

and

$$y_{k,l}(t) = e_{k,1}(t) - e_{k,l+1}(t), \quad l = 1, \dots, n-1. \quad (11)$$

The corresponding variable vectors are denoted by  $e_k(t)$  and  $y_k(t)$ , respectively.

At time  $t$ ,  $q(t)$  denotes the probability that the system has an output. From 0 to  $T_{min}$ ,  $q(t) = 1$ . From  $T_{min}$  to  $T_{max}$ ,  $q(t) \in (0, 1)$ .  $\varepsilon_k(t)$  obeys Bernoulli distribution. Specially,  $\varepsilon_k(t) = 1$  represents the system (1) will work more than the time  $t$  for iteration  $k$ , whose probability is  $q(t)$ , and vice versa.

A special state error is as below

$$e_k^*(t) = \begin{cases} e_k(t), & 0 \leq t \leq T_k \\ 0, & T_k < t \leq T_d \end{cases} \quad (12)$$

Obviously, we can get

$$e_k^*(t) = \varepsilon_k(t)e_k(t). \quad (13)$$

An iterative learning protocol is presented as follows

$$u_{k+1,l}(t) = u_{k,l}(t) + \Gamma \sum_{j \in N_{k,l}(t)} a_{k,lj}(t)(\dot{e}_{k,j}^*(t) - \dot{e}_{k,l}^*(t)). \quad (14)$$

Equation (14) is rewritten as

$$u_{k+1}(t) = u_k(t) - (L_{\sigma_k(t)} \otimes \Gamma)\dot{e}_k^*(t), \quad (15)$$

of which  $\Gamma \in \mathbb{R}^{m_1 \times m}$  denotes a gain matrix,  $\sigma_k(t)$  and  $L_{\sigma_k(t)}$  are the switching signal function and the Laplacian matrix of graph  $\mathfrak{G}_{\sigma_k(t)}$ , respectively.

### 4 Convergence Analysis

**Theorem 1.** *Apply (15) to the system (1) with randomly varying length, and let Assumptions 1 and 2 hold. If  $\Gamma \in \mathbb{R}^{m_1 \times m}$  satisfies*

$$\sup_{t,k} \|I - (SL_{\sigma_k(t)}M \otimes W(t)\Gamma)\| < 1, \quad (16)$$

$y_k(t)$  will converge to zero with iteration increasing.

*Proof.* Combining with Assumption 2, we have

$$\begin{aligned}
y_{k+1}(t) - y_k(t) &= \int_0^t (S \otimes I_m)[H(g_{k+1}(\tau)) - H(g_k(\tau))]d\tau \\
&\quad + \int_0^t (S \otimes W(\tau))[u_{k+1}(\tau) - u_k(\tau)]d\tau \\
&= \int_0^t (S \otimes I_m)[H(g_{k+1}(\tau)) - H(g_k(\tau))]d\tau \\
&\quad - \int_0^t (S \otimes W(\tau))(L_{\sigma_k(t)} \otimes \Gamma)\dot{e}_k^*(\tau)d\tau \\
&= \int_0^t (S \otimes I_m)[H(g_{k+1}(\tau)) - H(g_k(\tau))]d\tau \\
&\quad - (SL_{\sigma_k(t)} \otimes W(t)\Gamma)e_k^*(t) + \int_0^t \psi_k(\tau)e_k^*(\tau)d\tau,
\end{aligned} \tag{17}$$

where  $\psi_k(\tau) = D^+[SL_{\sigma_k(\tau)} \otimes W(\tau)\Gamma]$ .

According to (8),

$$\int_0^t \psi_k(\tau)e_k^*(\tau)d\tau = \int_0^t \psi_k(\tau)(M \otimes I_m)\varepsilon_k(\tau)y_k(\tau)d\tau. \tag{18}$$

Since  $L_{\sigma_k(t)}\mathbf{1} = \mathbf{0}$ ,

$$\begin{aligned}
y_{k+1}(t) - y_k(t) &= \int_0^t (S \otimes I_m)[H(g_{k+1}(\tau)) - H(g_k(\tau))]d\tau \\
&\quad - (SL_{\sigma_k(t)} \otimes W(t)\Gamma)[\mathbf{1} \otimes (\varepsilon_k(t)e_{k,1}(t)) + (M \otimes I_m)\varepsilon_k(t)y_k(t)] \\
&\quad + \int_0^t \psi_k(\tau)(M \otimes I_m)\varepsilon_k(\tau)y_k(\tau)d\tau \\
&= \int_0^t (S \otimes I_m)[H(g_{k+1}(\tau)) - H(g_k(\tau))]d\tau \\
&\quad - (SL_{\sigma_k(t)}M \otimes W(t)\Gamma)\varepsilon_k(t)y_k(t) \\
&\quad + \int_0^t \psi_k(\tau)(M \otimes I_m)\varepsilon_k(\tau)y_k(\tau)d\tau.
\end{aligned} \tag{19}$$

Taking Euclidean norm yields

$$\begin{aligned}
\|y_{k+1}(t)\| &\leq \int_0^t k_f \|S\| \|g_{k+1}(\tau) - g_k(\tau)\| d\tau \\
&\quad + \|(I - \varepsilon_k(\tau)(SL_{\sigma_k(t)}M \otimes W(t)\Gamma))y_k(t)\| \\
&\quad + \int_0^t \|\psi_k(\tau)(M \otimes I_m)\varepsilon_k(\tau)y_k(\tau)\| d\tau.
\end{aligned} \tag{20}$$

Taking mathematical expectation, we know

$$\begin{aligned} E\|y_{k+1}(t)\| &\leq \int_0^t k_f \|S\| \cdot E\|g_{k+1}(\tau) - g_k(\tau)\| d\tau \\ &\quad + \sup_{t,k} E\|I - \varepsilon_k(t)(SL_{\sigma_k(t)}M \otimes W(t)\Gamma)\| \cdot E\|y_k(t)\| \\ &\quad + \int_0^t \varphi \cdot E\|y_k(\tau)\| d\tau, \end{aligned} \quad (21)$$

where  $\varphi = \sup_{\tau,k} E\|\psi_k(\tau)(M \otimes I_m)\varepsilon_k(\tau)\|$ .

Adding  $e^{-\lambda t}$  to both sides of above equation follows

$$\begin{aligned} e^{-\lambda t} E\|y_{k+1}(t)\| &\leq \int_0^t k_f \|S\| \cdot e^{-\lambda t} E\|g_{k+1}(\tau) - g_k(\tau)\| d\tau \\ &\quad + \sup_{t,k} E\|I - \varepsilon_k(t)(SL_{\sigma_k(t)}M \otimes W(t)\Gamma)\| \cdot e^{-\lambda t} E\|y_k(t)\| \\ &\quad + \int_0^t \varphi \cdot e^{-\lambda t} E\|y_k(\tau)\| d\tau \\ &\leq (k_f \|S\| \cdot \|g_{k+1}(\tau) - g_k(\tau)\|_\lambda + \varphi \cdot \|y_k(t)\|_\lambda) \cdot \frac{1 - e^{-\lambda T_d}}{\lambda} \\ &\quad + \sup_{t,k} E\|I - \varepsilon_k(t)(SL_{\sigma_k(t)}M \otimes W(t)\Gamma)\| \cdot \|y_k(t)\|_\lambda. \end{aligned} \quad (22)$$

Further, noticing Assumption 2, we obtain

$$\begin{aligned} g_{k+1}(t) - g_k(t) &= \int_0^t [H(g_{k+1}(\tau)) - H(g_k(\tau))] d\tau \\ &\quad - (L_{\sigma_k(t)} \otimes W(t)\Gamma)e_k^*(t) + \int_0^t \varrho_k(\tau)e_k^*(\tau)d\tau \\ &= \int_0^t [H(g_{k+1}(\tau)) - H(g_k(\tau))] d\tau \\ &\quad - (L_{\sigma_k(t)}M \otimes W(t)\Gamma)\varepsilon_k(t)y_k(t) \\ &\quad + \int_0^t \varrho_k(\tau)(M \otimes I_m)\varepsilon_k(\tau)y_k(\tau)d\tau, \end{aligned} \quad (23)$$

with  $\varrho_k(\tau) = D^+[L_{\sigma_k(\tau)} \otimes W(\tau)\Gamma]$ .

Applying the Euclidean norm leads to

$$\begin{aligned} \|g_{k+1}(t) - g_k(t)\| &\leq \int_0^t \|H(g_{k+1}(\tau)) - H(g_k(\tau))\| d\tau \\ &\quad + \sup_{t,k} \|\varepsilon_k(t)(L_{\sigma_k(t)}M \otimes W(t)\Gamma)\| \cdot \|y_k(t)\| \\ &\quad + \sup_{\tau,k} \|\varepsilon_k(\tau)\varrho_k(\tau)(M \otimes I_m)\| \int_0^t \|y_k(\tau)\| d\tau. \end{aligned} \quad (24)$$

Similar to (22),

$$\begin{aligned} \|g_{k+1}(t) - g_k(t)\|_\lambda &\leq (k_f \|g_{k+1}(t) - g_k(t)\|_\lambda + \theta \cdot \|y_k(t)\|_\lambda) \cdot \kappa(\lambda^{-1}) \\ &\quad + \eta \cdot \|y_k(t)\|_\lambda \end{aligned} \quad (25)$$

with  $\theta = \sup_{\tau, k} E\|\varepsilon_k(\tau)\varrho_k(\tau)(M \otimes I_m)\|$ ,  $\eta = \sup_{t, k} E\|\varepsilon_k(t)(L_{\sigma_k(t)} M \otimes W(t)\Gamma)\|$  and  $\kappa(\lambda^{-1}) = \frac{1-e^{-\lambda T_d}}{\lambda}$ .

The above inequality is written as

$$\|g_{k+1}(t) - g_k(t)\|_\lambda \leq \frac{\kappa(\lambda^{-1})\theta + \eta}{1 - \kappa(\lambda^{-1})k_f} \|y_k(t)\|_\lambda. \quad (26)$$

Substituting (26) into (22) follows

$$\begin{aligned} \|y_{k+1}(t)\|_\lambda &\leq \left( \frac{\kappa(\lambda^{-1})\theta + \eta}{1 - \kappa(\lambda^{-1})k_f} \cdot k_f \|S\| + \varphi \right) \cdot \kappa(\lambda^{-1}) \cdot \|y_k(t)\|_\lambda \\ &\quad + \sup_{t, k} E\|I - \varepsilon_k(t)(SL_{\sigma_k(t)} M \otimes W(t)\Gamma)\| \cdot \|y_k(t)\|_\lambda. \end{aligned} \quad (27)$$

If  $\Gamma$  satisfies (16), then equivalently

$$\sup_{t, k} [ \|I - (SL_{\sigma_k(t)} M \otimes W(t)\Gamma)\| q(t) + 1 - q(t) ] < 1. \quad (28)$$

We can choose a suitable  $\lambda$  to achieve

$$\left( \frac{\kappa(\lambda^{-1})\theta + \eta}{1 - \kappa(\lambda^{-1})k_f} \cdot k_f \|S\| + \varphi \right) \cdot \kappa(\lambda^{-1}) + \sup_{t, k} E\|I - \varepsilon_k(t)(SL_{\sigma_k(t)} M \otimes W(t)\Gamma)\| < 1.$$

It means

$$\lim_{k \rightarrow \infty} \|y_k(t)\|_\lambda = 0, \quad \forall t.$$

Further,

$$\lim_{k \rightarrow \infty} E\|y_k(t)\| = 0, \quad \forall t.$$

Obviously,

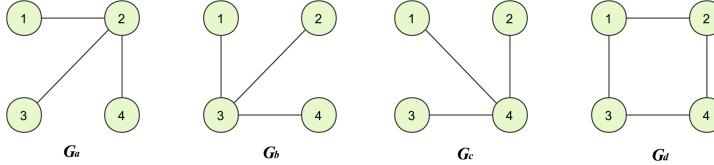
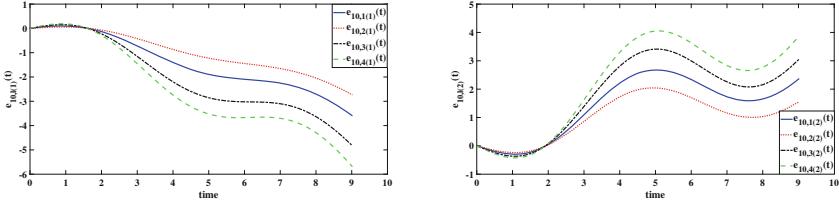
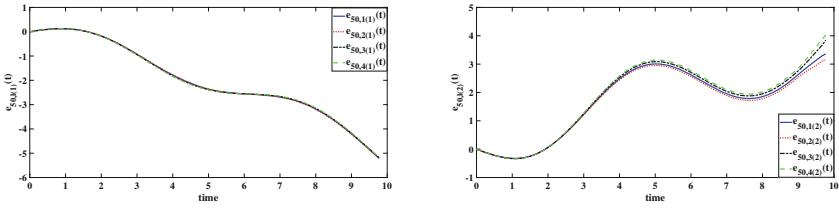
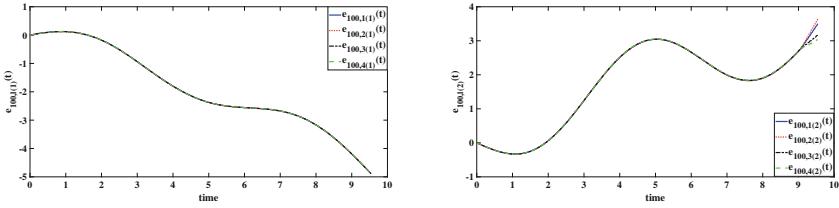
$$\lim_{k \rightarrow \infty} \|y_k(t)\| = 0, \quad \forall t.$$

This completes the proof. ■

## 5 Illustrative Simulations

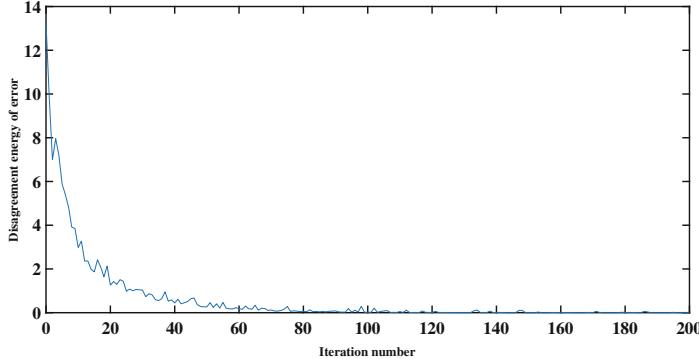
The system (1) is formed by four agents, and  $T_d$  is 10 s,

$$W(t) = \begin{bmatrix} 1 & 0 \\ -0.5 + \sin(t) & 3 + 0.5 \cos(t) \end{bmatrix}, h(g_l(t)) = \begin{bmatrix} 0.1g_{l1} \cos(0.5t) \\ -0.2g_{l2} \cos(t^2) \end{bmatrix},$$

**Fig. 1.** Switching interaction graphs.**Fig. 2.** State errors of all agents at iteration 10.**Fig. 3.** State errors of all agents at iteration 50.**Fig. 4.** State errors of all agents at iteration 100.

where  $g_l(t) = [g_{l1}(t) \ g_{l2}(t)]^T \in \mathbb{R}^2$ . The desired relative states are

$$\begin{aligned} d_1(t) &= \begin{bmatrix} 0 \\ 10 - \cos(t) \end{bmatrix}, \quad d_2(t) = \begin{bmatrix} 5 + \cos(t) \\ 10 - \cos(t) \end{bmatrix}, \\ d_3(t) &= \begin{bmatrix} 5 + \cos(t) \\ 0 \end{bmatrix}, \quad d_4(t) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned} \tag{29}$$



**Fig. 5.** The formation learning process.

The interaction graph switches within the graph set  $\mathfrak{G}_s = \{G_a, G_b, G_c, G_d\}$  in Fig. 1.

Denote  $g_0 = [0 \ 10 \ 0 \ 0 \ 5 \ 0 \ 5 \ 10]^T$  and  $u_0(t) = 0$ ,  $t = 0, \dots, T_d$ . Moreover, assume that  $T_{min} = 9$  and  $T_k \in [9, 10]$  satisfies continuous uniform distribution. The state errors  $e_{k,l}(t)$  are shown in Figs. 2, 3 and 4 and coincide with Theorem 1. The disagreement of state errors are given by  $z_k(t) = (L_c \otimes I_m)e_k(t)$ . The diagonal elements of  $L_c \in \Re^{n \times n}$  are  $\frac{n-1}{n}$  and else elements are  $\frac{-1}{n}$ . So we set the Y-axis as  $\|z_k(t)\|_2$  in Fig. 5. When  $z_k(t) = \mathbf{0}$ , the formation error converges to origin and the formation is realized with iteration increasing.

## 6 Conclusions

From what has been discussed above, we address the formation problem with stochastic variable lengths. When the proposed protocol is applied, the expected formation can be achieved. The ILC scheme relaxes the condition that each iterative trial should be the same.

**Acknowledgements.** This work was supported by the NSFC (61473015, 61520106010, 61976013) and the National Basic Research Program of China (973 Program: 2012CB821201).

## References

1. Huang, H., Ma, G., Zhuang, Y., Lv, Y.: Optimal spacecraft formation reconfiguration with collision avoidance using particle swarm optimization. *Inf. Technol. Control* **41**(2), 143–150 (2012)
2. Wang, J., Liu, Z., Hu, X.: Consensus control design for multi-agent systems using relative output feedback. *Int. J. Syst. Sci.* **27**(2), 237–251 (2014)

3. Fu, J., Wang, Q., Wang, J.: Robust finite-time consensus tracking for second-order multi-agent systems with input saturation under general directed communication graphs. *Int. J. Control.* **92**(8), 1785–1795 (2019)
4. Ren, W., Atkins, E.: Distributed multi-vehicle coordinated control via local information exchange. *Int. J. Robust Nonlinear Control* **17**(10–11), 1002–1033 (2010)
5. Liu, Y., Jia, Y.: An iterative learning approach to formation control of multi-agent systems. *Syst. Control Lett.* **61**(1), 148–154 (2012)
6. Liu, Y., Jia, Y.: Formation control of discrete-time multi-agent systems by iterative learning approach. *Int. J. Control Autom. Syst.* **10**(5), 913–919 (2012)
7. Liu, Y., Jia, Y.: Robust formation control of discrete-time multi-agent systems by iterative learning approach. *Int. J. Syst. Sci.* **46**(4), 625–633 (2015)
8. Meng, D., Jia, Y., Du, J., Zhang, J.: On iterative learning algorithms for the formation control of nonlinear multi-agent systems. *Automatica* **50**(1), 291–295 (2014)
9. Shen, D., Xu, J.X.: Iterative learning control algorithm with gain adaptation for stochastic systems. *IEEE Trans. Autom. Control* **65**(3), 1280–1287 (2020)
10. Li, X., Shen, D.: Two novel iterative learning control schemes for systems with randomly varying trial lengths. *Syst. Control Lett.* **107**, 9–16 (2017)
11. Li, X., Xu, J., Huang, D.: Iterative learning control for nonlinear dynamic systems with randomly varying trial lengths. *Int. J. Adapt. Control Signal Process.* **29**(11), 1341–1353 (2015)
12. Liu, C., Shen, D., Wang, J.R.: Adaptive learning control for general nonlinear systems with nonuniform trial lengths, initial state deviation, and unknown control direction. *Int. J. Robust Nonlinear Control.* **29**(17), 6227–6243 (2019)
13. Seel, T., Werner, C., Schauer, T.: The adaptive drop foot stimulator-multivariable learning control of foot pitch and roll motion in paretic gait. *Med. Eng. Phys.* **38**(11), 1205–1213 (2016)



# Backstepping-Based Adaptive Neural Control of Constrained Nonlinear Systems

Penghao Chen, Tianping Zhang<sup>(✉)</sup>, Houbin Qian, and Yang Yi

College of Information Engineering, Yangzhou University, Yangzhou 225127, China  
tpzhang@yzu.edu.cn

**Abstract.** In this article, the problem of backstepping-based adaptive neural control is discussed for nonlinear systems in strict-feedback form with full state restrictions and dynamical uncertainties. The constrained affine nonlinear system is transformed into a nonaffine nonlinear system without state restrictions by constructing invertible asymmetric nonlinear mapping. The uncertainties are disposed of by the aid of first-order auxiliary dynamic system. Based on modified backstepping design and using the property of Gaussian function, an improved adaptive backstepping control scheme is proposed for the novel unconstrained nonaffine nonlinear systems. The computation complex generated by the conventional backstepping design and mean value theorem is avoided. All signals of the whole system are proved to be semi-globally uniformly ultimately boundedness (SGUUB).

**Keywords:** Backstepping · State constraints · Neural networks · Unmodeled dynamics · Strict-feedback systems

## 1 Introduction

As we know, unmodeled dynamics had been widely studied in order to dispose of its negative effect. Many effective design methods were proposed in [1–9] such as auxiliary dynamic signal method [1,3,5,6], Lyapunov function description [2,9], normalization signal method [7–9]. With the help of mean value theorem, adaptive dynamic surface control (DSC) was proposed for nonaffine systems including unmodeled dynamics in [4].

On the other hand, output and state constraints usually have also effected on the system performance. A lot of constrained control problems were discussed in [10–21]. Backstepping design and dynamic surface control technique were usually employed to design adaptive control scheme of constrained nonlinear systems in [10–21]. Output constraints [10,11] or time-varying output constraints [12,13] or full states constraints [14,15] were handled by the aid of barrier Lyapunov function (BLF). A symmetric nonlinear mapping was used to dispose of the output restriction problems for special strict-feedback system with one as control coefficient in [16] and the switched stochastic nonaffine systems with output

constraints in [17]. However, the output constraint concept in probability had not been defined in [17]. By introducing asymmetric invertible change, adaptive control was discussed for pure-feedback systems with time-varying asymmetric output restrictions in [18]. In [19, 20], based on invertible asymmetric nonlinear mapping and dynamic surface control technique, two adaptive control strategies were proposed for affine and nonaffine systems with whole state restrictions. Adaptive neural DSC was presented based on integral BLF (iBLF) for nonaffine systems subject to all state restrictions in [21]. However, each virtual coefficient and the derivative were supposed to be bounded.

In this article, by the aid of improved backstepping, a novel neural control scheme with simpler structure is developed for uncertain constrained affine nonlinear systems. The main contributions of the paper are listed as follows:

- (i) By the aid of asymmetric invertible nonlinear mapping and using modified backstepping, backstepping-based adaptive control is developed for uncertain strict-feedback nonlinear systems with full state constraints, and the constrained conditions are never violated.
- (ii) The restrictions of the virtual coefficients are relaxed by using modified backstepping method. With the help of the value of Gaussian function belonging to the open interval (0,1), the number of input variable in neural networks is effectively reduced. The computation complex produced by the conventional backstepping design and mean value theorem are avoided.

## 2 Problem Formulation and Basic Assumptions

Consider the following nonlinear systems in strict-feedback form:

$$\begin{cases} \dot{\zeta} = Q(t, \zeta, x) \\ \dot{x}_i = f_i(\bar{x}_i) + g_i(\bar{x}_i)x_{i+1} + d_i(t, \zeta, x) \\ \quad 1 \leq i \leq n-1 \\ \dot{x}_n = f_n(x) + g_n(x)u + d_n(t, \zeta, x) \\ y = x_1 \end{cases} \quad (1)$$

where  $x = [x_1, x_2, \dots, x_n]^T \in R^n$  is the state,  $\bar{x}_i = [x_1, x_2, \dots, x_i]^T$ ,  $\zeta \in R^{n_0}$  is unmodeled dynamics,  $u \in R$  is the input,  $y \in R$  is the output.  $f_i(\cdot)$ ,  $g_i(\cdot)$ ,  $Q(\cdot)$  are the unknown smooth nonlinear functions,  $d_1(\cdot), \dots, d_n(\cdot)$  are the Lipschitz functions. All the states  $x_i$  belong to the open sets  $\Omega_{x_i} = \{x_i : -k_{b_{i1}} < x_i < k_{b_{i2}}\}$ , where  $k_{b_{i1}} > 0, k_{b_{i2}} > 0$  are known constants.

The control goal is to design the input  $u(t)$  for plant (1) such that  $y$  tracks  $y_d$ , and  $x_i \in \Omega_{x_i}$  holds, and the number of input variable of neural networks and computation burden will be effectively reduced compared with conventional backstepping.

**Assumption 1.** *The unmodeled dynamics  $\zeta$  is said to be exponentially input-state-practically stable (exp-ISpS), i.e., for system  $\dot{\zeta} = Q(t, \zeta, x)$ , if there exist class  $K_\infty$  functions  $\bar{\alpha}_1$  and  $\bar{\alpha}_2$  and a Lyapunov function  $W(\zeta)$  such that*

$$\bar{\alpha}_1(\|\zeta\|) \leq W(\zeta) \leq \bar{\alpha}_2(\|\zeta\|) \quad (2)$$

and there exist two constants  $c > 0, d \geq 0$  and a class  $K_\infty$  function  $\gamma$  such that

$$\frac{\partial W(\zeta)}{\partial \zeta} Q(t, \zeta, x) \leq -cW(\zeta) + \gamma(|y|) + d, \quad \forall [t, \zeta, x] \in R_+ \times R^{n_0} \times R^n \quad (3)$$

where  $c > 0$  and  $d \geq 0$  are known,  $\gamma(\cdot)$  is a known class  $K_\infty$  function.

**Assumption 2.** The dynamic disturbances  $d_i(t, \zeta, x)$  satisfy the following inequalities

$$|d_i(t, \zeta, x)| \leq \varphi_{i1}(\|\bar{x}_i\|) + \varphi_{i2}(\|\zeta\|), \quad \forall (t, \zeta, x) \in R_+ \times R^{n_0} \times R^n \quad (4)$$

where  $i = 1, \dots, n$ ,  $\varphi_{i1}(\cdot) \geq 0$  are continuous functions, and  $\varphi_{i2}(\cdot) \geq 0$  are monotone increasing continuous functions.

**Assumption 3.** Suppose  $\bar{y}_{di} = [y_d, \dot{y}_d, \dots, \dot{y}_d^{(i)}]^T \in \Omega_{di} \subset R^{i+1}$ ,  $i = 1, \dots, n$  are available, where the compact sets  $\Omega_{di}$  are known.

**Assumption 4.** Suppose  $0 < b_i \leq g_i(\bar{x}_i)$ ,  $i = 1, \dots, n$ , and  $g_n(\bar{x}_n) \geq b_n > 0$  are true, where  $b_i$  are some unknown positive constants.

**Lemma 1.** If  $W$  is an exp-ISpS Lyapunov function for a system  $\dot{\zeta} = Q(t, \zeta, x)$ , i.e. (2) and (3) hold, then,  $\forall \bar{c} \in (0, c)$ , any initial condition  $\zeta(0) = \zeta_0$ ,  $r_0 > 0$ , for any continuous function  $\bar{\gamma}$  such that  $\bar{\gamma}(|x_1|) \geq \gamma(|x_1|)$ , an auxiliary system designed by

$$\dot{r} = -\bar{c}r + \bar{\gamma}(|x_1|) + d, \quad r(0) = r_0 \quad (5)$$

such that  $W(\zeta) \leq r(t) + D(t)$ , and  $D(t) = 0$  for  $t \geq T_0$  with  $D(t) = \max\{0, e^{-ct} \times W(\zeta_0) - e^{-\bar{c}t} r_0\}$ , and  $T_0 = \max\{0, \ln[\frac{W(\zeta_0)}{r_0}] / (c - \bar{c})\} \geq 0$ .

In this article,  $\|\cdot\|$  stands for Euclidian norm,  $(\tilde{\cdot}) = (\cdot)^* - (\hat{\cdot})$ .

### 3 Adaptive Neural Control

To handle the problem of restrictions, invertible nonlinear mapping (NM) is defined as follows:

$$s_i = \log \frac{k_{b_{i1}} + x_i}{k_{b_{i2}} - x_i}, \quad i = 1, \dots, n \quad (6)$$

From (6), its inverse change is

$$x_i = k_{b_{i2}} - \frac{k_{b_{i2}} + k_{b_{i1}}}{e^{s_i} + 1}, \quad i = 1, \dots, n \quad (7)$$

Therefore, we obtain

$$\dot{s}_i = \frac{e^{s_i} + e^{-s_i} + 2}{k_{b_{i1}} + k_{b_{i2}}} \dot{x}_i \quad (8)$$

System (1) can be rewritten as follows:

$$\begin{cases} \dot{\zeta} = Q(t, \zeta, x) \\ \dot{s}_1 = F_1(s_1, s_2) + s_2 + D_1(t, \zeta, \bar{s}_n) \\ \vdots \\ \dot{s}_{n-1} = F_{n-1}(\bar{s}_{n-1}, s_n) + s_n + D_{n-1}(t, \zeta, \bar{s}_n) \\ \dot{s}_n = F_n(\bar{s}_n) + \kappa_n(s_n)G_n(\bar{s}_n, u)u + D_n(t, \zeta, \bar{s}_n) \end{cases} \quad (9)$$

where  $\bar{s}_i = [s_1, \dots, s_i]^T, i = 1, \dots, n$ ,

$$\kappa_i(s_i) = \frac{e^{s_i} + e^{-s_i} + 2}{k_{b_{i1}} + k_{b_{i2}}}, \quad i = 1, \dots, n \quad (10)$$

$$\begin{aligned} F_i(\bar{s}_{i+1}) &= \kappa_i(s_i) \left( f_i(\bar{x}_i) + g_i(\bar{x}_i)x_{i+1} \right) - s_{i+1}, \\ i &= 1, \dots, n \end{aligned} \quad (11)$$

$$F_n(\bar{s}_n) = \kappa_n(s_n)f_n(\bar{x}_n) \quad (12)$$

$$G_n(\bar{s}_n) = g_n(\bar{x}_n) \quad (13)$$

$$D_i(t, \zeta, \bar{s}_n) = \kappa_i(s_i)d_i(t, \zeta, x), \quad i = 1, \dots, n \quad (14)$$

Let

$$\begin{cases} z_1 = s_1 - \hat{y}_d \\ z_2 = s_2 - \alpha_1 \\ \vdots \\ z_n = s_n - \alpha_{n-1} \end{cases} \quad (15)$$

where  $\hat{y}_d = \log \frac{k_{b_{11}} + y_d}{k_{b_{12}} - y_d}$ ,  $\alpha_i$  will be designed later.

Let  $\Omega_{Z_i} \subset R^{N_i}$  be a given compact set. Using neural network  $W_i^{*T}S_i(Z_i)$  approximates unknown continuous functions  $E_i(Z_i)$  over the compact set  $\Omega_{Z_i}$  as addressed in [22], where  $Z_i$  and  $E_i(Z_i)$  will be given later. Then, we have

$$E_i(Z_i) = W_i^{*T}S_i(Z_i) + \delta_i(Z_i) \quad (16)$$

where  $S_i(Z_i) = [s_{i1}(Z_i), \dots, s_{il_i}(Z_i)]^T \in R^{l_i}$  is the basis vector, the basis function  $s_{ij}(Z_i)$  is taken as follows:

$$s_{ij}(Z_i) = \exp \left[ -\frac{(Z_i - \mu_{ij})^T(Z_i - \mu_{ij})}{\phi_{ij}^2} \right] \quad (17)$$

$j = 1, \dots, l_i$ ,  $i = 1, \dots, n$ ,  $\mu_{ij} = [\mu_{ij1}, \mu_{ij2}, \dots, \mu_{ijq_{ij}}]^T$ ,  $q_{ij} = N_i$  and  $\phi_{ij}$ ,  $s_{ij}(Z_i)$ ;  $W_i^*$  is defined as follows:

$$W_i^* = \arg \min_{W_i \in R^{l_i}} \left[ \sup_{Z_i \in \Omega_{Z_i}} |W_i^T S_i(Z_i) - E_i(Z_i)| \right] \quad (18)$$

the reconstruction error  $\delta_i(Z_i)$  satisfies  $|\delta_i(Z_i)| \leq \varepsilon_i$ ,  $\forall Z_i \in \Omega_{Z_i}$ , here  $\varepsilon_i > 0$  is unknown.

The virtual control laws of  $\alpha_i$  and the adaptation laws of  $\hat{\lambda}_i$  are designed as follows:

$$\alpha_i = -\frac{\hat{\lambda}_i}{2\eta_i^2} \|S_i(X_i)\|^2 z_i - k_i z_i \quad (19)$$

$$\dot{\hat{\lambda}}_i = \frac{\gamma_i}{2\eta_i^2} \|S_i(X_i)\|^2 z_i^2 - \sigma_i \hat{\lambda}_i \quad (20)$$

where  $k_i, \eta_i, \sigma_i$  and  $\gamma_i$  are some known constants, and  $k_1, \dots, k_{n-1} > \frac{1}{2}$ ,  $k_n > \frac{1}{2b_n}$ ,  $\hat{\lambda}_i$  is the estimate of  $\lambda_i^*$ ,  $\tilde{\lambda}_i = \lambda_i^* - \hat{\lambda}_i$ ,  $\lambda_i^* = b_{\min}^{-1} \|W_i^*\|^2$ ,  $i = 1, 2, \dots, n$ ,  $b_{\min} = \min\{1, b_n\}$ ,  $X_i \in R^{i+2}$  will be given later. It should be pointed out that all the components of  $X_i$  are part of the variable  $Z_i \in R^{2i+4}$ .

**Step 1** ( $i = 1$ ): Let  $V_{z_1} = \frac{z_1^2}{2}$ . According to system (9) and noting  $z_1 = s_1 - \hat{y}_d$ , it yields

$$\dot{V}_{z_1} = z_1 \left[ F_1(\bar{s}_2) + s_2 + D_1(t, \zeta, \bar{s}_n) - \dot{y}_d \right] \quad (21)$$

From Assumption 1 and Lemma 1, we have

$$\begin{aligned} |z_1 D_1(t, \zeta, \bar{s}_n)| &\leq |z_1| \kappa_1(s_1) [\varphi_{11}(|x_1|) + \varphi_{12}(\|\zeta\|)] \\ &\leq |z_1| \kappa_1(s_1) [\varphi_{11}(|s_1|) + \varphi_{12}(\bar{\alpha}_1^{-1}(r + D_0))] \\ &\leq z_1^2 \kappa_1^2(s_1) \psi_1(s_1, r) + \frac{1}{2} \end{aligned} \quad (22)$$

where  $\psi_1(s_1, r) = \varphi_{11}^2(|x_1|) + \varphi_{12}^2(\bar{\alpha}_1^{-1}(r + D_0))$ .

Substituting (22) into (21), we obtain

$$\begin{aligned} \dot{V}_{z_1} &\leq z_1 [F_1(\bar{s}_2) + s_2 + z_1 \kappa_1(s_1)^2 \psi_1(s_1, r) - \dot{y}_d] + \frac{1}{2} \\ &\leq z_1 E_1(Z_1) - \frac{1}{2} z_1^2 z_2^2 + z_1 (z_2 + \alpha_1) + \frac{1}{2} \end{aligned} \quad (23)$$

where

$$E_1(Z_1) = F_1(\bar{s}_2) + z_1 \kappa_1^2(s_1) \psi_1(s_1, r) - \dot{y}_d + \frac{1}{2} z_1 z_2^2 \quad (24)$$

$$Z_1 = [s_1, s_2, \hat{y}_d, \dot{\hat{y}}_d, r]^T \in R^5 \quad (25)$$

From (9) and (21), we have

$$\dot{V}_{z_1} \leq W_1^{*T} S_1(Z_1) z_1 + \delta_1(Z_1) z_1 + z_1 \alpha_1 + 1 \quad (26)$$

Let

$$V_1 = V_{z_1} + \frac{b_{\min}}{2\gamma_1} \tilde{\lambda}_1^2 \quad (27)$$

Using Young's inequality and the property of Gaussian function, one yields

$$W_1^{*T} S(Z_1) z_1 \leq \frac{b_{\min} \lambda_1^*}{2\eta_1^2} \|S_1(X_1)\|^2 z_1^2 + \frac{\eta_1^2}{2}$$

where

$$X_1 = [s_1, \hat{y}_d, r]^T \in R^3 \quad (28)$$

Differentiating  $V_1$  for  $t$  in (27), we have

$$\begin{aligned} \dot{V}_1 &\leq \frac{b_{\min} \lambda_1^*}{2\eta_1^2} \|S_1(X_1)\|^2 z_1^2 + \delta_1(Z_1) z_1 \\ &\quad + z_1 \alpha_1 - \frac{1}{\gamma_1} \tilde{\lambda}_1 \dot{\lambda}_1 + \frac{\eta_1^2}{2} + 1 \end{aligned} \quad (29)$$

Noting (19), we have

$$z_1 \alpha_1 \leq -k_1 z_1^2 - \frac{b_{\min} \hat{\lambda}_1}{2\eta_1^2} \|S_1(X_1)\|^2 z_1^2 \quad (30)$$

Substituting (30) and (20) into (29), we obtain

$$\dot{V}_1 \leq -k_1 z_1^2 + \delta_1(Z_1) z_1 + \frac{b_{\min} \sigma_1 \tilde{\lambda}_1}{\gamma_1} \hat{\lambda}_1 + \frac{\eta_1^2}{2} + 1 \quad (31)$$

Due to  $|\delta_1(Z_1) z_1| \leq \frac{z_1^2}{2} + \frac{\varepsilon_1^2}{2}$  and  $\frac{\sigma_1 \tilde{\lambda}_1}{\gamma_1} \hat{\lambda}_1 \leq -\frac{\sigma_1 \tilde{\lambda}_1^2}{2\gamma_1} + \frac{\sigma_1 \lambda_1^{*2}}{2\gamma_1}$ , it yields

$$\dot{V}_1 \leq -(k_1 - \frac{1}{2}) z_1^2 - \frac{b_{\min} \sigma_1 \tilde{\lambda}_1^2}{2\gamma_1} + \mu_{10} \quad (32)$$

where  $\mu_{10} = \frac{\varepsilon_1^2}{2} + \frac{b_{\min} \sigma_1 \lambda_1^{*2}}{2\gamma_1} + \frac{\eta_1^2}{2} + 1$ .

**Step i** ( $n-1 \geq i \geq 2$ ): Let  $V_{z_i} = \frac{z_i^2}{2}$ . According to system (9) and utilizing  $z_i = s_i - \alpha_{i-1}$ , we get

$$\begin{aligned} \dot{V}_{z_i} &= z_i \left[ F_i(\bar{s}_{i+1}) + s_{i+1} + D_i(t, \zeta, \bar{s}_n) - \dot{\alpha}_{i-1} \right] \\ &= z_i s_{i+1} + z_i \left[ F_i(\bar{s}_{i+1}) + D_i(t, \zeta, \bar{s}_n) \right. \\ &\quad \left. - \sum_{j=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial s_j} D_j(t, \zeta, \bar{s}_n) - W_{i-1} \right] \end{aligned} \quad (33)$$

where  $\dot{\alpha}_{i-1} = W_{i-1} + \sum_{j=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial s_j} D_j(t, \zeta, \bar{s}_n)$ ,

$$W_{i-1} = \sum_{j=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial s_j} \left( F_j(\bar{s}_{j+1}) + s_{j+1} \right) + \omega_{i-1} \quad (34)$$

$$\omega_{i-1} = \sum_{j=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial \hat{\lambda}_j} \dot{\hat{\lambda}}_j + \frac{\partial \alpha_{i-1}}{\partial \bar{y}_{d,i-1}^T} \dot{\bar{y}}_{d,i-1} + \frac{\partial \alpha_{i-1}}{\partial r} \dot{r} \quad (35)$$

where  $\bar{y}_{di} = [\hat{y}_d, \dot{\hat{y}}_d, \dots, \hat{y}_d^{(i)}]^T$ .

From Lemma 1, we have

$$\begin{aligned} |z_i D_i(t, \zeta, \bar{s}_n)| &\leq |z_i| \kappa_i(s_i) [\varphi_{i1}(\|\bar{x}_i\|) + \varphi_{i2}(\|\zeta\|)] \\ &\leq z_i^2 \kappa_i^2(s_i) \psi_i(\bar{s}_i, r) + \frac{1}{2} \end{aligned} \quad (36)$$

where

$$\psi_i(\bar{s}_i, r) = \varphi_{i1}^2(\|\bar{x}_i\|) + \varphi_{i2}^2(\bar{\alpha}_1^{-1}(r + D_0)) \quad (37)$$

Similarly, we have

$$|z_i \frac{\partial \alpha_{i-1}}{\partial s_j} D_j(t, \zeta, \bar{s}_n)| \leq z_i^2 \left[ \frac{\partial \alpha_{i-1}}{\partial s_j} \right]^2 \kappa_j^2(s_j) \psi_j(\bar{s}_j, r) + \frac{1}{2}, \quad j = 1, \dots, i-1 \quad (38)$$

From (36), we have

$$-z_i \sum_{j=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial s_j} D_j(t, \zeta, \bar{s}_n) \leq z_i^2 \sum_{j=1}^{i-1} \left( \frac{\partial \alpha_{i-1}}{\partial s_j} \right)^2 \kappa_j^2(s_j) \psi_j(\bar{s}_j, r) + \frac{i-1}{2} \quad (39)$$

Substituting (36) and (39) into (33), we obtain

$$\begin{aligned} \dot{V}_{z_i} &\leq z_i \left[ F_i(\bar{s}_{i+1}) + z_i \kappa_i^2(s_i) \psi_i(\bar{s}_i, r) \right. \\ &\quad \left. + z_i \sum_{j=1}^{i-1} \left( \frac{\partial \alpha_{i-1}}{\partial s_j} \right)^2 \kappa_j^2(s_j) \psi_j(\bar{s}_j, r) - W_{i-1} \right] + z_i s_{i+1} + \frac{i}{2} \end{aligned} \quad (40)$$

Furthermore, we have

$$\dot{V}_{z_i} \leq z_i E_i(Z_i) + z_i s_{i+1} - \frac{1}{2} z_i z_{i+1}^2 + \frac{i}{2} \quad (41)$$

where

$$\begin{aligned} E_i(Z_i) &= F_i(\bar{s}_{i+1}) + z_i \sum_{j=1}^{i-1} \left( \frac{\partial \alpha_{i-1}}{\partial s_j} \right)^2 \kappa_j^2(s_j) \psi_j(\bar{s}_j, r) \\ &\quad + z_i \kappa_i^2(s_i) \psi_i(\bar{s}_i, r) - W_{i-1} + \frac{1}{2} z_i z_{i+1}^2 \end{aligned} \quad (42)$$

$$Z_i = [\bar{s}_{i+1}^T, \alpha_{i-1}, \frac{\partial \alpha_{i-1}}{\partial \bar{s}_i^T}, \omega_{i-1}, r]^T \in R^{2i+4} \quad (43)$$

From (9) and (41), we obtain

$$\begin{aligned}\dot{V}_{z_i} &\leq W_i^{*T} S_i(Z_i) z_i + \delta_i(X_i) z_i + z_i \alpha_i + \frac{i+1}{2} \\ &\leq \frac{b_{\min} \lambda_i^*}{2 \eta_i^2} \|S_i(X_i)\|^2 z_i^2 + \frac{\eta_i^2}{2} + \frac{\varepsilon^2}{2} + \frac{1}{2} z_i^2 + z_i \alpha_i + \frac{i+1}{2}\end{aligned}\quad (44)$$

where

$$X_i = [\bar{s}_i^T, \alpha_{i-1}, r]^T \in R^{i+2} \quad (45)$$

Let

$$V_i = V_{z_i} + \frac{b_{\min}}{2 \gamma_i} \tilde{\lambda}_i^2$$

Substituting (19) and (20) into (44), we obtain

$$\dot{V}_i \leq -(k_i - \frac{1}{2}) z_i^2 - \frac{b_{\min} \sigma_i \tilde{\lambda}_i^2}{2 \gamma_i} + \mu_{i0} \quad (46)$$

where  $\mu_{i0} = \frac{\varepsilon_i^2}{2} + \frac{b_{\min} \sigma_i \lambda_i^{*2}}{2 \gamma_i} + \frac{\eta_i^2}{2} + \frac{i+1}{2}$ .

**Step n:** Let  $V_n = V_{z_n} + \frac{b_{\min}}{2 \gamma_n} \tilde{\lambda}_n^2$ , where  $V_{z_n} = \frac{z_n^2}{2}$ . From system (9) and using  $z_n = s_n - \alpha_{n-1}$ .

Design the control law as follows:

$$u = -\frac{1}{\kappa_n(s_n)} \left[ k_n z_n + \frac{\hat{\lambda}_n}{2 \eta_n^2} \|S_n(X_n)\|^2 z_n \right] \quad (47)$$

where  $Z_n = [\bar{s}_n^T, \alpha_{n-1}, \frac{\partial \alpha_{n-1}}{\partial \bar{s}_n^T}, \omega_{n-1}, r]^T \in R^{2n+3}$ ,  $X_n = [\bar{s}_n^T, \alpha_{n-1}, r]^T \in R^{n+2}$ ,  $\alpha_{n-1}$ ,  $\hat{\lambda}_n$  and  $\omega_{n-1}$  are determined by (19), (20) and (35) for  $i = n$ ,  $k_n > \frac{1}{2 b_n}$  and  $\eta_n > 0$ .

$$\begin{aligned}\dot{V}_{z_n} &= z_n \left[ F_n(\bar{s}_n) + \kappa_n(s_n) G_n(\bar{s}_n) u + D_n(t, \zeta, \bar{s}_n) \right. \\ &\quad \left. - \sum_{j=1}^{n-1} \frac{\partial \alpha_{n-1}}{\partial s_j} (D_j(t, \zeta, \bar{s}_n)) - W_{n-1} \right]\end{aligned}\quad (48)$$

where  $W_{n-1}$  is given by (34) for  $i = n$ . Similarly, we get

$$\dot{V}_n \leq -(b_n k_n - \frac{1}{2}) z_n^2 - \frac{b_{\min} \sigma_n \tilde{\lambda}_n^2}{2 \gamma_n} + \mu_{n0} \quad (49)$$

where  $\mu_{n0} = \frac{\varepsilon_n^2}{2} + \frac{b_{\min} \sigma_n \lambda_n^{*2}}{2 \gamma_n} + \frac{\eta_n^2}{2} + \frac{n+1}{2}$ .

## 4 Main Results

**Theorem 1.** If Assumptions 1–4 hold for system (1), the nonlinear mapping is given by (6), and the control law is determined by (47), and the updating laws are determined by (19) and (20), then for bounded initial conditions, the following properties hold.

- (i) All signals in the whole system are bounded under  $\hat{\lambda}_i(0) \geq 0, i = 1, \dots, n$ .
- (ii)  $z \in \Omega_z = \{z \mid z_1^2 + z_2^2 + \dots + z_n^2 \leq A_z\}$ ,  
 $\hat{\lambda} \in \Omega_{\hat{\lambda}} = \{\hat{\lambda} \mid \hat{\lambda}_1^2 + \hat{\lambda}_2^2 + \dots + \hat{\lambda}_n^2 \leq A_{\hat{\lambda}}\}$ ,  
where  $z = [z_1, \dots, z_n]^T$ ,  $\hat{\lambda} = [\hat{\lambda}_1, \dots, \hat{\lambda}_n]^T$ ,  $A_z$  and  $A_{\hat{\lambda}}$  are two positive constants.
- (iii) Every state  $x_i \in \Omega_{x_i}$  is never triggered.

*Proof.* To save space, the proof is omitted.

## 5 Conclusions

By introducing invertible nonlinear change, the uncertain affine system with whole state restrictions is transformed into a novel uncertain unconstrained nonaffine system. Using modified backstepping design and the characteristic of Gaussian function, a novel neural control method with simpler structure has been presented for the transformed uncertain system. By the aid of Young's inequality, only one adjusting parameter is updated online for the approximated function at each step of the recursion. By theoretical analysis, all signals involved in the whole system are bounded, and the whole state restrictions are never triggered.

**Acknowledgements.** This work was partially supported by the National Natural Science Foundation of China (61573307), the Natural Science Foundation of Jiangsu Province (BK20181218) and Yangzhou University Top-level Talents Support Program (2016).

## References

1. Jiang, Z.P., Praly, L.: Design of robust adaptive controllers for nonlinear systems with dynamic uncertainties. *Automatica* **34**(7), 825–840 (1998). [https://doi.org/10.1016/S0005-1098\(98\)00018-1](https://doi.org/10.1016/S0005-1098(98)00018-1)
2. Jiang, Z.P., Hill, D.J.: A robust adaptive backstepping scheme for nonlinear systems with unmodeled dynamics. *IEEE Trans. Autom. Control* **44**(9), 1705–1711 (1999). [https://doi.org/0018-9286\(99\)07141-X](https://doi.org/0018-9286(99)07141-X)
3. Jiang, Z.P.: A combined backstepping and small-gain approach to adaptive output feedback control. *Automatica* **35**(6), 1131–1139 (1999). [https://doi.org/0005-1098\(99\)00015-1](https://doi.org/0005-1098(99)00015-1)
4. Zhang, X.Y., Lin, Y.: Adaptive tracking control for a class of pure-feedback nonlinear systems including actuator hysteresis and dynamic uncertainties. *IET Control Theory Appl.* **5**(16), 1868–1880 (2011). <https://doi.org/10.1049/iet-cta.2010.0711>
5. Zhang, T.P., Xia, X.N.: Decentralized adaptive fuzzy output feedback control of stochastic nonlinear large-scale systems with dynamic uncertainties. *Inf. Sci.* **315**, 17–38 (2015). <https://doi.org/10.1016/j.ins.2015.04.002>
6. Zhang, T.P., Xia, X.N.: Adaptive output feedback tracking control of stochastic nonlinear systems with dynamic uncertainties. *Int. J. Robust Nonlinear Control* **25**(9), 1282–1300 (2015). <https://doi.org/10.1002/rnc.3139>
7. Krstic, M., Sun, J., Kokotovic, P.V.: Robust control of nonlinear systems with input unmodeled dynamics. *IEEE Trans. Autom. Control* **41**(6):913–920 (1996). [https://doi.org/0018-9286\(96\)02823-1](https://doi.org/0018-9286(96)02823-1)

8. Arcak, M., Kokotovic, P.V.: Robust nonlinear control of systems with input unmodeled dynamics. *Syst. Control Lett.* **41**(2), 115–122 (2000). [https://doi.org/10.167-6911\(00\)00044-X](https://doi.org/10.167-6911(00)00044-X)
9. Xia, X.N., Zhang, T.P., Yi, Y., Shen, Q.K.: Adaptive prescribed performance control of output feedback systems including input unmodeled dynamics. *Neurocomputing* **190**, 226–236 (2016). <https://doi.org/10.1016/j.neucom.2016.01.014>
10. Tee, K.P., Ge, S.S., Tay, E.H.: Barrier Lyapunov functions for the control of output-constrained nonlinear systems. *Automatica* **45**(4), 918–927 (2009). <https://doi.org/10.1016/j.automatica.2008.11.017>
11. Ren, B.B., Tee, K.P., Lee, T.H.: Adaptive neural control for output feedback nonlinear systems using a barrier Lyapunov function. *IEEE Trans. Neural Netw.* **21**(8), 1339–1345 (2010). <https://doi.org/10.1109/TNN.2010.2047115>
12. Tee, K.P., Ren, B.B., Ge, S.S.: Control of nonlinear systems with time-varying output constraints. *Automatica* **47**(11), 2511–2516 (2011). <https://doi.org/10.1016/j.automatica.2011.08.044>
13. Qiu, Y.N., Liang, X.G., Dai, Z.Y., Cao, J.X., Chen, Y.Q.: Backstepping dynamic surface control for a class of nonlinear systems with time-varying output constraints. *IET Control Theory Appl.* **9**(15), 2312–2319 (2015). <https://doi.org/10.1049/iet-cta.2015.0019>
14. Liu, Y.J., Tong, S.C., Philip Chen, C.L.: Neural network control-based adaptive learning design for nonlinear systems with full-state constraints. *IEEE Trans. Neural Netw. Learn Syst.* **27**(7), 1562–1571 (2016). <https://doi.org/10.1109/TNNLS.2015.2508926>
15. Liu, Y.J., Tong, S.C.: Barrier Lyapunov functions-based adaptive control for a class of nonlinear pure feedback systems with full state constraints. *Automatica* **64**, 70–75 (2016). <https://doi.org/10.1016/j.automatica.2015.10.034>
16. Guo, T., Wu, X.W.: Backstepping control for output-constrained nonlinear systems based on nonlinear mapping. *Neural Comput. Appl.* **25**(7–8), 1665–1674 (2014). <https://doi.org/10.1007/s00521-014-1650-9>
17. Yin, S., Yu, H., Shahnazi, R., Haghani, A.: Fuzzy adaptive tracking control of constrained nonlinear switched stochastic pure-feedback systems. *IEEE Trans. Cybern.* **47**(3), 579–588 (2017). <https://doi.org/10.1109/TCYB.2016.2521179>
18. Meng, W.C., Yang, Q.M., Si, S.N., Sun, Y.X.: Adaptive neural control of a class of output-constrained non-affine systems. *IEEE Trans. Cybern.* **46**(1), 85–95 (2016). <https://doi.org/10.1109/TCYB.2015.2394797>
19. Zhang, T.P., Xia, M.Z., Yi, Y.: Adaptive neural dynamic surface control of strict-feedback nonlinear systems with full state constraints and unmodeled dynamics. *Automatica* **81**, 232–239 (2017). <https://doi.org/10.1016/j.automatica.2017.03.033>
20. Zhang, T.P., Xia, M.Z., Yi, Y., Shen, Q.K.: Adaptive neural dynamic surface control of pure-feedback nonlinear systems with full state constraints and dynamic uncertainties. *IEEE Trans Syst. Man Cybern. Syst.* **47**(8), 2378–2387 (2017). <https://doi.org/10.1109/TSMC.2017.2675540>
21. Kim, B.S., Yoo, S.G.: Approximation-based adaptive control of uncertain nonlinear pure-feedback systems with full state constraints. *IET Control Theory Appl.* **8**(17), 2070–2081 (2014). <https://doi.org/10.1049/iet-cta.2014.0254>
22. Ge, S.S., Hang, C.C., Lee, T.H., Zhang, T.: Stable Adaptive Neural Network Control. Kluwer Academic, Boston (2001)



# A Music Generation Model Based on Generative Adversarial Networks with Bayesian Optimization

Yijie Xu<sup>1</sup>, Xueqing Yang<sup>2(✉)</sup>, Yiming Gan<sup>3(✉)</sup>, Wuneng Zhou<sup>4</sup>,  
Hangyang Cheng<sup>1</sup>, and Xuehui He<sup>1</sup>

<sup>1</sup> College of Information Science and Technology, Donghua University,  
Shanghai 201620, China

<sup>2</sup> Educational Technology Center, Donghua University, Shanghai 200051, China  
[etdaqing@163.com](mailto:etdaqing@163.com)

<sup>3</sup> Guangdong Polytechnic College, Foshan 528041, China  
[ygan@ultraliv.com](mailto:ygan@ultraliv.com)

<sup>4</sup> Engineering Research Center of Digitized Textile and Fashion Technology,  
Donghua University, Shanghai, China

**Abstract.** In recent years, a huge number of neural networks have been applied in music generation, many of which use generative adversarial networks (GAN). In this paper, a novel melody generation framework is proposed to create motivation for composers, which contains a generator made by bidirectional long short-term memory (Bi-LSTM) and a discriminator made by long short-term memory (LSTM). We change the traditional optimization policy of GAN by bringing Bayesian optimization in our model. In last, we conduct a user study that show better preference of our generated melodies over that produced by several recent other music generation models.

**Keywords:** Algorithmic composition · Generative adversarial networks · Bayesian optimization

## 1 Introduction

Deep learning has revolutionized artistic domain for a long time, and music is born to be suitable for neural networks to learn due to its inseparable relationship with mathematics. Algorithmic composition is an academic study that aims to complete a stylized process that minimizes the degree of involvement by person when composing [1]. Goodfellow proposed generative adversarial networks (GAN) in 2014 [2]. GAN trains a generator and a discriminator to promote each other in the two-person zero-sum game. In 2016, for the first time, C-RNN-GAN applied GAN in algorithmic composition to generate classical music [3].

In this paper, the proposed model changes the optimization of traditional GAN, we presented a deep time series prediction model based on bidirectional

long short-term memory (Bi-LSTM) as a generator, which means let the computer learn and summarize the music theory. During the training process, we use Bayesian optimization to indicate the training process.

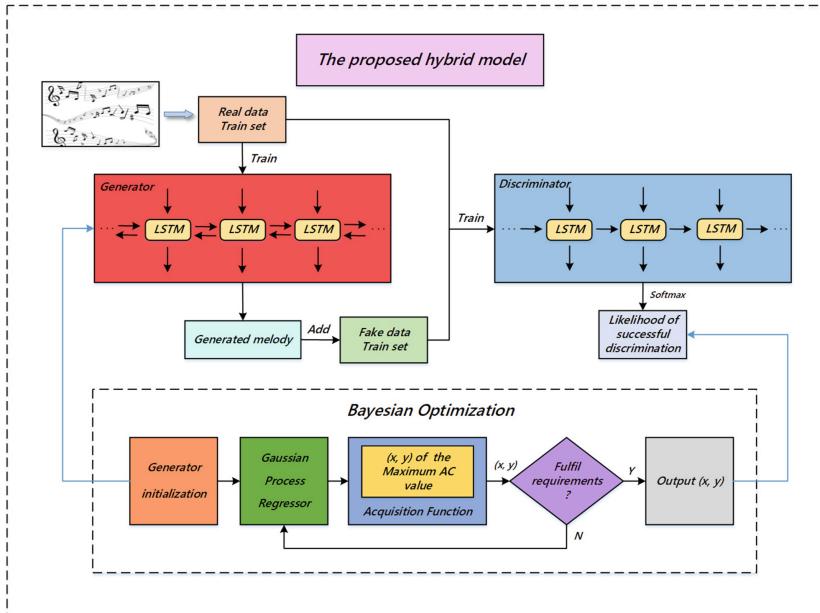
Our research aims to generate a short melody section based on a short pop music phrase. A short melody section can become a motivation for human composition and can also provide inspiration for musicians. The core of this paper is to generate notes in a predictive way, so we refer to this new model as the ForecastGAN.

The principal contributions of this paper are as follows.

- (1) We propose a music data modeling and symbol representation method through processing XML format files;
- (2) We propose a melody generation model based on generative adversarial networks, but we change the adversarial training strategy by using Bayesian optimization;
- (3) In addition to the traditional user study, our experiments conduct another new evaluation policy: calculating the proportion of generated notes in the same tuning scale.

## 2 Models

In this section, we will introduce our model.



**Fig. 1.** Framework of the proposed model.

As showed in Figure 1, the technical details of each major component of the proposed model are follows.

## 2.1 Generator Bi-LSTM and Discriminator LSTM

Long short-term memory (LSTM) contains input gates, forget gates and output gates, which are used as interfaces for information propagation within the networks [4, 10, 11]. Besides, LSTM adds memory cells into the model so that it is able to decide whether it is necessary to memorize information.

The formula of LSTM is presented as follows

$$i_t = \sigma (W_i \cdot [\omega_t, h_{t-1}] + b_i), \quad (1)$$

$$f_t = \sigma (W_f \cdot [\omega_t, h_{t-1}] + b_f), \quad (2)$$

$$o_t = \sigma (W_o \cdot [\omega_t, h_{t-1}] + b_o), \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot [\omega_t, h_{t-1}] + b_c), \quad (4)$$

$$h_t = o_t \odot \tanh(c_t), \quad (5)$$

$c_t$  is the current LSTM cell state,  $h_t$  is the current hidden layer state,  $i_t$ ,  $f_t$ ,  $o_t$  denote input gates, forgetting gates and output gates respectively.  $\sigma$  is the sigmoid function, which is used to control the input and output information for each iteration.  $\{W_i, W_f, W_o, W_c, b_i, b_f, b_o, b_c\}$  are the parameters to be learned during training, where the dimension of each value is equal to that of word vector.

The discriminator learns to discriminate between real data and fake data that generated by the computer, and the generator learns to generate the melody that can fool the discriminator. We use a typical two-class LSTM network as the discriminator, but we choose Bi-LSTM as the generator. Because Bi-LSTM has more excellent performance in utilizing future information. The basic principle of Bi-LSTM is that the data of the input layer is calculated in the forward and backward directions [5]. The Bi-LSTM network update formula can be computed as

$$\vec{h}_t = H \cdot (W_1 \cdot x_t + W_2 \cdot \vec{h}_{t-1} + \vec{b}), \quad (6)$$

$$\overleftarrow{h}_t = H \cdot (W_3 \cdot x_t + W_5 \cdot \overrightarrow{h}_{t-1} + \overleftarrow{b}), \quad (7)$$

$$y_t = W_4 \cdot \vec{h}_t + W_6 \cdot \overleftarrow{h}_t + b_y, \quad (8)$$

the six unique weights are reused at each time step, and the six weights correspond to: input to the forward and backward hidden layers ( $W_1$ ,  $W_3$ ), and hidden layers to the hidden layer itself ( $W_2$ ,  $W_5$ ), forward and backward hidden layer to output layer ( $W_4$ ,  $W_6$ ).

## 2.2 Bayesian Optimization and Adversarial Training Policy

Hyperparameters can not learned from the training process directly because they are parameters of the algorithm itself. Manual parameter settings are not only inefficient, but are always affected by human bias. Another reason we choose Bayesian optimization to update the generator is the special characteristic of the music generation task, the difficulty in evaluating the quality of generated samples makes our GAN difficult to train [6].

The basic method of Bayesian optimization is to estimate the posterior distribution of the objective function from the data, and then select next sample's hyperparameters combination according to the distribution with Bayes' theorem. It takes advantage of the information from previous sample points, then it optimizes by analysing the shape of the objective function and adjusting the parameters that minimize the result to the global minimum [7]. The application scenario of Bayesian optimization in this paper is

$$x^* = \arg \min_{x \in S} (1 - D(G(z))), \quad (9)$$

in the formula,  $S$  is a candidate set of  $x$ , and the optimization goal is to select an  $x$  from  $S$  such that the value of  $D(G(z))$  will be the largest. The specific formula of  $D(G(z))$  is unknown, and is equivalent to a black box function.

Bayesian optimization need to tradeoff the exploration and exploitation for the purpose of avoiding the local optima. Our acquisition function chooses expected improvement (EI) which provides a single measure of the usefulness of trying any given point

$$EI(x) = \begin{cases} (\mu(x) - f(x^+)) \cdot \phi(Z) + \sigma(x) \cdot \phi(Z), & \sigma(x) > 0, \\ 0, & \sigma(x) = 0, \end{cases} \quad (10)$$

$$Z = \frac{\mu(x) - f(x^+)}{\sigma(x)}. \quad (11)$$

As shown in the Fig. 1, in each iteration, the generator will be trained to update its parameters separately. Then generate a sequence and input it into the trained discriminator to get a probability, that is, the possibility of discriminating is fake data. After that, adjust the hyperparameters of the generator in the Bayesian optimization iteration continuously, and the optimization goal is to minimize the probability of the output.

So far, we have strengthened the generator, but this process does not include the idea of adversarial training. So we propose one another kind of strategy to strengthen the discriminator: in each iteration, we add the samples generated in the previous round into the discriminator's training set, so that the performance of identifying real and fake data can be improved.

## 3 Experiment

The model in this paper is implemented by Tensorflow. We train the generator 128 batches each time, the learning rate and the number of neurons of the hidden

layer are selected by Bayesian optimization, and the initial values are set to 0.001 and 100. For the discriminator, we set 64 neurons in the hidden layer, and the learning rate is 0.001. In the training phase, we tag 420 real samples as 1, and randomly generate 579 time series with the label be tagged as 0. In the column of “note length”, the smallest value is 0.25 that represents the a quarter beat, so we randomly generate an integer between [1 ,16] and multiply it by 0.25; in the column of the “scale degree”, we just need to generate the value among [0, 107] randomly. Therefore, the fake data is equivalent to a combination of notes that are disorder, which is enough to distinguish it from the real melodies.

### 3.1 Dataset Preprocessing

This paper presents a creative music data manipulation method through processing XML format files, instead of converting the note events in the MIDI file into one-hot matrix, which solves the problem that MidiNet cannot distinguish one long note from two repeated short notes [8]. We crawled the training data in XML format required by the experiment from TheoryTab<sup>1</sup>, each XML file contains information such as song titles, velocities, and more dimensions, but we only extract the “note length”, “scale degree” and “octave” of the main melody as three features.

The notes in music theory define the basic unit of music composition. Music follows the 12-tone system, where 12 is the period length of all notes. And the 12 tones are: C, C#/Db, D, D#/Eb, E, F, F#/Gb, G, G#/Ab, A, A#/Bb, B. Scale is defined as a subset of notes. For simplicity, we normalize all the data into C major scale. In our time series, use 60–71 to denote C major scale. 60 is the pitch of the central C (C4) in MIDI music, each additional 12, the pitch rises by one octave. So that we can calculate the pitch of a note that is not in the same octave. After preprocessing, we can ignore the value of “octave” and focus on “note length” and “scale degree”. In addition, the pause of the melody in the xml file is expressed in the vacancy of the “scale degree”. We will fill the vacancy value to −100, which symbolizes that the pitch of this time’s melody is infinitely small and close to silence.

We only take a melody of length 30 from each sequence, and remove data with less than 30 notes. After the preprocessing process, we finally get 420 sequences of length 30, which are combined into a two-dimensional time series of length 12600, it is equivalent to a very long C major melody.

### 3.2 User Study

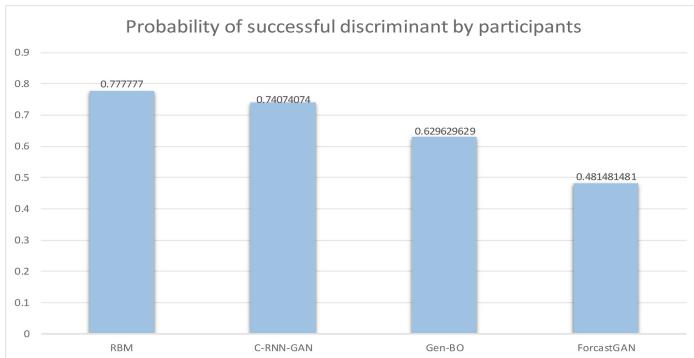
To evaluate the aesthetic quality of our model’s generation results, we conduct a user study with 19 participants. Eight of them understand basic music theory and own the experience of playing instruments or being an amateur musician, so we considered them as people with musical backgrounds, or simply professionals. We consider three baselines: our method versus RBM model [9], our method

---

<sup>1</sup> <https://www.hooktheory.com/theorytab>.

versus C-RNN-GAN, and our method versus only a LSTM forecasting model similar to our generator but without the discriminator and adversarial training (Gen-BO). We randomly picked three melodies out of the several generation results from each of four considered models. To ForecastGAN and the LSTM forecasting model, we randomly selected a different sequence each time from the training set as the priming melodies to guide their generation.

First, participants were asked to judge whether the music was artificial or generated by machine. In order to avoid human bias, we told them that some might be composed by human, although all of the melodies were actually created by models. Followed the mechanism in MidiNet [8], then they were told to rate the samples in terms of three metrics: how pleasing, how realistic, and how interesting, from 1 (low) to 5 (high), we add one extra metric for the participants with musical backgrounds: the possibility of being extended into a complete song by human arrangement as a motivation.

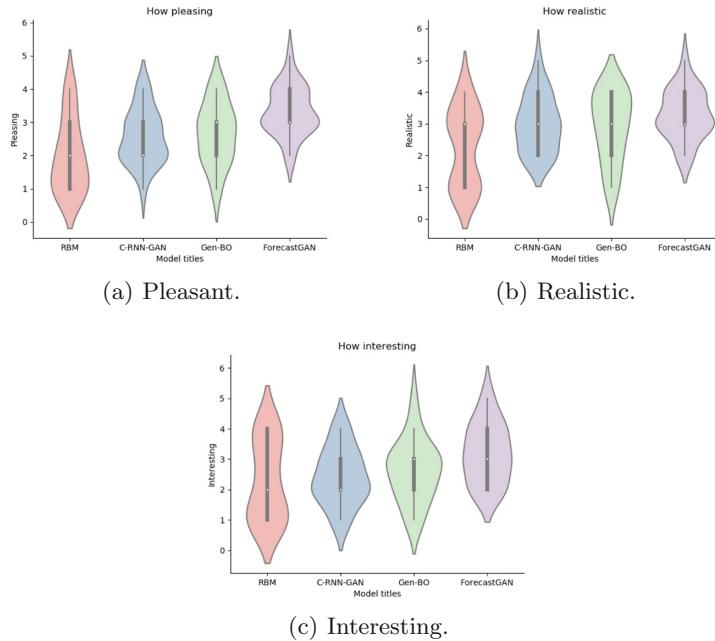


**Fig. 2.** User's listening judgment results.

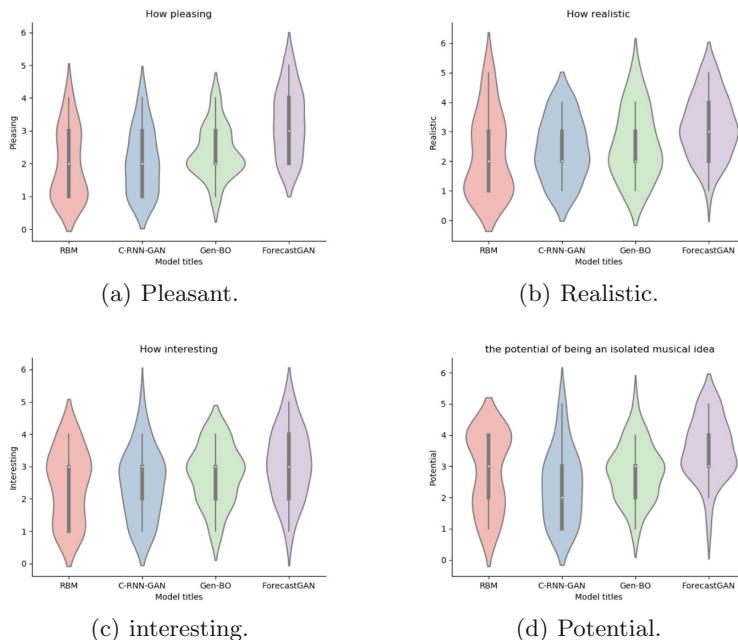
We can see from the Fig. 2 that the participants think that the probability that the generated samples from our two proposed models (ForecastGAN & Gen-BO) are written by humans is higher than the RBM model and C-RNN-GAN. And in contrast, ForecastGAN is further enhanced Gen-BO, showing the positive impact of adding discriminator and adversarial training.

**Table 1.** Models' performance comparison by common participants

Model	Pleasant	Realistic	Interesting
RBM model	2	2.28571	2.28571
C-RNN-GAN	2.57143	3	2.42857
Gen-BO	2.66667	2.90476	2.61905
ForecastGAN	3.33333	3.28571	3.19048



**Fig. 3.** Evaluation results from the common participants.



**Fig. 4.** Evaluation results from the professional participants.

**Table 2.** Models' performance comparison by professional participants

Model	Pleasant	Realistic	Interesting	Potential
RBM model	2	2.27273	2.45455	2.81818
C-RNN-GAN	2.09091	2.42424	2.48485	2.39394
Gen-BO	2.36364	2.57576	2.75758	2.818818
ForecastGAN	3.27273	3.18182	3.09091	3.54545

The five-point Likert scale can be used to measure some of the multi-dimensional complex concepts or attitudes that other scales cannot measure. Figure 3, 4 and Table 1, 2 show the results given by the common participants and the professional participants. Similarly, ForecastGAN is much better than baselines, both in terms of the mean values of being pleasant, realistic, interesting, and the potential of being a musical motivation.

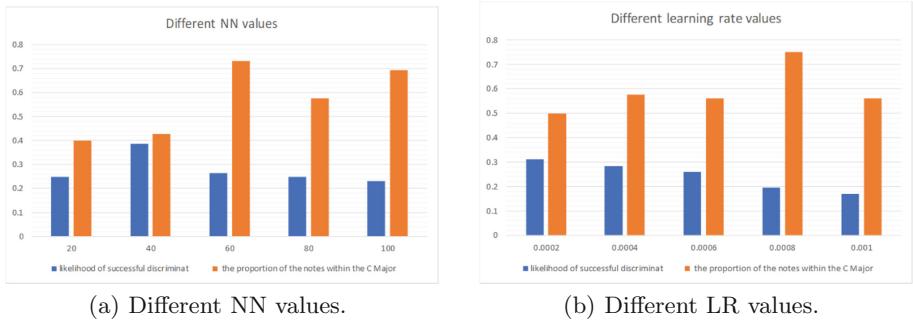
However, in our user study, the participants also reflected some of the problems. First, although some notes are indeed in the C Major, but the octave span is too large, which is not in line with human singing habits during composition. Those melodies sound a bit weird, not aesthetically comparable to mature human works; Second, there are some results that sound very similar to their priming melody, even almost a variation of the original song. Strictly speaking, these results can only be regarded as a simple imitation of human works in the prediction of the model, and the innovation is not enough.

### 3.3 Parameters' Case Study and a Novel Indicator

As mentioned above, the enhancement of the generator in the adversarial training is achieved by means of Bayesian optimization. In this section, neurons number (NN), learning rate (LR) in Bi-LSTM are selected as adjustment parameters to test ForecastGAN's performance.

First, the probability of fake data output by the discriminator can be used as a specific indicator to evaluate models. Secondly, we propose a new evaluation parameter: proportion of generated notes in the same tuning scale. Since we normalize all the training data to C major, the ratio of the notes in C major scale to the total number of notes in the generated melody can be used to assess the degree to which the melody follows music theory. The ratio does not need to be 100%, but if there are too many notes outside the scale, the melody will not sound pleasant absolutely.

It can be seen from the Fig. 5 that different parameter choices have a huge impact on melody's quality. If the value of NN is too small, the proportion of the generated notes within the C Major will not be very high, indicating that the effect of model learning is not particularly good. When the values of neurons number are 60 and 100, the proportion is higher than the other initial values' results, and at the same time the probability of successful discriminant is almost



**Fig. 5.** Parameters' case study results.

the lowest. As for the learning rate, extremely large values also have a significant impact on the quality of the generated samples.

Since the proportion of the generated notes within the scale can only be used as a fuzzy reference to evaluate the computational results of the model, we still tend to choose the initial value that minimizes the likelihood of successful discriminant, that is, the NN value is set as 100 and the LR value is 0.001.

### 3.4 Listening Impressions

Some generated samples can be listened and downloaded from <https://github.com/JaguarXu/ForecastGAN>.

## 4 Conclusion

In this paper, we have presented ForecastGAN, a melody generation model, effectively trained by the approach based on GAN with Bayesian optimization. Our research show the advantages of ForecastGAN compared to some existing baselines. Although more experiments need to be done, we believe the results are still promising. We have noticed that adversarial training helps the Bi-LSTM generate music that more pleasing and realistic. The novel changes proposed by this paper to the structure of GAN also can be considered as an attempt worthy of reference. However, our work can only generate single-track melodies. For future work, we will study the generation of chords progression. Algorithmic composition calls for a deeper study, our work is still only scratching the surface.

**Acknowledgements.** This work is partially supported by the National Natural Science Foundation of China (No. 61705127). This work is supported by the Natural Science Foundation of Shanghai under grant no. 20ZR1402800.

## References

1. Doornbusch, P.: Gerhard Mierhaus: Algorithmic composition: paradigms of automated music generation. *Comput. Music J.* **34**(3), 70–74 (2010). [https://doi.org/10.1162/comj\\_r\\_00008](https://doi.org/10.1162/comj_r_00008)
2. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets. In: 27th Neural Information Processing Systems, Montreal, pp. 2672–2680 (2014). [arXiv:1406.2661](https://arxiv.org/abs/1406.2661)
3. Mogren, O.: C-RNN-GAN: continuous recurrent neural networks with adversarial training. In: 29th Neural Information Processing Systems, Barcelona (2016). [arXiv:1611.09904](https://arxiv.org/abs/1611.09904)
4. F.A. Gers: Learning to forget: continual prediction with LSTM. In: 9th International Conference on Artificial Neural Networks, IET (1999). <https://doi.org/10.1049/cp:19991218>
5. Chen, T., Xu, R., He, Y., Wang, X.: Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Syst.* **72**, 221–230 (2017). <https://doi.org/10.1016/j.eswa.2016.10.065>
6. Yu, L., Zhang, W., Wang, J., Yu, Y.: SeqGAN: sequence generative adversarial nets with policy gradient. In: 31th AAAI Conference on Artificial Intelligence, San Francisco, vol. 31, pp. 4–9 (2017). [arXiv:1609.05473](https://arxiv.org/abs/1609.05473)
7. Wang, C., Liu, S., Zhu, M.: Bayesian network learning algorithm based on unconstrained optimization and ant colony optimization. *J. Syst. Eng. Electron.* **23**(5), 784–790 (2012). <https://doi.org/10.1109/jsee.2012.00096>
8. Yang, L.C., Chou, S.Y., Yang, Y.H.: MidiNet: a convolutional generative adversarial network for symbolic-domain music generation. In: 20th International Society of Music Information Retrieval Conference, Suzhou (2017). [arxiv.org/abs/1703.10847](https://arxiv.org/abs/1703.10847)
9. Use TensorFlow to generate short sequences of music with a restricted Boltzmann machine. [https://github.com/lISourceII/Music\\_Generator\\_Demo](https://github.com/lISourceII/Music_Generator_Demo)
10. Jia, Y.: Robust control with decoupling performance for steering and traction of 4WS vehicles under velocity-varying motion. *IEEE Trans. Control Syst. Tech.* **8**(3), 554–569 (2000)
11. Jia, Y.: Alternative proofs for improved LMI representations for the analysis and the design of continuous-time systems with polytopic type uncertainty: a predictive approach. *IEEE Trans. Autom. Control* **48**(8), 1413–1416 (2003)



# Interactive Attention and Position-Aware Mechanism for Aspect-Level Sentiment Analysis

Xuehui He<sup>1</sup>, Yiming Gan<sup>1,2(✉)</sup>, Xueqing Yang<sup>3(✉)</sup>, Wuneng Zhou<sup>1,4(✉)</sup>,  
and Yuanhong Ren<sup>1</sup>

<sup>1</sup> College of Information Science and Technology, Donghua University,  
Shanghai 201620, China

[zhouwuneng@163.com](mailto:zhouwuneng@163.com)

<sup>2</sup> Guangdong Polytechnic College, Foshan 528041, China  
[ygan@ultraliv.com](mailto:ygan@ultraliv.com)

<sup>3</sup> Educational Technology Center, Donghua University, Shanghai 200051, China  
[etdaqing@163.com](mailto:etdaqing@163.com)

<sup>4</sup> Engineering Research Center of Digitized Textile and Fashion Technology,  
Donghua University, Shanghai, China

**Abstract.** Aspect-level sentiment classification is a fine-grained work in sentiment analysis with the goal of predicting the sentiment categories of target words in a given text. In this paper, a framework of an interactive attention based on bidirectional LSTM networks and position-aware mechanism for aspect-level sentiment classification has been proposed, in which a fine-grained attention is introduced to capture the intersection between context and aspect words. Meanwhile, the position awareness mechanism is introduced to make the weight distribution of cross attention more reasonable. Experiments on the SemEval-2014 datasets show that our proposed model is always superior to the state-of-the-art methods of all datasets.

**Keywords:** Interactive attention · Position-aware mechanism · Aspect-level sentiment analysis

## 1 Introduction

The sentiment analysis task, as a fundamental work in Natural Language Processing (NLP), has attracted the attention of many researchers. 99% of the research papers on the topic of sentiment analysis have appeared since 2004, making sentiment analysis one of the fastest-growing research areas.

Aspect-level sentiment analysis, as a fine-grained research, aims to extract aspect polarity from the comment texts for a specific aspect word [1–3]. For example, “The staff service is very good and the cake tastes normal, but I am not satisfied with the drinks here.”, for aspect words “service”, “cake” and “drinks”, the corresponding sentiment polarities of which are positive, neutral and negative.

Neural network models are the most commonly-used technique, which has proven to be capable of achieving state-of-art performance in many NLP tasks. Long Short-term Memory Networks (LSTM) [4], which is broadly used in NLP fields, have achieved fruitful results. Tang et al. propose the Target-Dependent LSTM (TD-LSTM) to capture the connection between the target word and its context when generating the sentence representation [5].

Attention mechanisms have achieved the state-of-the-art results in many NLP tasks, particularly in the areas of image processing and machine translation. Recently, attention mechanisms have been widely used in natural language processing [6, 11, 12]. In the aspect-level sentiment analysis task, Wang et al. proposed Attention-based LSTM with Aspect Embedding (ATAE-LSTM), whose key idea is to learn aspect words embedding and let them participate in the calculation of attention weight [7]. Ma et al. proposed Interactive Attention Networks (IAN) to learn contextual attention and target attention [8]. The main idea of IAN is to use interactive modeling of target and context to generate attention weight and well generate the representation of target and context.

Although the above methods have proved their effectiveness, the interaction between aspect words and contextual information has been ignored. In this paper, we proposed interactive attention based on the bidirectional LSTM network and position-aware mechanism framework. We introduce an attention mechanism of position in the LSTM network, which is a position weight function based on the relationship between aspect words and context. Our work can be summarized as:

- 1: A fine-grained interactive attention mechanism is designed to capture the cross-relationship between context vectors and aspect vectors.
- 2: Our model takes full account of the importance of position information, and encodes sentences with position awareness. Position information can improve the accuracy of the model.
- 3: Context information is connected with aspect word information as input embedding vectors. We embed aspect word vector directly into the context information of sentences to improve the importance of aspect word information.

## 2 Models

In this section, we introduce our model architecture.

### 2.1 Word Representation

We assume that a sentence contains  $n$  words and aspect terms contain  $m$  words. We define the sentiment analysis sentence as  $S = \{\omega_s^1, \omega_s^2, \dots, \omega_s^i, \dots, \omega_s^n\}$  and use  $S$  to represents the word embedding vector of the sentence, where the total

length of the sentence is  $n$ , so that the sentence can be represented as a dimension matrix of  $\mathbb{R}^{d_{em} \times n}$ . Define the aspect words as  $A = \{\omega_a^1, \omega_a^2, \dots, \omega_a^i, \dots, \omega_a^m\}$  and use  $A$  to represents the embedding vector of aspect words. The total length of aspect words is  $m$ , so the aspect words can be represented as a dimension matrix of  $\mathbb{R}^{d_{em} \times m}$ .

LSTM, as a unique RNN, has its unique memory unit to replace the traditional RNN neuron hidden layer. In the standard LSTM neural network memory module, it is mainly divided into three unit doors, namely: input gate, output gate and forgetting gate. So LSTM neural network can selectively transfer useful data to the next neuron unit and abandon the worthless information flow. LSTM network can update the state as follows

$$\begin{aligned} i_t &= \sigma(W_i \cdot [\omega_t, h_{t-1}] + b_i), \\ f_t &= \sigma(W_f \cdot [\omega_t, h_{t-1}] + b_f), \\ o_t &= \sigma(W_o \cdot [\omega_t, h_{t-1}] + b_o), \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot [\omega_t, h_{t-1}] + b_c), \\ h_t &= o_t \odot \tanh(c_t), \end{aligned}$$

where  $i_t$ ,  $f_t$ ,  $o_t$  represent input gates, forgetting gates and output gates respectively.  $\sigma$  is the sigmoid function, which is used to control the input and output information for each iteration.  $\{W_i, W_f, W_o, W_c, b_i, b_f, b_o, b_c\}$  are the parameters to be learned during training, where the dimension of each value is equal to that of word vector.

However, the standard LSTM can not capture this relationship-dependent problem, so we use Bi-LSTM to solve this problem. Bi-LSTM processes sequence information from two directions, and obtains context information from forward and backward, respectively. We input the context sentence  $S$  into the Bi-LSTM neural network, we can get two hidden states, namely the forward hidden state  $\overrightarrow{h}_s^i \in \mathbb{R}^{d_h}$  and the backward hidden state  $\overleftarrow{h}_s^i \in \mathbb{R}^{d_h}$ , where  $d_h$  represents the number of cells in the hidden layer. Finally, we can join the forward hidden state and the backward hidden state to get  $h_s^i = [\overrightarrow{h}_s^i, \overleftarrow{h}_s^i]$ , where  $h_s^i \in \mathbb{R}^{2d_h}$ . Similarly, for the aspect term, we can also get its hidden layer state information through Bi-LSTM neural network, where  $h_a^i = [\overrightarrow{h}_a^i, \overleftarrow{h}_a^i]$  and  $h_a^i \in \mathbb{R}^{2d_h}$ .

## 2.2 Position Representation

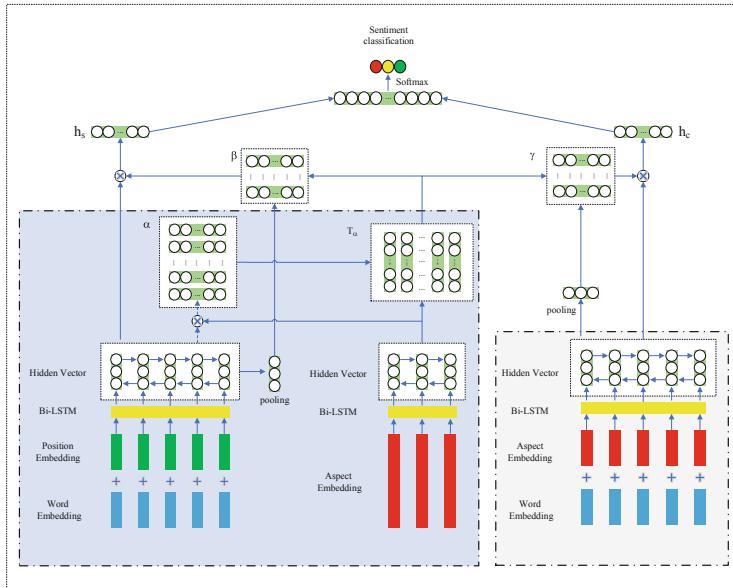
Consequently, to better enhance the position weight information of the aspect statement, we also introduce position awareness to highlight the aspect information. We define a position sequence relative to the aspect term

$$p_i = \begin{cases} i - s_1 & i < s_1 \\ 0 & s_1 < i < s_2 \\ i - s_2 & i > s_2, \end{cases} \quad (9)$$

where  $p_i$  is the relative distance from the  $i$ -th word in  $S$  to the aspect term, and  $s_1$  and  $s_2$  represent the beginning and ending indices of the aspect term. The corresponding positional embedding of  $S$  can be found through the position-aware embedding matrix, the position-aware embedded matrix can be represented as  $\mathbb{R}^{d_p \times N}$  and  $N$  is the total length of  $S$ .

### 2.3 Model

As shown in the Fig. 1, we use the embedded sentence vector combined with position-aware as the first part of the input layer while the aspect word as the second part of the input layer, then the two parts are input to the Bi-LSTM neural network to obtain the hidden vectors  $h_s^i = [h_s^i, h_s^i]$  and  $h_a^i = [h_a^i, h_a^i]$  respectively.



**Fig. 1.** The framework of proposed model

#### Phase 1: Interactive attention stage.

Since the initial sentence  $S$  does not add the position sensing mechanism, it needs to combine the sentence vector embedding of  $S$  and the position-aware embedding. Where the position-aware embedding method is as shown above. After obtaining the hidden vector representation of the aspect words and sentences, we can complete the calculation of the cross attention. The formula derivation process is as follows

$$\alpha_i = \frac{\exp(\tanh(h_s^i \cdot W_S \cdot h_a^i))}{\sum_{j=1}^m \exp(\tanh(h_s^j \cdot W_S \cdot h_a^j))}, \quad (11)$$

where  $\alpha_i$  is the normalized attention weight and  $W_S$  is the weight matrix, and the tanh is a nonlinear activation function. Afterwards, we can obtain  $T_\alpha$ , which is the target attention representation, through a weighted combination of aspect term hidden states

$$T_\alpha = \sum_{i=1}^m \alpha_i \cdot h_a^i. \quad (12)$$

Phase 2: Calculate the position-aware context vector representation.

The attention score of the position-aware context is calculated based on the attention  $T_\alpha$  of the obtained aspect word

$$\bar{h}_s = \frac{1}{n} \sum_{i=1}^n h_s^i, \quad (13)$$

$$\beta_i = \frac{\exp(\tanh(\bar{h}_s \cdot W_A \cdot T_\alpha))}{\sum_{j=1}^n \exp(\tanh(\bar{h}_s \cdot W_A \cdot T_\alpha))}, \quad (15)$$

where  $\bar{h}_s$  is the initial representation of the sentence and  $W_A$  is the weight matrix. After getting the attention score of  $\beta_i$ , the final expected context representation state can be obtained according to the weighted combination of the hidden state of the context

$$h_S = \sum_{i=1}^n \beta_i \cdot h_s^i. \quad (16)$$

Phase 3: Calculating the attention score of context information combined with aspect words.

We embed aspect words vectors directly into the context information of sentences and get  $C = \{\omega_{sa}^1, \omega_{sa}^2, \dots, \omega_{sa}^i, \dots, \omega_{sa}^n\}$ . Then we put  $C$  into the Bi-LSTM neural network to get hidden state  $\overrightarrow{h}_{sa}^i = [\overrightarrow{h}_{sa}^i, \overleftarrow{h}_{sa}^i]$ , and  $h_{sa}^i \in \mathbb{R}^{2d_h}$ . In the hidden state  $h_{sa}^i$  obtained by combining the context information with the aspect words, each item contains information embedded from the aspect words. After getting hidden state  $h_{sa}^i$ , we can use the interactive attention generated by phase 1 to calculate the final vector representation combined by the context information and the aspect words. The process of calculating the attention score of context information combined with aspect words is as follows

$$\bar{h}_{sa} = \frac{1}{n} \sum_{i=1}^n h_{sa}^i a, \quad (17)$$

$$\gamma_i = \frac{\exp(\tanh(\bar{h}_{sa} \cdot W_C \cdot T_\alpha))}{\sum_{j=1}^n \exp(\tanh(\bar{h}_{sa} \cdot W_C \cdot T_\alpha))}, \quad (19)$$

where  $\bar{h}_{sa}$  is the initial representation of the sentence and  $W_C$  is the weight matrix. After getting the attention score of  $\gamma_i$ , the final expected context representation state can be obtained according to the weighted combination of hidden states

$$h_C = \sum_{i=1}^n \gamma_i \cdot h_{sa}^i. \quad (20)$$

Phase 4: sentiment classification.

We add the  $h_s$  and the  $h_c$  together to get the final output hidden state vector

$$h = h_s + h_c = \sum_{i=1}^n \beta_i \cdot h_s^i + \sum_{i=1}^n \gamma_i \cdot h_{sa}^i. \quad (21)$$

then we input  $h$  into the nonlinear layer to get the output sequence representation

$$x = \tanh(W_X \cdot h + b_X), \quad (22)$$

where  $W_X$  is the corresponding weight coefficients and  $b_X$  is the corresponding bias vectors. Finally, we put  $x$  into a linear layer and a *softmax* layer is followed to compute the probability of the labeling sentence belonging to the positive, the negative or the neutral polarity

$$y = \text{softmax}(W_R \cdot x + b_R), \quad (23)$$

where  $W_R$  is the corresponding weight coefficient and  $b_R$  is the corresponding bias vector respectively.

## 2.4 Model Training

In our model architecture,  $y$  is the final sentiment polarity output, which can be represented by an one-hot vector, where  $y$  is the correct sentiment output and  $\hat{y}$  is the sentiment polarity of the predicted content. We use cross entropy as a loss function

$$\text{loss} = - \sum_{i=1}^S y \log(\hat{y}) + \frac{1}{2} \lambda \|\theta\|^2, \quad (24)$$

where  $\lambda$  is a regularization factor and  $\theta$  is used as a parameter set. For example,  $\theta = \{W_{[S,A,C,R]}, BiLSTM\}$ , where *BiLSTM* represents the internal parameters of the LSTM.

## 3 Experiment

### 3.1 Experimental Preparation and Evaluation Method

In our model, the embedding vectors of the words in the experiment are all trained in the glove vector [9]. The vector dimension we used in the experiment is 300. The weight matrix in the model is obtained in a standard uniform distribution  $U(-0.1, 0.1)$ , and the deviations in the model are all set to zero. The experiment uses the accuracy and the F-measure as evaluation criteria.

### 3.2 Datasets

We conduct experiments on the SemEval-2014 Task 4 to evaluate our proposed model, the SemEval-2014 datasets consist of two open source datasets, the “*Laptop*” dataset and the “*Restaurant*” dataset [10]. Table 1 shows the train and test sample numbers for each sentiment polarity.

**Table 1.** Samples of SemEval 2014 Dataset

Domain	Positive		Negative		Neutral	
	Train	Test	Train	Test	Train	Test
Restaurant	2164	728	805	196	633	196
Laptop	987	341	866	128	460	169

**Table 2.** Comparison with baselines. Accuracy and F1-Measure on Three-class prediction about “*Restaurant*” and “*Laptop*” dataset, and Three-class denotes {positive; negative; neutral}.

Method	Restaurant		Laptop	
	Accuracy	F-Measure	Accuracy	F-Measure
LSTM	0.7430	0.5932	0.6724	0.5944
TD-LSTM	0.7643	0.6673	0.6724	0.6843
TC-LSTM	0.7601	*	0.6562	*
ATAE-LSTM	0.7718	0.6702	0.6776	0.6393
IAN	0.7860	0.6631	0.7210	0.6592
MemNet	0.7816	0.6583	0.7033	0.6409
<b>Our</b>	<b>0.8113</b>	<b>0.7106</b>	<b>0.7357</b>	<b>0.6870</b>

### 3.3 Result Comparisons

We have obtained a comparison of the sentiment analysis accuracy data of our proposed model and other standard baseline models. The results are shown in Table 2, we can see that our model achieves the best results on both standard data sets. For example, when testing the sentiment analysis on the restaurant dataset in three-class, our model yielded an accuracy that was about 3% higher than the accuracy of other baseline models.

The TD-LSTM and TC-LSTM models combine the aspect vector at the input with the context vector, illustrating the importance of the target keyword in multiple aspects of sentiment analysis. The Bi-LSTM sentiment analysis model

reads the context information in both directions, so the effect is slightly better than the TD-LSTM and TC-LSTM models, indicating that the target keyword is still lacking. Bi-LSTM also significantly improves the effect of reading context information from both directions than reading context information in a single direction.

The IAN model differs from the above models in that context information and target information are input into LSTM neural network respectively, where context information and target information are exchanged interactively. Considering the importance of aspect words to context and the relevance of context information to aspect words, The IAN model achieves advanced results in aspect-level sentiment analysis results.

For the MemNet model, it has achieved very good results in accuracy and F-measure since it uses a multilayer attention structure.

For our proposed model, considering the importance of position information for aspect words, we indexed and coded aspect words to highlight the importance of aspect words and generated attention mechanism by using aspect words and context sentence vectors with position awareness, which provided attention weight for two kinds of input: position-aware context input and aspect words embedded in context input separately. Taking full account of position information and aspect word information, we have achieved the state-of-the-art results under a new interactive attention mechanism.

### 3.4 Model Analysis

In this subsection, we designed the following experiments to analyze the structure of the proposed model, which is divided into the following aspects:

**Model w/o PM:** To verify the influence of the position-aware mechanism on the model. In this part, we run our model without using the position-aware mechanism.

**Model w/o W + A:** The word + aspect input module is an effective complement to the entire model. In this part, we only use two embedding inputs including sentence input and aspect input.

From Table 3, we first compared the model without the position-aware mechanism and the model we proposed. The results show that the proposed model performance is significantly better than the model without position-aware mechanism since the position index can effectively emphasize the role of the aspect words in the sentence and share more weight to the context near the aspect words.

Then, we only use the sentence input with position-aware and aspect input in this experiment. The results demonstrate a small decline in each data set, which proved that the joint embedding of sentences and aspects at the input port could enhance the performance of the proposed model.

**Table 3.** Accuracy and F1-Measure for model analysis

Method	Restaurant		Laptop	
	Accuracy	F-Measure	Accuracy	F-Measure
Model w/o PM	0.7903	0.6984	0.7006	0.6698
Model w/o W + A	0.8055	0.7011	0.7241	0.6854
Our model	<b>0.8113</b>	<b>0.7106</b>	<b>0.7357</b>	<b>0.6870</b>

## 4 Conclusion

In this paper, a framework of interactive attention based on bidirectional LSTM network and position-aware mechanism for aspect-level sentiment classification has been proposed. The position index awareness mechanism can use the position information between the aspect words and other words to determine the terms related to the sentiment polarity. At the attention mechanism level, the main idea is to extract the features of the aspect words and sentences and learn it alternately to obtain a more reasonable distribution of attention weight. Besides, sentence vectors and aspect words vectors are embedded together to improve the final sentiment classification performance in the form of the content supplement. The experimental results show that our model achieves remarkable performance and are superior to all baseline models.

**Acknowledgements.** This work is partially supported by the National Natural Science Foundation of China (No. 61573095). This work is supported by the Natural Science Foundation of Shanghai under grant no. 20ZR1402800.

## References

- Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found. Trends Inf. Retrieval* **2**(1–2), 1–135 (2008)
- Liu, B.: Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* **5**(1), 1–167 (2012)
- Schouten, K., Frasincar, F.: Survey on aspect-level sentiment analysis. *IEEE Trans. Knowl. Data Eng.* **28**(3), 813–830 (2015)
- Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
- Tang, D., Qin, B., Feng, X., Liu, T.: Effective LSTMs for target-dependent sentiment classification. arXiv preprint [arXiv:1512.01100](https://arxiv.org/abs/1512.01100) (2015)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
- Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 606–615 (2016)
- Ma, D., Li, S., Zhang, X., Wang, H.: Interactive attention networks for aspect-level sentiment classification. arXiv preprint [arXiv:1709.00893](https://arxiv.org/abs/1709.00893) (2017)

9. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
10. Wagner, J., Arora, P., Cortes, S., Barman, U., Bogdanova, D., Foster, J., Tounsi, L.: DCU: aspect-based polarity classification for SemEval task 4 (2014)
11. Jia, Y.: Robust control with decoupling performance for steering and traction of 4WS vehicles under velocity-varying motion. *IEEE Trans. Control Syst. Technol.* **8**(3), 554–569 (2000)
12. Jia, Y.: Alternative proofs for improved LMI representations for the analysis and the design of continuous-time systems with polytopic type uncertainty: a predictive approach. *IEEE Trans. Autom. Control* **48**(8), 1413–1416 (2003)



# Neural Network-Based Exponential Stability of Affine Nonlinear Systems by Event-Triggered Approach

Fan Liu<sup>1</sup>, Yiming Gan<sup>1,2</sup>, Xueqing Yang<sup>3</sup>, and Wuneng Zhou<sup>1,4(✉)</sup>

<sup>1</sup> Donghua University, Shanghai 201620, China

[NielsbergLiu@foxmail.com](mailto:NielsbergLiu@foxmail.com)

<sup>2</sup> Guangdong Polytechnic College, Foshan 528041, China

[ygan@ultraliv.com](mailto:ygan@ultraliv.com)

<sup>3</sup> Educational Technology Center, Donghua University, Shanghai 200051, China  
[etdaqing@163.com](mailto:etdaqing@163.com)

<sup>4</sup> Engineering Research Center of Digitized Textile and Fashion Technology,  
Donghua University, Shanghai, China  
[zhouwuneng@163.com](mailto:zhouwuneng@163.com)

**Abstract.** In this paper, the event-triggered approach is considered as the implementation of exponential stability for a class of nonlinear systems with an asymptotic control input which is generated by an neural network-based feedback network. Under additional mild assumptions, we first rigorously conduct discussion on the transformation between affine nonlinear systems and linear systems via some existing results. Subsequently, we present a novel weight update law for the feedback neural network (NN) so as to achieve the process of finding the optimal weight matrix. Afterward, we derive an appropriate condition using the Lyapunov-Krasovskii method, that guarantees the practical convergence of the closed-loop system toward an equilibrium point (zero point) and is used to design relevant coefficients for the event-triggering scheme. Finally, a numerical example substantiates the achievement of exponential stability and the reduction of loss of the closed-loop system.

**Keywords:** Delayed nonlinear system · Exponential stability · Neural network · Event-triggered control

## 1 Introduction

Stability problems for linear systems have been widely investigated under various event-triggering mechanisms in the past several decades[1–4], in which the practicality of event-triggering mechanisms has been strongly improved. However, in this paper, we want to apply event-triggering scheme to some special nonlinear systems which may be transformed into corresponding linear systems if certain constraint conditions are satisfied. In particular, we focus on a class of nonlinear system called affine nonlinear system, which has been proved to

possess some special traits, for example, its transformation into linear system while the aforementioned constraint conditions are involved. Furthermore, we note that the existing event-triggered methods applied to affine nonlinear systems are rarely based on the practical time-varying delays, instead of simply assuming those systems with no delays or constant delays [5].

Similar with other works where event-triggering instants are transmitted when the error measurements of system states exceed the thresholds [6], in this note, error measurements are also used to synthesize the event-triggered control condition. In [6], a new event-triggering mechanism is presented to determine whether the sampled states will be sent out, which and whose modifications now are widely adopted to decide the triggering instants of closed-loop systems. Besides, comparing with traditional sample-based control methods, some remarkable properties of event-triggering schemes are illustrated in [7], for instance, the computation and usage of bandwidth can effectively reduce. Another advantage about event-triggered techniques is that they can guarantee a lower boundary of inter-event intervals, which in nature excludes the Zeno behavior as shown in [3]. As a consequence, such trait leads to not only the achievement of system stability, but also the control input of closed-loop systems only updates when event-triggered control condition is satisfied. In this paper, the nonlinear feedback network formed as NN-based dynamics and event-triggered controller are selected to execute the approximation of ideal input control. Such attempt to apply the event-triggered control techniques on nonlinear models can be seen in some creative works [5]. We find that stabilization of closed-loop systems with NN-based control input feedback networks is discussed in [8], where ones proved that NN-based approximation method is applicable.

The objective of the present paper is to realize the exponential stability of a kind of nonlinear uncertain systems using the event-triggered approach. And the feedback network of such kind of systems is approximated by an NN and exists time-varying communication delay. In addition, the control input of nonlinear system or the output of the feedback network is designed to update when the threshold of the even-triggering scheme is violated. Thus, the main contributions of this paper include: 1) nonlinear uncertain systems in affine form are investigated, whose feedback network is approximated by a designed NN and include time-varying delay of communication; 2) a novel update law of the weight matrix of event-based NN is presented to reduce workload of communication channel and ensure the system stability; and 3) a new stability criterion of the closed-loop nonlinear systems is introduced based on Lyapunov-Krasovskii method for delayed affine systems, which thereby leads to the criterion is rewritten in the form of linear matrix inequality (LMI).

## 2 Preliminaries and System Description

### 2.1 Relevant Lemmas and Definition

**Lemma 1 ([9]).** *Let  $R \in \mathbb{R}^{n \times n}$  be a symmetric positive definite matrix. If there exists  $X_i, Y_i, i = 1, 2$  with appropriate dimensions, such that inequality*

$\begin{bmatrix} R & 0 \\ 0 & R \end{bmatrix} - \alpha \begin{bmatrix} X_1 & Y_1 \\ Y_1^T & 0 \end{bmatrix} - (1 - \alpha) \begin{bmatrix} 0 & Y_2 \\ Y_2^T & X_2 \end{bmatrix} \succeq 0$  holds for all  $\alpha = 0, 1$ , then for all  $\alpha \in (0, 1)$ , there holds  $\begin{bmatrix} \frac{1}{\alpha}R & 0 \\ 0 & \frac{1}{1-\alpha}R \end{bmatrix} \succeq \begin{bmatrix} R & 0 \\ 0 & R \end{bmatrix} + (1 - \alpha) \begin{bmatrix} X_1 & Y_2 \\ Y_2^T & 0 \end{bmatrix} + \alpha \begin{bmatrix} 0 & Y_1 \\ Y_1^T & X_2 \end{bmatrix}$ .

**Lemma 2** ([10]). *For any prescribed symmetric positive definite matrix  $R \in \mathbb{R}^{n \times n}$  and a differentiable function  $x : [a, b] \rightarrow \mathbb{R}^n$ , the following inequalities hold*

$$\begin{cases} h \int_a^b \dot{x}^T(s) R \dot{x}(s) ds \geq \Phi_1^T R \Phi_1 + 3\Phi_2^T R \Phi_2 + 5\Phi_3^T R \Phi_3 \\ \int_a^b \int_{\rho}^b \dot{x}^T(s) R \dot{x}(s) ds d\rho \geq 2\Phi_4^T R \Phi_4 + 4\Phi_5^T R \Phi_5 + 6\Phi_6^T R \Phi_6 \end{cases}$$

where  $h = b - a$ ,  $\Phi_1 = x(b) - x(a)$ ,  $\Phi_2 = x(b) + x(a) - \frac{2}{h} \int_a^b x(s) ds$ ,  $\Phi_3 = x(b) - x(a) + \frac{6}{h} \int_a^b x(s) ds - \frac{12}{h^2} \int_a^b \int_{\rho}^b x(s) ds d\rho$ ,  $\Phi_4 = x(b) - \frac{1}{h} \int_a^b x(s) ds$ ,  $\Phi_5 = x(b) + \frac{2}{h} \int_a^b x(s) ds - \frac{6}{h^2} \int_a^b \int_{\rho}^b x(s) ds d\rho$ , and  $\Phi_6 = x(b) - \frac{3}{h} \int_a^b x(s) ds + \frac{24}{h^2} \int_a^b \int_{\rho}^b x(s) ds d\rho - \frac{60}{h^3} \int_a^b \int_r^b \int_{\rho}^b x(s) ds d\rho dr$ .

**Definition 1** ([13]). *Suppose there exists positive scalars  $\ell_1, \ell_2, \gamma$  and a functional  $V(t, z(t), \dot{z}(t))$ , which is continuous from the right for  $z(t)$ , absolutely continuous for  $t \neq s_k$ , such that  $\lim_{t \rightarrow s_k^-} V(t) \geq V(s_k^-)$  and  $\ell_1 |z(0)|^2 \leq$*

*$V(t, z(t), \dot{z}(t)) \leq \ell_2 \left[ \max_{t \in [-a, 0]} |z(t)| + \left( \int_{-a}^0 |\dot{z}(\theta)|^2 d\theta \right)^{\frac{1}{2}} \right]^2$  with  $a \geq \tau_M$ , and almost for all  $t$ , there holds  $\dot{V}(t, z(t), \dot{z}(t)) + 2\gamma V(t, z(t), \dot{z}(t)) \leq 0$ , then  $z(t)$  is exponentially stable with the decay rate  $\gamma$ .*

## 2.2 Problem Description

We choose a class of nonlinear systems in affine form:

$$\dot{x} = f(x) + \mathcal{G}(x)u, \quad x(0) = x_0, \quad t \geq 0 \quad (1)$$

where  $x = [x_1, x_2, \dots, x_n] \in \mathfrak{D} \subseteq \mathbb{R}^n$  and  $u = [u_1, u_2, \dots, u_m] \in \mathbb{R}^m$  are system state vectors and control input, respectively;  $f : \mathfrak{D} \rightarrow \mathbb{R}^n$  and  $\mathcal{G} : \mathfrak{D} \rightarrow \mathbb{R}^{n \times m}$  are internal nonlinear dynamics and control coefficient matrix, respectively. We assume that system (1) is feedback linearizable [11].

**Assumption 1.** System (1) is linearizable, i.e., the state of system (1) can be transformed into  $z \in \mathbb{R}^n$  and

$$\dot{z} = Az + B\beta(x)[u - \alpha(x)]. \quad (2)$$

Dynamical system (2) clearly possesses a linear form such that it is easy to deal with if  $A$  and  $B$  are specified. Note that not all the nonlinear systems in the form of (1) are linearizable, but fortunately, a large number of affine nonlinear systems satisfy Assumption 1 in practical systems, such as pendulum system, damping system, and many others.

Now our mission is to design a feedback control input  $u(x)$  for the transformation achievement from (1) to (2). It cannot obtain the control input directly

from system (1), since  $f(x)$  and  $\mathcal{G}(x)$  are unknown. But any measurable function can be approximated to any desired degree of accuracy using multi-layer NNs with at least one hidden layer and sufficiently many hidden units [12]. So in this paper, we consider only one hidden layer NNs with  $l$  hidden units, a constant input weight matrix  $V \in \mathbb{R}^{n \times l}$  and a variable output weight matrix  $W \in \mathbb{R}^{l \times m}$  to approximate  $u(x)$ . Thus,  $u(x)$  is represented as

$$u(x) = W^T \phi(V^T x) \quad (3)$$

where  $\phi(\cdot) = [\phi_1(\cdot), \phi_2(\cdot), \dots, \phi_l(\cdot)] : \mathbb{R}^l \rightarrow \mathbb{R}^l$  is the neuron activation function. Besides, in order to obtain effective approximation property of NNs, following assumption is needed for  $\phi(\cdot)$ . The reader can find more details in [12].

**Assumption 2.**  $\phi_i(\cdot), i = 1, 2, \dots, l$ , is a non-decreasing function, mapping  $\mathbb{R}$  to  $[0, 1]$ , and satisfies  $\lim_{x \rightarrow +\infty} \phi_i(x) = 1$  and  $\lim_{x \rightarrow -\infty} \phi_i(x) = 0$ .

Noticing  $\phi(\cdot)$  is bounded and nonnegative under Assumption 2, thus for any constant  $W$ , it leads to a bounded control input  $u(x)$ . Taking into account that the bandwidth of feedback channels is limited, we use following event-triggering scheme to broadcast system's state at every time instant  $s_k$ ,  $k \in \mathbb{N}$  [6].

$$s_{k+1} = \min\{t > s_k | (z(t) - z(s_k))^T \Omega (z(t) - z(s_k)) > \sigma z^T(s_k) \Omega z(s_k)\} \quad (4)$$

Define time-varying variable  $\tau(t)$  as the so-called delay at time instant  $t$ , so that  $\tau(t) = t - s_k$ ,  $t \in [s_k, s_{k+1})$ . Moreover, taking the practice into consideration, we assume  $\tau(t)$  satisfies  $0 \leq \tau_m \leq \tau(t) \leq \tau_M$ . Hence, Eq. (4) can be further rewritten as following one.

$$\begin{aligned} s_{k+1} = \min\{t > s_k | & (z(t) - z(t - \tau(t)))^T \Omega (z(t) - z(t - \tau(t))) \\ & > \sigma z^T(t - \tau(t)) \Omega z(t - \tau(t))\} \end{aligned} \quad (5)$$

Then we derive, in light of the discussion above, that the closed-loop system can be described as follows.

$$\begin{cases} \dot{z} = Az + B\beta(x)[u(x(t - \tau(t))) - \alpha(x)] \\ u(x(t - \tau(t))) = W^T \phi(V^T x(t - \tau(t))) \end{cases} \quad t \geq 0 \quad (6)$$

### 3 Main Results

#### 3.1 Weight Update Law

Let  $lm = l \times m$ . We choose following weight update law to perform the update of output weight matrix  $W$  in (6) at the event-triggering instants.

$$\text{vec}W(k+1) = G\text{vec}W(k) + H(e_\phi(k) \otimes I_m)D, \quad k = 0, 1, 2, \dots, \quad (7)$$

where  $\text{vec}W(k)$  is the vectorization of weight matrix  $W$  at event-triggering instant  $s_k$ ;  $G \in \mathbb{R}^{lm \times lm}$  is a nonsingular matrix. Furthermore, the spectral

radius of  $G$  is less than 1.  $H \in \mathbb{R}^{lm \times lm}$  denotes the gain matrix to be designed,  $e_\phi(k) = \phi(V^T x(s_{k+1})) - \phi(V^T x(s_k))$ , and  $D \in \mathbb{R}^m$  is a constant vector to match the dimension.

It is clear that (7) is asymptotically stable since the eigenvalues of  $G$  are located inside the unit circle and  $(e_\phi(k) \otimes I_m) D$  are bounded due to boundedness of  $e_\phi(k)$ , which is guaranteed by Assumption 2. By noticing this we know that  $\text{vec}W(k)$  will converge to a constant vector which we denote by  $\text{vec}W_f$ . We now define the weight error dynamics as  $e_\phi(k+1) = \text{vec}W_f - \text{vec}W(k)$  and in view of (7), it is given by

$$e_w(k+1) = Ge_w(k) - H(e_\phi(k) \otimes I_b)D + (I_{lb} - G)\text{vec}W_f, \quad k = 0, 1, 2, \dots \quad (8)$$

### 3.2 Stability Analysis Under Event-Triggering Scheme

**Theorem 1.** *For prescribed scalars  $\gamma \geq 0$ ,  $\delta > 0$ ,  $\tau_M \geq \tau_m \geq 0$ , the closed-loop system (6) under event-triggering scheme (5) is exponentially stable with the decay rate  $\gamma$  if there exists matrices  $P \succeq 0$ ,  $R_i \succeq 0$ ,  $Q_i \succeq 0$ ,  $S \succeq 0$ ,  $\Omega \succeq 0$ , and  $X_i, Y_i$ ,  $i = 1, 2$  with appropriate dimensions such that, for  $\tau = \tau_m$  and  $\tau = \tau_M$ , following inequalities always hold.*

$$\begin{aligned} & \begin{bmatrix} \Xi_i(\tau) & \Xi_2 e_1^T A^T M \\ \Xi_3 & 0 \\ * & -M \end{bmatrix} \prec 0, \\ & \begin{bmatrix} \tilde{R}_2 & 0 \\ X_1 & Y_1 \\ 0 & \tilde{R}_1 \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \tilde{R}_2 & 0 \\ 0 & \tilde{R}_2 \end{bmatrix} - \begin{bmatrix} 0 & Y_2 \\ Y_1^T & X_2 \end{bmatrix} \succeq 0 \end{aligned} \quad (9)$$

where

$$\begin{aligned} \Xi_1(\tau) &= e_1^T \left( PA + A^T P + \gamma P + \gamma P^T + Q_1 - \Omega \right) e_1 + e_1^T \Omega e_3 + e_3^T \Omega e_1 \\ &\quad + e^{-2\gamma\tau_m} e_2^T (Q_2 - Q_1) e_2 + (\sigma - 1) e_3^T \Omega e_3 - e^{-2\gamma\tau_M} e_4^T Q_2 e_4 \\ &\quad - e^{-2\gamma\tau_m} \Pi_1^T \tilde{R}_1 \Pi_1 - e^{-2\gamma\tau_M} \begin{bmatrix} \Psi_1 \\ \Psi_2 \end{bmatrix}^T \begin{bmatrix} \tilde{S} & 0 \\ 0 & \tilde{S} \end{bmatrix} \begin{bmatrix} \Psi_1 \\ \Psi_2 \end{bmatrix} \\ &\quad - e^{-2\gamma\tau_M} \begin{bmatrix} \Pi_2 \\ \Pi_3 \end{bmatrix}^T \begin{bmatrix} \tilde{R}_2 & 0 \\ 0 & \tilde{R}_2 \end{bmatrix} + \frac{\tau_M - \tau}{\tau_h} \begin{bmatrix} X_1 & Y_2 \\ Y_2^T & 0 \end{bmatrix} + \frac{\tau - \tau_m}{\tau_h} \begin{bmatrix} 0 & Y_1 \\ Y_1^T & X_2 \end{bmatrix} \begin{bmatrix} \Pi_2 \\ \Pi_3 \end{bmatrix} \\ \Xi_2 &= e_1^T \left[ PB + A^T (\tau_m^2 R_1 + \tau_h^2 R_2 + \frac{\tau_h^2}{2} S) B \right] \\ \Xi_3 &= B^T \left[ \tau_m^2 R_1 + \tau_h^2 (R_2 + \frac{1}{2} S) \right] B \end{aligned}$$

$$\Pi_i = \text{col}\{e_i - e_{i+1}, e_i + e_{i+1} - 2e_{i+4}, e_i - e_{i+1} + 6e_{i+4} - 12e_{i+7}\}$$

$$\Psi_i = \text{col}\{e_{i+1} - e_{i+5}, e_{i+1} + 2e_{i+5} - 6e_{i+8}, e_{i+1} - 3e_{i+5} + 24e_{i+8} - 60e_{i+10}\}$$

$$\tilde{R}_1 = \text{diag}\{R_1, 3R_1, 5R_1\}, \tilde{R}_2 = \text{diag}\{R_2, 3R_2, 5R_2\}, \tilde{S} = \text{diag}\{2S, 4S, 6S\}$$

$$M = \tau_m^2 R_1 + \tau_h^2 R_2 + \frac{\tau_h^2}{2} S, e_i = [0_{n \times (i-1)n}, I_n, 0_{n \times (12-i)n}] .$$

*Proof.* Let  $v(t) = \beta(x)[u(x(t - \tau(t))) - \alpha(x)]$ ,  $\tau_h = \tau_M - \tau_m$ . We choose the Lyapunov-Krasovskii functional as follows:

$$V(t) = \sum_7^{i=4} V_i(t) \quad (10)$$

where

$$\begin{aligned} V_4(t) &= z^T(t) P z(t) \\ V_5(t) &= \int_{t-\tau_m}^t e^{2\gamma(s-t)} z^T(s) Q_1 z(s) ds + \int_{t-\tau_M}^{t-\tau_m} e^{2\gamma(s-t)} z^T(s) Q_2 z(s) ds \\ V_6(t) &= \tau_m \int_{-\tau_m}^0 \int_{t+\rho}^t e^{2\gamma(s-t)} \dot{z}^T(s) R_1 \dot{z}(s) ds d\rho \\ &\quad + \tau_h \int_{-\tau_M}^{-\tau_m} \int_{t+\rho}^t e^{2\gamma(s-t)} \dot{z}^T(s) R_2 \dot{z}(s) ds d\rho \\ V_7(t) &= \int_{-\tau_M}^{-\tau_m} \int_r^{-\tau_m} \int_{t+\rho}^t e^{2\gamma(s-t)} \dot{z}^T(s) S \dot{z}(s) ds d\rho dr. \end{aligned}$$

By differentiating these functionals above, we derive

$$\begin{aligned} \dot{V}_4(t) + 2\gamma V_4(t) &= 2\dot{z}^T(t) P z(t) + 2\gamma z^T(t) P z(t) \\ \dot{V}_5(t) + 2\gamma V_5(t) &= z^T(t) Q_1 z(t) + e^{-2\gamma\tau_m} z^T(t - \tau_m) (Q_2 - Q_1) z(t - \tau_m) \\ &\quad - e^{-2\gamma\tau_M} z^T(t - \tau_M) Q_2 z(t - \tau_M) \\ \dot{V}_6(t) + 2\gamma V_6(t) &\leq \tau_m^2 \dot{z}^T(t) R_1 \dot{z}(t) + \tau_h^2 \dot{z}^T(t) R_2 \dot{z}(t) - \tau_m e^{-2\gamma\tau_m} \\ &\quad \times \int_{t-\tau_m}^t \dot{z}^T(\rho) R_1 \dot{z}(\rho) d\rho - \tau_h e^{-2\gamma\tau_M} \int_{t-\tau_M}^{t-\tau_m} \dot{z}^T(\rho) R_2 \dot{z}(\rho) d\rho \\ \dot{V}_7(t) + 2\gamma V_7(t) &\leq \frac{1}{2} \tau_h^2 \dot{z}^T(t) S \dot{z}(t) - e^{-2\gamma\tau_M} \int_{t-\tau_M}^{t-\tau_m} \int_r^{t-\tau_m} \dot{z}^T(\rho) S \dot{z}(\rho) d\rho dr. \end{aligned}$$

Let  $\tau_1 = \tau(t) - \tau_m$ ,  $\tau_2 = \tau_M - \tau(t)$  and choose  $\chi(t) =$

$$\text{col} \left\{ \begin{bmatrix} z(t) \\ z(t-\tau_m) \\ z(t-\tau(t)) \\ z(t-\tau_M) \end{bmatrix}, \begin{bmatrix} \frac{1}{\tau_m} \int_{-\tau_m}^t z(s) ds \\ \frac{1}{\tau_1} \int_{-\tau_m}^{t-\tau_m} z(s) ds \\ \frac{1}{\tau_2} \int_{-\tau_M}^{t-\tau(t)} z(s) ds \\ \frac{1}{\tau_m^2} \int_{-\tau_m}^t \int_\rho^t z(s) ds d\rho \end{bmatrix}, \begin{bmatrix} \frac{1}{\tau_1^2} \int_{-\tau(t)}^{t-\tau_m} \int_\rho^{t-\tau_m} z(s) ds d\rho \\ \frac{1}{\tau_2^2} \int_{-\tau_M}^{t-\tau(t)} \int_\rho^{t-\tau(t)} z(s) ds d\rho \\ \frac{1}{\tau_1^3} \int_{-\tau_m}^{t-\tau_m} \int_t^{t-\tau_m} \int_\rho^{t-\tau_m} z(s) ds d\rho dr \\ \frac{1}{\tau_2^3} \int_{-\tau_M}^{t-\tau(t)} \int_t^{t-\tau(t)} \int_\rho^{t-\tau(t)} z(s) ds d\rho dr \end{bmatrix} \right\}.$$

Then, we obtain further, based on Lemma 2, that  $-\tau_m \int_{t-\tau_m}^t \dot{z}^T(\rho) R_1 \dot{z}(\rho) d\rho \leq -\chi^T(t) \Pi_1^T \tilde{R}_1 \Pi_1 \chi(t)$ , and  $-\int_{t-\tau_M}^{t-\tau_m} \dot{z}^T(\rho) R_2 \dot{z}(\rho) d\rho = -\int_{t-\tau(t)}^{t-\tau_m} \dot{z}^T(\rho) R_2 \dot{z}(\rho) d\rho - \int_{t-\tau_M}^{t-\tau(t)} \dot{z}^T(\rho) R_2 \dot{z}(\rho) d\rho \leq -\chi^T(t) \begin{bmatrix} \Pi_2 \\ \Pi_3 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\tau_1} \tilde{R}_2 & 0 \\ 0 & \frac{1}{\tau_2} \tilde{R}_2 \end{bmatrix} \begin{bmatrix} \Pi_2 \\ \Pi_3 \end{bmatrix} \chi(t)$ . Moreover, in light of Lemma 1, if the inequalities  $\begin{bmatrix} \tilde{R}_2 & 0 \\ 0 & \tilde{R}_1 \end{bmatrix} - \begin{bmatrix} X_1 & Y_1 \\ Y_1^T & 0 \end{bmatrix} \succeq 0$ , and  $\begin{bmatrix} \tilde{R}_2 & 0 \\ 0 & \tilde{R}_2 \end{bmatrix} -$

$\begin{bmatrix} 0 & Y_2 \\ Y_2^T & X_2 \end{bmatrix} \succeq 0$  hold, we arrive at that  $-\tau_h \int_{t-\tau_M}^{t-\tau_m} \dot{z}^T(\rho) R_2 \dot{z}(\rho) d\rho \leq$

$$-\chi^T(t) \begin{bmatrix} \Pi_2 \\ \Pi_3 \end{bmatrix}^T \left\{ \begin{bmatrix} \tilde{R}_2 & 0 \\ 0 & \tilde{R}_2 \end{bmatrix} + \frac{\tau_2}{\tau_h} \begin{bmatrix} X_1 & Y_2 \\ Y_2^T & 0 \end{bmatrix} + \frac{\tau_1}{\tau_h} \begin{bmatrix} 0 & Y_1 \\ Y_1^T & X_2 \end{bmatrix} \right\} \begin{bmatrix} \Pi_2 \\ \Pi_3 \end{bmatrix} \chi(t). \quad (11)$$

It is well known that the second and third terms of the curly brackets in (11) are convex in  $\tau(t)$ . With this end in view, if (11) is feasible for  $\tau(t) = \tau_m$  and  $\tau(t) = \tau_M$ , it leads to (11) is feasible for all  $\tau(t) \in [\tau_m, \tau_M]$ .

Similarly, by Lemma 2, we have

$$\begin{aligned} & - \int_{t-\tau_M}^{t-\tau_m} \int_r^{t-\tau_m} \dot{z}^T(\rho) S \dot{z}(\rho) d\rho dr \\ & \leq - \int_{t-\tau(t)}^{t-\tau_m} \int_r^{t-\tau_m} \dot{z}^T(\rho) S \dot{z}(\rho) d\rho dr - \int_{t-\tau_M}^{t-\tau(t)} \int_r^{t-\tau(t)} \dot{z}^T(\rho) S \dot{z}(\rho) d\rho dr \\ & \leq -\chi^T(t) \Psi_1^T \tilde{S} \Psi_1 \chi(t) - \chi^T(t) \Psi_2^T \tilde{S} \Psi_2 \chi(t). \end{aligned}$$

By introducing the additional term  $\sigma z^T(t - \tau(t)) \Omega z(t - \tau(t)) - (z(t) - z(t - \tau(t)))^T \Omega (z(t) - z(t - \tau(t))) \geq 0$ , we yield, in view of (6), that

$$\begin{aligned} & \dot{V}(t) + 2\gamma V(t) \leq \dot{V}(t) + 2\gamma V(t) \\ & + \sigma z^T(t - \tau(t)) \Omega z(t - \tau(t)) - (z(t) - z(t - \tau(t)))^T \Omega (z(t) - z(t - \tau(t))) \\ & \leq \chi^T(t) \left( e_1^T (2PA + 2\gamma P + Q_1 - \Omega) e_1 + 2e_1^T \Omega e_3 + e^{-2\gamma\tau_m} e_2^T (Q_2 - Q_1) e_2 \right. \\ & + (\sigma - 1)e_3^T \Omega e_3 - e^{-2\gamma\tau_M} e_4^T Q_2 e_4 - e^{-2\gamma\tau_m} \Pi_1^T \tilde{R}_1 \Pi_1 - e^{-2\gamma\tau_M} \\ & \times \begin{bmatrix} \Pi_2 \\ \Pi_3 \end{bmatrix}^T \left\{ \begin{bmatrix} \tilde{R}_2 & 0 \\ 0 & \tilde{R}_2 \end{bmatrix} + \frac{\tau_M - \tau(t)}{\tau_h} \begin{bmatrix} X_1 & Y_2 \\ Y_2^T & 0 \end{bmatrix} + \frac{\tau(t) - \tau_m}{\tau_h} \begin{bmatrix} 0 & Y_1 \\ Y_1^T & X_2 \end{bmatrix} \right\} \begin{bmatrix} \Pi_2 \\ \Pi_3 \end{bmatrix} \\ & \left. - e^{-2\gamma\tau_M} \begin{bmatrix} \Psi_1 \\ \Psi_2 \end{bmatrix}^T \begin{bmatrix} S & 0 \\ 0 & \tilde{S} \end{bmatrix} \begin{bmatrix} \Psi_1 \\ \Psi_2 \end{bmatrix} \right) \chi(t) + \chi^T(t) \{ e_1^T [2PB + A^T(2\tau_m^2 R_1 \right. \\ & \left. + 2\tau_h^2 R_2 + \tau_h^2 S)B] \} v(t) + v^T(t) \{ B^T[\tau_m^2 R_1 + \tau_h^2(R_2 + \frac{1}{2}S)]B \} v(t) \\ & + \chi^T(t) \left[ e_1^T A^T(\tau_m^2 R_1 + \tau_h^2 R_2 + \frac{\tau_h^2}{2}S)Ae_1 \right] \chi(t). \end{aligned}$$

We thus have  $\dot{V}(t) + 2\gamma V(t) \leq \eta^T(t) \left\{ \begin{bmatrix} \Xi_1(\tau) & \Xi_2 \\ \Xi_2^T & \Xi_3 \end{bmatrix} + [Ae_1 \ 0]^T M [Ae_1 \ 0] \right\} \eta(t)$  where the augmented vector  $\eta(t) = [\Xi^T(t), v^T(t)]^T$ . Then, by applying Schur complement, the result can be derived directly, which completes the proof.

Based on Definition 1, Theorem 1 implies that the closed-loop system (6) can achieve exponential stability under the event-triggering scheme (5).

## 4 Numerical Validation

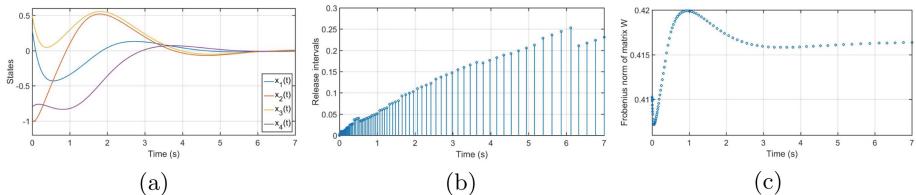
We implement the validation using a dynamic model called fault-considered rocket fairing structural acoustic model, where the inner structural parameters are given as

$$f(x) = A_0 x + \begin{bmatrix} -0.3 \sin(2x_2) + x_4 \\ -0.5x_3 \cos^3(1.5x_1) - x_2^3 + 0.2x_1^3 \\ -0.4 \sin(2x_1) \\ -0.6 \sin(x_1) \cos(3x_4) \end{bmatrix}, \quad \mathcal{G}(x) = \begin{bmatrix} 0 & 0.775 & 0.525 \\ 0 & 0.4 & 0.595 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

The matrix  $A_0$  is given by

$$A_0 = \begin{bmatrix} 0 & 1 & 0.0802 & 1.0415 \\ -0.198 & -0.115 & -0.0318 & 0.3 \\ -3.05 & 1.188 & -0.465 & 0.9 \\ 0 & 0.0805 & 1 & 0 \end{bmatrix}.$$

It is simple to find that  $n = 4$ ,  $m = 3$  and the number of neurons of NN's hidden layer  $l$  is chosen as 16 for better asymptotic property according to the study of Hornik et al. [12]. According to Assumption 2, choose a common squashing function as the activation function which we denote by  $\phi(\cdot) = e^{(\cdot)}/(e^{(\cdot)} + 1)$ . The input weight matrix  $V$  belonging to  $\mathbb{R}^{4 \times 16}$  are randomly initialized from the uniform distribution in the interval  $[0, 1]$ , and the corresponding initial output matrix  $W \in \mathbb{R}^{16 \times 3}$  is populated with random variables from the uniform distribution in the interval  $[0, 0.1]$ . As to the initial state vector, it is chosen as  $x_0 = [0.3, -1, 0.5, -0.8]^T$ , and following the linearization programming in [11], we can obtain  $A$  that is given in Table 1 while the constant matrix  $B$  are designed to be an identity matrix with appropriate dimensions. Moreover, suppose that  $\tau_m$  and  $\tau_M$  are 0 and 0.1, respectively, we can obtain the value of  $\Omega$  that is given in Table 1 by applying Theorem 1 with  $\gamma = 2.5$  and  $\delta = 0.5$ .



**Fig. 1.** (a) Trajectory of system state  $x = [x_1, x_2, x_3, x_4]^T$ . (b) Release instants and release intervals. (c) Frobenius norm of matrix  $W$  at event-triggering instants.

Simulation result implies that the states exponentially converge to zero with the usage of event-triggering scheme and asymptotic property of NN, which is shown in Fig. 1(a). Furthermore, event-triggering release instants and release

**Table 1.** Matrices  $A$  and  $\Omega$ 

$A$	$\Omega$
-5.796 1.4185 1.9622 1.8029	6.8598 -1.5521 -2.7852 -2.2355
-3.9566 0.0356 1.1246 0.6606	-1.5521 0.9949 0.6950 0.5485
-3.05 1.188 -0.465 0.9	-2.7852 0.6950 1.7632 0.9883
0 0.0805 1 0	-2.2355 0.5485 0.9883 1.4167

**Table 2.** The upper bounds of  $\tau_M$  with respect to  $\sigma$  when  $\gamma = 2.5$  and  $\tau_m = 0$ .

$\sigma$	0.05	0.1	0.2	0.3	0.4	0.6
$\bar{\tau}_M$	0.3017	0.2611	0.2214	0.1724	0.1368	0.0503

intervals are displayed in Fig. 1(b), which suggests that the event-triggering controller can guarantee the significant reduction of transmission rate and computational complexity. Subsequently, for the clarity of the dynamic of output matrix  $W$ , Fig. 1(c) shows that the Frobenius norm of  $W$ , i.e.  $\|W\|$ . Finally, the upper bound  $\bar{\tau}_M$  of  $\tau_M$  with fixed  $\gamma$  and  $\tau_m$  and different  $\sigma$  is investigated in this work, we report the result in Table 2 where the parameters are carefully hand tuned. To check by eye, this plot indicates the existence of an ideal weight matrix. Similarly, the relationship of  $\tau_M$  and  $\gamma$  or  $\sigma$  and  $\gamma$  can be investigated based on Theorem 1 using the same method, but we omit that here.

## 5 Conclusion

The exponential stability of nonlinear systems in affine form has been realized under event-triggered control in this paper using the approximate property of the NN, which is utilized as the implementation of the feedback network where exists time-varying delay. Besides, a more effective and feasible weight update law of the NN has been developed to execute the update of the NN's weight when the event-triggering condition are satisfied, as well as a significant criterion for the closed loop stability of nonlinear system has been obtained to maintain the stability. Finally, the simulation results suggests that the event-triggering scheme and the NN's approximate property can be employed to other time-varying systems.

**Acknowledgement.** This work is supported by the Natural Science Foundation of Shanghai under Grant No. 20ZR1402800.

## References

1. Jia, Y.: Robust control with decoupling performance for steering and traction of 4WS vehicles under velocity-varying motion. *IEEE Trans. Control Syst. Technol.* **3**, 554–569 (2000)

2. Jia, Y.: Alternative proofs for improved LMI representations for the analysis and the design of continuous-time systems with polytopic type uncertainty: a predictive approach. *IEEE Trans. Autom. Control* **8**, 1413–1416 (2003)
3. Hu, S., Yue, D.: Event-triggered control design of linear networked systems with quantizations. *ISA Trans.* **1**, 153–162 (2012)
4. Shi, D., et al.: Event-Based State Estimation: A Stochastic Perspective. Springer Press, Cham (2016)
5. Liu, D., Yang, G.: Event-based model-free adaptive control for discrete-time nonlinear processes. *IET Control Theory Appl.* **11**, 2531–2538 (2017)
6. Yue, D., et al.: A delay system method for designing event-triggered controllers of networked control systems. *IEEE Trans. Autom. Control* **58**, 475–481 (2013)
7. Mazo, M., Tabuada, P.: Decentralized event-triggered control over wireless sensor/actuator networks. *IEEE Trans. Autom. Control* **56**, 2456–2461 (2011)
8. Hu, S., et al.: Stabilization of neural-network-based control systems via event-triggered control with nonperiodic sampled data. *IEEE Trans. Neural Netw. Learn. Syst.* **29**, 573–585 (2018)
9. Park, P., et al.: Reciprocally convex approach to stability of systems with time-varying delays. *Automatica* **47**, 235–238 (2011)
10. Zheng, W., et al.: New stability criteria for asymptotic stability of time-delay systems via integral inequalities and Jensen inequalities. *J. Inequal. Appl.* **2019**, 30 (2019)
11. Khalil, H.K., Grizzle, J.W.: Nonlinear systems. Prentice Hall, Upper Saddle River (2002)
12. Hornik, K., et al.: Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**, 359–366 (1989)
13. Fridman, E.: A refined input delay approach to sampled-data control. *Automatica* **46**, 421–427 (2010)



# Adaptive Finite-Time Leader-Following Consensus of Multi-agent Systems Against Time-Varying Actuator Faults

Yanhui Yin<sup>1,2</sup>, Fuyong Wang<sup>1,2</sup>, Zhongxin Liu<sup>1,2(✉)</sup>, and Zengqiang Chen<sup>1,2</sup>

<sup>1</sup> College of Artificial Intelligence, Nankai University,  
Tianjin 300350, People's Republic of China  
[1zhx@nankai.edu.cn](mailto:1zhx@nankai.edu.cn)

<sup>2</sup> Tianjin Key Laboratory of Intelligent Robotics, Nankai University,  
Tianjin 300350, People's Republic of China

**Abstract.** This paper investigates the fault-tolerant consensus problem for leader-following multi-agent systems. Time-varying actuator faults including loss of effectiveness, stuck, and outage are considered. By combining adaptive technique with linear matrix inequality approach, a distributed control protocol is put forward, under which the consensus error will be restricted to a small neighborhood around zero within finite time. The relationship between the steady-state consensus error and the control parameters is established by the practical finite-time stability theory. The proposed protocol is totally distributed. All the signals in the proposed scheme are bounded. A numerical example is presented to show the effectiveness of the theoretical results.

**Keywords:** Fault-tolerant control · Multi-agent systems · Finite-time control

## 1 Introduction

In the past decade, there has been an increasing interest in the cooperative control of multi-agent systems (MASs) because of its wide application in various fields, for instance, formation control [1], containment control [2], flocking [3]. The basic problem of distributed cooperative control is consensus, whose research content involves designing distributed control strategy by local information so as to the states of the agents achieve a common value. In recent years many results have been achieved on consensus [4–6].

Convergence time is an crucial indicator for evaluating the performance of MASs. Finite-time convergence theory has got ever-increasing attention. In [7], the leaderless consensus problem is studied with single-integrator dynamics. The finite-time consensus can be reached if the communication topology is undirected or directed with a detailed balance condition. In [8], the event-triggered method is utilized to design a distributed finite-time protocol. Two sufficient conditions

are put forward to guarantee the consensus result for MASs. In [9], the consensus of second-order switched MASs is considered. The radial basis function network is applied to help to construct the nonlinear protocol, under which the practical finite-time consensus can be ensured. In [10], the unknown dynamics and input saturation are both considered under directed topology. The proposed protocol is developed by a sliding-mode observer, and the finite-time saturated consensus can be reached.

Many modern systems have put forward higher requirements on the reliable operation of MASs. Fault-tolerant control (FTC) is becoming one of research hot spots in MASs. In [11], the authors investigate the fault-tolerant consensus problem with loss of effectiveness (LOE) faults. An active FTC protocol is constructed based on the estimation of fault severity. In [12], the information of the faults is estimated by some adaptive parameters, based on which, the distributed FTC schemes are designed for both leaderless MASs and leader-following MASs. Up until now, there are few studies on finite-time FTC for MASs. A finite settling time is not easy to be guaranteed, especially when the bounds of the faults are unknown.

The contributions are summarised as follows:

- A finite-time FTC scheme for general MASs is put forward under multiple faults and disturbances. The practical finite-time consensus can be ensured even if more than one agent gets faults.
- The proposed protocol is totally distributed and only the information from neighbours is needed.
- The relationship between the steady-state tracking error and the control parameters is established. The steady error and settling time can be controlled by designers.

The rest of the paper is organized as follows. In Sect. 2, preliminaries on graph theory are briefly presented and the FTC problem for MASs is formulated. The finite-time FTC protocol is presented and consensus results are analysed in Sect. 3. Section 4 gives a simulation example to verify the theoretical results. Some conclusions are provided in Sect. 5.

*Notions:* Let  $A^T$  denote the transpose of the matrix  $A$ .  $\text{col}\{a_1, a_2, \dots, a_n\}$  denotes  $[a_1^T, a_2^T, \dots, a_n^T]^T$ .  $\text{diag}\{b_1, b_2, \dots, b_n\}$  represents a diagonal matrix with diagonal elements  $b_1, b_2, \dots, b_n$ .  $I$  represents the identity matrix with appropriate dimensions.  $r(A)$  denotes the rank of  $A$ .

## 2 Preliminaries and Problem Statement

### 2.1 Preliminaries

In a leader-following system, the communication of the followers can be denoted by  $\mathcal{G} = (\mathcal{V}, \mathcal{A}, \mathcal{E})$ .  $\mathcal{V} = \{1, 2, \dots, N\}$  denotes the set of agents.  $(i, j) \in \mathcal{E}$  means node  $i$  can receive direct information from node  $j$ .  $\mathcal{A} = [a_{ij}]_{N \times N}$  denotes the weights of the exchanged information. That is,  $a_{ij} > 0$  when  $(i, j) \in \mathcal{E}$ ,  $a_{ij} = 0$

otherwise. Specially  $a_{ii} = 0$  for  $i \in \mathcal{V}$ . Let  $\mathcal{L} = \mathcal{D} - \mathcal{A}$ , where  $\mathcal{D}$  is a diagonal matrix with  $\sum_{j=1}^N a_{ij}$  on its main diagonal. A path between agent  $i_l$  and agent  $i_m (m > l)$  is a sequence of edges  $(i_{l+p}, i_{l+p+1})$ , where  $p = 0, 1, \dots, m-l$ . The graph is called undirected if  $a_{ij} = a_{ji}, i, j \in \mathcal{V}$ . If at least one path exists between any different agents, then  $\mathcal{G}$  is called connected. In addition, let  $\mathcal{K} = \mathcal{L} + \mathcal{H}$ , where  $\mathcal{H} = \text{diag}\{h_1, h_2, \dots, h_N\}$ .  $h_i \geq 0$  denotes the weight of the information from the leader to agent  $i$ .

## 2.2 Problem Statement

The dynamics of the MAS systems are

$$\dot{x}_0 = Ax_0 + Bu_0, \quad (1)$$

$$\dot{x}_i = Ax_i + Bu_i^F + Dd_i, i \in \mathcal{V}, \quad (2)$$

where  $x_i$  denotes the  $n$ -dimensional state,  $u_i$  denotes the  $m$ -dimensional control input,  $u_0$  is unknown but bounded.  $d_i \in \mathbb{R}^p$  is the time-varying disturbance satisfying  $\|d_i\| \leq \bar{d}_i$ .  $\bar{d}_i > 0$  is the unknown bound of the disturbance.  $A, B, D$  are all known matrices.  $D$  satisfied the matched condition  $r(B) = r(B, D)$ . The actuator faults in this paper are described by

$$u_i^F = \rho_i u_i + \sigma_i u_{is}, i = 1, 2, \dots, N, \quad (3)$$

where

$$\begin{cases} \rho_i = \text{diag}\{\rho_{i1}, \rho_{i2}, \dots, \rho_{im}\}, \\ \sigma_i = \text{diag}\{\sigma_{i1}, \sigma_{i2}, \dots, \sigma_{im}\}, \\ u_{is} = \text{col}\{u_{i1s}, u_{i2s}, \dots, u_{ims}\}. \end{cases}$$

$\rho_{ij}$  is a time-varying scalar satisfying  $0 \leq \rho_{ij} \leq 1$ .  $u_{is}$  is a constant representing the stuck value. The value  $\sigma_i \in \{0, 1\}$  depends on the fault mode. If  $\rho_{ij} = 1$ , the  $j$ th actuator is faultless; If  $0 < \rho_{ij} < 1$ , the  $j$ th actuator suffers LOE; If  $\rho_{ij} = 0, \sigma_{ij} = 0$ , the  $j$ th actuator suffers outage; If  $\rho_{ij} = 0, \sigma_{ij} = 1$ , the  $j$ th actuator gets stuck at  $u_{is}$ . Let  $\psi_i = \sigma_i u_{is} + Ed_i - u_0$ , where  $E$  is defined by  $BE = D$ . By substituting (3) into (1) and (2), one can get

$$\dot{e}_i = Ae_i + B\rho_i u_i + B\psi_i, i \in \mathcal{V}. \quad (4)$$

$e_i = x_i - x_0$  represents the tracking error of agent  $i$ . If  $u_i$  is bounded,  $u_{is}$  is bounded. Let  $\bar{\psi}_i$  denote the unknown upper bound of  $\|\psi_i\|$  such that  $\|\psi_i\| \leq \bar{\psi}_i$ . The following assumptions and lemmas are necessary to facilitate analysis.

**Assumption 1.**  $\mathcal{G}$  is undirected connected and  $\mathcal{H} \neq 0$ .

**Assumption 2.** For any fault mode  $r(B) = r(B\rho_i), i \in \mathcal{V}$ .

**Assumption 3.** For any  $i \in \mathcal{V}$ , up until  $m-1$  actuators suffer  $\rho_{ij} = 0$ , the FTC objective is still achievable.

**Lemma 1** [13]. Let Assumption 1 hold.  $\mathcal{K} > 0$  and there exists a non-singular matrix  $W$  satisfying  $W^T \mathcal{K} W = \Lambda$ , where  $W^{-1} = W^T$ ,  $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_N\}$ ,  $\lambda_1 \leq \dots \leq \lambda_N$ .

**Lemma 2** [14]. Let  $V(x) : \mathbb{R}^n \rightarrow \mathbb{R}$  be differentiable positive definite. If there exist real numbers  $\lambda \in (0, \infty)$ ,  $\gamma \in (0, 1)$ , and  $\eta \in (0, \infty)$  satisfying  $\dot{V}(x) \leq -\lambda V^\gamma(x) + \eta$ , the trajectories of  $x$  will reach the residual set  $\left\{x \mid V^\gamma(x) \leq \frac{\eta}{\lambda(1-\alpha)}\right\}$ , where  $0 < \alpha < 1$ . The settling time is estimated by  $t^* \leq \frac{V^{1-\gamma}(x(t_0))}{\lambda\alpha(1-\gamma)}$ .

**Lemma 3** [15]. For any  $a \in \mathbb{R}$  and  $0 < q \leq 1$ , then  $(\sum_{i=1}^N |a_i|)^q \leq \sum_{i=1}^N |a_i|^q$ .

**Lemma 4** [16]. Suppose that Assumption 2 holds. For any fault mode there exists a  $\mu_i > 0$  satisfying  $B\rho_i B^T \geq \mu_i BB^T$ .

### 3 Practical Finite-Time FTC for General MASs

The controller can be constructed by

$$u_i = \begin{cases} -\hat{\kappa}_{i1}B^TP\xi_i - \hat{\kappa}_{i2}\frac{B^TP\xi_i}{\|B^TP\xi_i\|}, & \|B^TP\xi_i\| \neq 0, \\ 0, & \|B^TP\xi_i\| = 0, \end{cases} \quad (5)$$

where  $P > 0$  is given by:

$$A^T P + PA - 2PBB^T P < 0.$$

$\xi_i$  is the local information defined as  $\xi_i = \sum_{j=1}^N a_{ij}(e_i - e_j) + h_i e_i$ ,  $\hat{\kappa}_{i1}$  and  $\hat{\kappa}_{i2}$  are adaptive parameters with positive initial values and updated by

$$\dot{\hat{\kappa}}_{i1} = \gamma_{i1}(-\omega_{i1}\hat{\kappa}_{i1} + \|B^TP\xi_i\|^2), \quad (6)$$

$$\dot{\hat{\kappa}}_{i2} = \gamma_{i2}(-\omega_{i2}\hat{\kappa}_{i2} + \|B^TP\xi_i\|), \quad (7)$$

where  $\gamma_{i1}$ ,  $\gamma_{i2}$ ,  $\omega_{i1}$ ,  $\omega_{i2}$  are positive learning rates.

The global tracking error can be denoted by  $e = \text{col}\{e_1, e_2, \dots, e_N\}$ . From (4), one can get

$$\dot{e} = (I \otimes A)e + (I \otimes B)\rho u + (I \otimes B)\psi, \quad (8)$$

where

$$\rho = \text{diag}\{\rho_1, \rho_2, \dots, \rho_N\}, u = \text{col}\{u_1, u_2, \dots, u_N\}, \psi = \text{col}\{\psi_1, \psi_2, \dots, \psi_N\}.$$

**Theorem 1.** Consider a leader-following system described by (1) and (2). Let Assumptions 1–3 hold. The FTC problem can be solved with the protocol (5).

The tracking error  $e$  will be restricted to a small residual set  $\left\{e \mid V^\gamma(e) \leq \frac{\eta}{\zeta(1-\alpha)}\right\}$ , where  $0 < \gamma < 1$ ,  $0 < \alpha < 1$  are design parameters. Other notions are given by:

$$\begin{aligned} V(e) &= e^T(\mathcal{K} \otimes P)e + \sum_{i=1}^N \frac{\mu_i}{\gamma_{i1}} \tilde{\kappa}_{i1}^2 + \sum_{i=1}^N \frac{\mu_i}{\gamma_{i2}} \tilde{\kappa}_{i2}^2, \\ \eta &= \frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)} + 2 \sum_{i=1}^N \sum_{j=1}^2 \omega_{ij} \mu_i (\kappa_{ij}^2 + 1), \\ \zeta &= \min\left\{\frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)}, \beta_{ij}\right\}, \end{aligned}$$

$$\tilde{\kappa}_{i1} = \hat{\kappa}_{i1} - \kappa_{i1}, \quad \tilde{\kappa}_{i2} = \hat{\kappa}_{i2} - \kappa_{i2}, \quad \kappa_{i1} = \frac{1}{\mu_i \lambda_1}, \quad \kappa_{i2} = \frac{\psi_i}{\mu_i}, \quad \beta_{ij} = 2\omega_{ij} \mu_i^{1-\gamma} \gamma_{ij}^\gamma, \\ i \in \mathcal{V}, j = 1, 2, \quad -Q = PA + A^T P - 2PBB^TP.$$

The setting time satisfies

$$t^* \leq \frac{V^{1-\gamma}(e(t_0))}{\zeta \alpha (1-\gamma)}.$$

*Proof.* Taking the time derivative of  $V$ , one obtains

$$\dot{V} = 2e^T(\mathcal{K} \otimes P)\dot{e} + 2 \sum_{i=1}^N \frac{\mu_i \tilde{\kappa}_{i1}}{\gamma_{i1}} \dot{\kappa}_{i1} + 2 \sum_{i=1}^N \frac{\mu_i \tilde{\kappa}_{i1}}{\gamma_{i2}} \dot{\kappa}_{i2}. \quad (9)$$

From (8), it has

$$e^T(\mathcal{K} \otimes P)\dot{e} = e^T(\mathcal{K} \otimes PA)e + e^T(\mathcal{K} \otimes PB)\rho u + e^T(\mathcal{K} \otimes PB)\psi.$$

Define  $\xi = \text{col}\{\xi_1, \xi_2, \dots, \xi_N\}$ . It's easy to see that  $\xi = (\mathcal{K} \otimes I)e$ . Then we have

$$e^T(\mathcal{K} \otimes P)\dot{e} = e^T(\mathcal{K} \otimes PA)e + \sum_{i=1}^N \xi_i^T PB\rho_i u_i + \sum_{i=1}^N \xi_i^T PB\psi_i. \quad (10)$$

By Lemma 4, from (5) one obtains

$$\begin{aligned} \sum_{i=1}^N \xi_i^T PB\rho_i u_i &= - \sum_{i=1}^N \hat{\kappa}_{i1} \xi_i^T PB\rho_i B^T P \xi_i - \hat{\kappa}_{i2} \frac{\xi_i^T PB\rho_i B^T P \xi_i}{\|B^T P \xi_i\|} \\ &\leq - \sum_{i=1}^N \mu_i \hat{\kappa}_{i1} \|B^T P \xi_i\|^2 - \mu_i \hat{\kappa}_{i2} \|B^T P \xi_i\|. \end{aligned}$$

Then we continue (10) as follows

$$\begin{aligned} e^T(\mathcal{K} \otimes P)\dot{e} &\leq e^T(\mathcal{K} \otimes PA)e - \sum_{i=1}^N \mu_i \tilde{\kappa}_{i1} \|B^T P \xi_i\|^2 - \sum_{i=1}^N \mu_i \tilde{\kappa}_{i2} \|B^T P \xi_i\| \\ &\quad - \sum_{i=1}^N \mu_i \kappa_{i1} \|B^T P \xi_i\|^2, \end{aligned} \quad (11)$$

where we use the fact that  $\mu_i k_{i2} \geq \|\psi_i\|$ . From (6), (7), (9), and (11), it has

$$\dot{V} \leq 2e^T(\mathcal{K} \otimes PA)e - 2 \sum_{i=1}^N \mu_i \kappa_{i1} \|B^T P \xi_i\|^2 - 2 \sum_{i=1}^N \mu_i \omega_{i1} \tilde{\kappa}_{i1} \hat{\kappa}_{i1} - 2 \sum_{i=1}^N \mu_i \omega_{i2} \tilde{\kappa}_{i2} \hat{\kappa}_{i2}.$$

Notice that  $\mu_i \kappa_{i1} \lambda_1 = 1$ , according to Lemma 1, we have

$$\begin{aligned} \dot{V} &\leq -\lambda_{\min}(Q) e^T (\mathcal{K} \otimes I) e - 2 \sum_{i=1}^N \sum_{j=1}^2 \mu_i \omega_{ij} \tilde{\kappa}_{ij} \hat{\kappa}_{ij}. \\ &\leq -\frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)} e^T (\mathcal{K} \otimes P) e - 2 \sum_{i=1}^N \sum_{j=1}^2 \mu_i \omega_{ij} \tilde{\kappa}_{ij} \hat{\kappa}_{ij}. \end{aligned}$$

Furthermore, if  $e^T(\mathcal{K} \otimes P)e \geq 1$ , we have the following inequality

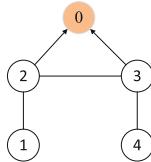
$$[e^T(\mathcal{K} \otimes P)e]^\gamma \leq e^T(\mathcal{K} \otimes P)e.$$

Otherwise if  $e^T(\mathcal{K} \otimes P)e < 1$ ,

$$[e^T(\mathcal{K} \otimes P)e]^\gamma \leq e^T(\mathcal{K} \otimes P)e + 1.$$

Then we have

$$\dot{V} \leq -\frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)} [e^T(\mathcal{K} \otimes P)e]^\gamma - 2 \sum_{i=1}^N \sum_{j=1}^2 \mu_i \omega_{ij} \tilde{\kappa}_{ij} \hat{\kappa}_{ij} + \frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)}. \quad (12)$$



**Fig. 1.** The communication topology.

Note that  $-\tilde{\kappa}_{ij} \hat{\kappa}_{ij} \leq -\tilde{\kappa}_{ij}^2 + \kappa_{ij}^2$ . Define  $\beta_{ij} = 2\omega_{ij} \mu_i^{1-\gamma} \gamma_{ij}^\gamma$ . If  $\|\tilde{\kappa}_{ij}^2\| \geq 1$ ,

$$\frac{\beta_{ij} \mu_i^\gamma}{\gamma_{ij}^\gamma} \tilde{\kappa}_{ij}^{2\gamma} \leq \frac{\beta_{ij} \mu_i^\gamma}{\gamma_{ij}^\gamma} \tilde{\kappa}_{ij}^2 = 2\omega_{ij} \mu_i \tilde{\kappa}_{ij}^2. \quad (13)$$

If  $\|\tilde{\kappa}_{ij}^2\| < 1$ ,

$$\frac{\beta_{ij} \mu_i^\gamma}{\gamma_{ij}^\gamma} \tilde{\kappa}_{ij}^{2\gamma} \leq \frac{\beta_{ij} \mu_i^\gamma}{\gamma_{ij}^\gamma} = 2\omega_{ij} \mu_i. \quad (14)$$

From (13) and (14), we have  $\frac{\beta_{ij}\mu_i^\gamma}{\gamma_{ij}^\gamma}\tilde{\kappa}_{ij}^{2\gamma} \leq 2\omega_{ij}\mu_i\tilde{\kappa}_{ij}^2 + 2\omega_{ij}\mu_i$ . Applying it into (12), one obtains

$$\dot{V} \leq -\frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)}[e^T(\mathcal{K} \otimes P)e]^\gamma - \sum_{i=1}^N \sum_{j=1}^2 \frac{\beta_{ij}\mu_i^\gamma}{\gamma_{ij}^\gamma}\kappa_{ij}^{2\gamma} + \eta,$$

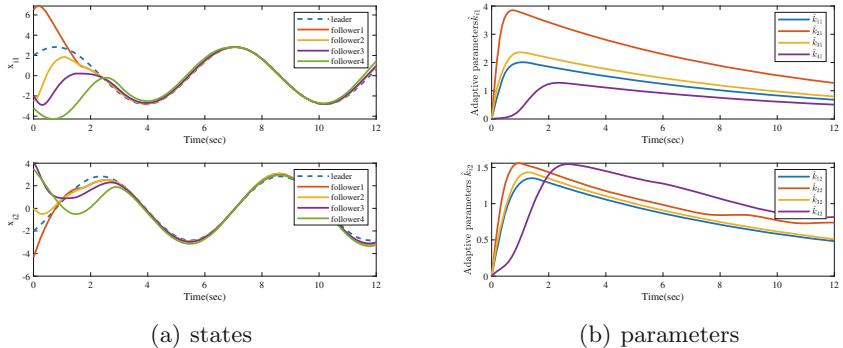
In view of the definitions of  $\zeta$  and  $\eta$ , based on Lemma 3, one can conclude  $\dot{V} \leq -\zeta V^\gamma + \eta$ . According to Lemma 2, the proof is completed.

## 4 Simulations

The considered MAS is composed of a leader and four followers. Figure 1 depicts the corresponding communication topology. The matrix  $\mathcal{K}$  is given by  $\mathcal{K} =$

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 3 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}. A, B, \text{ and } D \text{ are chosen as } A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}, D = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The time-varying disturbance  $d_i$  is given by  $d_i = [0, \sin(t)]^T, i \in \mathcal{V}$ . Suppose that agent 1 and agent 3 are faultless. At the beginning time, agent 2 gets LOE and outage,  $\rho_2$  is given by  $\text{diag}\{1, 1, 0, 0.5 + 0.2|\sin(t)|\}$ . Agent 4 gets LOE and stuck.  $\rho_4$  is given by  $\text{diag}\{1, 0.7 + 0.1\cos(t), 1, 0\}$ ,  $\sigma_{24} \neq 0$ . Set  $x_0(0) = [2, -2]^T$ .  $x_i(0), i = 1, 2, 3, 4$ . are chosen from the box  $[-10, 10] \times [-10, 10]$  randomly. Take  $\hat{\kappa}_{i1}(0) = \hat{\kappa}_{i2}(0) = 0$ ,  $\gamma_{i1} = \gamma_{i2} = 1$ , and  $\omega_{i1} = \omega_{i2} = 0.1$ . According to Theorem 1, we can obtain  $t^* \leq 43.2s$ . The trajectories of the states are depicted in Fig. 2(a), which show that the tracking error will converge to a small neighbourhood of zero. Figure 2(b) presents the trajectories of the adaptive parameters  $\hat{\kappa}_{i1}$  and  $\hat{\kappa}_{i2}$ . All the signals in our proposed scheme are bounded. This validates the theoretical results in Sect. 3.



**Fig. 2.** Response curves of the signals in the FTC scheme.

## 5 Conclusion

In this paper, the adaptive finite-time FTC for general MASs is studied with multiple actuator faults including LOE, stuck, and outage. Some adaptive updating laws are proposed to estimate the faults. Then a fully distributed control strategy is put forward, under which the tracking error will be restricted to a small neighborhood of zero and all the adaptive parameters are bounded. The fixed-time FTC for general MAS is an interesting topic and will be addressed in future.

**Acknowledgments.** This work is supported by the National Natural Science Foundation of China under Grant No.61973175, and the Fundamental Research Funds for the Central Universities, Nankai University under Grant No.63201196.

## References

1. Dong, X., Hu, G.: Time-varying formation control for general linear multi-agent systems with switching directed topologies. *Automatica* **73**, 47–55 (2016). <https://doi.org/10.1016/j.automatica.2016.06.024>
2. Wang, F., Liu, Z., Chen, Z., Wang, S.: Containment control for second-order nonlinear multi-agent systems with intermittent communications. *Int. J. Syst. Sci.* **50**(5), 919–934 (2019). <https://doi.org/10.1080/00207721.2019.1585997>
3. Zhao, X., Guan, Z.J., Zhang, X., Chen, C.: Flocking of multi-agent nonholonomic systems with unknown leader dynamics and relative measurements. *Int. J. Robust Nonlinear Cont.* **27**(17), 3685–3702 (2017). <https://doi.org/10.1002/rnc.3762>
4. Ren, W., Beard, R., Atkins, E.: A survey of consensus problems in multi-agent coordination. In: Proceedings of American Control Conference, Portland, pp. 1859–1864 (2005). <https://doi.org/10.1109/ACC.2005.1470239>
5. Zheng, Y., Ma, J., Wang, L.: Consensus of hybrid multi-agent systems. *IEEE Trans. Neural Netw. Learn. Syst.* **29**(4), 1359–1365 (2018). <https://doi.org/10.1109/TNNLS.2017.2651402>
6. Qin, J., Ma, Q., Shi, Y., Wang, L.: Recent advances in consensus of multi-agent systems: a brief survey. *IEEE Trans. Ind. Electron.* **64**(6), 4972–4983 (2017). <https://doi.org/10.1109/TIE.2016.2636810>
7. Jiang, F., Wang, L.: Finite-time information consensus for multiagent systems with fixed and switching topologies. *Physica D* **238**, 1550–1560 (2009). <https://doi.org/10.1016/j.physd.2009.04.011>
8. Zhang, H., Yue, D., Yin, X., Hu, S., Dou, C.X.: Finite-time distributed event-triggered consensus control for multi-agent systems. *Inf. Sci.* **339**, 132–142 (2016). <https://doi.org/10.1016/j.ins.2015.12.031>
9. Zou, W., Shi, P., Xiang, Z., Shi, Y.: Finite-time consensus of second-order switched nonlinear multi-agent systems. *IEEE Trans. Neural Netw. Learn. Syst.* **31**(5), 1757–1762 (2020). <https://doi.org/10.1109/TNNLS.2019.2920880>
10. Fu, J., Wang, Q., Wang, J.: Robust finite-time consensus tracking for second-order multi-agent systems with input saturation under general directed communication graphs. *Int. J. Control.* **92**(8), 1785–1795 (2017). <https://doi.org/10.1080/00207179.2017.1411609>
11. Zhang, G., Qin, J., Zheng, W.X., Kang, Y.: Fault-tolerant coordination control for second-order multi-agent systems with partial actuator effectiveness. *Inf. Sci.* **423**, 115–127 (2018). <https://doi.org/10.1016/j.ins.2017.09.043>

12. Chen, S., Ho, D.W.C., Li, L., Liu, M.: Fault-tolerant consensus of multi-agent system with distributed adaptive protocol. *IEEE Trans. Cybern.* **45**(10), 2142–2155 (2015). <https://doi.org/10.1109/tcyb.2014.2366204>
13. Hong, Y.G., Hu, H.P., Gao, L.X.: Tracking control for multi-agent consensus with an active leader and variable topology. *Automatica* **42**(7), 1177–1182 (2006). <https://doi.org/10.1016/j.automatica.2006.02.013>
14. Wang, H., Chen, B., Lin, C., Sun, Y., Wang, F.: Adaptive finite-time control for a class of uncertain high-order non-linear systems based on fuzzy approximation. *IET Control Theory Appl.* **11**(5), 677–684 (2017). <https://doi.org/10.1049/iet-cta.2016.0947>
15. Hardy, G., Littlewood, J., Polya, G.: Inequalities. Cambridge University Press, Cambridge (1952)
16. Wu, L.-B., Yang, G.-H., Ye, D.: Robust adaptive fault-tolerant control for linear systems with actuator failures and mismatched parameter uncertainties. *IET Control Theory Appl.* **8**(6), 441–449 (2014). <https://doi.org/10.1049/iet-cta.2013.0334>



# Domain Decomposition Strategies for Developing Parallel Unstructured Mesh Generation Software Based on PadMesh

Fengshun Lu, Xiong Jiang, Xinbiao Bao, Long Qi, and Yongheng Guo<sup>(✉)</sup>

China Aerodynamics Research and Development Center, Mianyang 621000, China  
[matrixspace@163.com](mailto:matrixspace@163.com)

**Abstract.** Mesh generation is a key preprocessing step in the computational fluid dynamics (CFD). The scale and quality of the generated mesh have an important influence on the CFD simulation results. In order to meet the requirements of large-scale unstructured mesh generation, the application of PadMesh (a distributed and parallel mesh generation software framework based on C/S architecture) in developing parallel unstructured mesh generation software is investigated. Two domain decomposition strategies have been proposed, namely the region growing strategy and the bounding box strategy. Three practical configurations have been utilized to verify the two proposed strategies. Experimental results show that the region growing strategy can generate subdomains with good load balance, while the bounding box strategy is suitable for the configurations with symmetrical features and can obtain subdomains with high load balance and good data locality.

**Keywords:** Unstructured mesh · Domain decomposition · Region growing strategy · Bounding box strategy · Software framework

## 1 Introduction

Mesh generation is one of the pre-processing steps for computational fluid dynamics (CFD) analysis. The size and quality of generated meshes greatly affect the CFD simulation results [1–4]. For the industrial CFD applications, the mesh scale has already reached hundreds of millions in the Reynolds-averaged Navier-Stokes (RANS) simulation scenario, even 10 billions in the fully-resolved large eddy simulation [5]. Furthermore, the computational meshes with hundreds of billions of grid points have been utilized in the fundamental research [6]. It can be noticed that the scale of computational meshes needed by large-scale CFD applications has become larger and larger, which poses a great challenge to the development of interactive mesh generation software.

The current mainstream interactive mesh generation software is usually installed on a desktop or laptop computer and usually runs as stand-alone software, such as Pointwise [7], ICEM [8], etc. It has the following two shortcomings.

© The Editor(s) (if applicable) and The Author(s), under exclusive license

to Springer Nature Singapore Pte Ltd. 2021

Y. Jia et al. (Eds.): CISC 2020, LNEE 705, pp. 194–205, 2021.

[https://doi.org/10.1007/978-981-15-8450-3\\_21](https://doi.org/10.1007/978-981-15-8450-3_21)

On the one hand, it only provides limited capacity bounded by the computational resources of the host computer and therefore can hardly satisfy the enormous demand for meshes from the large-scale CFD applications; on the other hand, users cannot access the mesh generation service at any time from any where. Consequently, parallel and distributing techniques [9–12] are crucial for developing mesh generation software that copes with the ever-increasing demand in mesh scale.

To address with the aforementioned issues, China Aerodynamics Research and Development Center has developed a Parallel And Distributed Mesh generation software framework (PadMesh) [13], which can be used as the infrastructure of constructing interactive mesh generation software. In this research, we mainly concentrate on the domain decomposition strategy for the development of parallel unstructured mesh generation software based on PadMesh.

The remainder of the paper is organized as follows. Section 2 gives a simple introduction of PadMesh. Section 3 demonstrates our proposed domain decomposition strategies. Section 4 describes the experimental validation by practical configurations. Finally, Sect. 5 concludes our paper and outlooks the future work.

## 2 PadMesh: A Parallel and Distributed Mesh Generation Software Framework

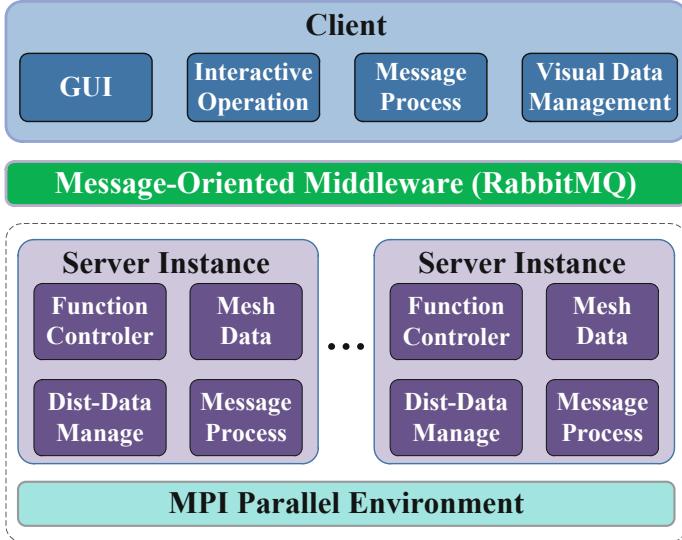
PadMesh can serve as the infrastructure for developing parallel and distributed interactive mesh generation software. Interested readers can refer to our previous work [13] fore more information. We hereby only give a simple instruction to PadMesh.

The design of PadMesh follows the software engineering principles such as modularity, high scalability, and high availability. As shown in Fig. 1, PadMesh has a hierarchical architecture and consists of four layers, namely client layer, message middleware layer, server layer and MPI parallel environment layer.

### 2.1 Client Layer

The client provides users with graphical tools for mesh generation and it includes four modules, namely the graphical user interface (GUI), interactive operation, message processing, and visual data management module. Their functionalities are listed as follows:

- The GUI module provides various interactive controls.
- The interactive operation module responds to keyboard or mouse events triggered by users.
- The message processing module is responsible for communicating with the message middleware for sending interactive commands and receiving newly generated visual data.
- The visual data management module efficiently manages the visual data with abundant volume.



**Fig. 1.** PadMesh architecture [13]

## 2.2 Message Middleware Layer

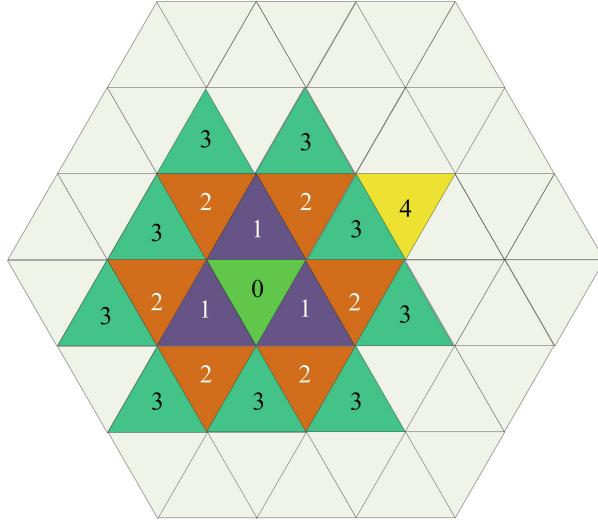
Message middleware layer is located between the client and the server, which performs as a bridge connecting these two layers. It establishes a communication layer that isolates interface details between application developers and operating systems/networks. Therefore, components independently developed and running on different platforms can communicate with each other.

Nowadays there are several message middleware products, such as RocketMQ [14], RabbitMQ [15], Redis [16] and Kafka [17]. Considering the low latency, data durability, programming language and other factors, we chose RabbitMQ as the message middleware of PadMesh.

## 2.3 Server Layer

According to the parallel execution configuration, the server layer may contain multiple server instances. Each server instance has the following four modules:

- The function controller module implements business logic for a variety of mesh generation operations.
- The mesh data module stores and maintains the entire mesh data structure.
- The distributed data management module deals with the storage of mesh data within multiple processes.
- The message processing module receives various commands from the message middleware and sends the newly generated visual data to it.



**Fig. 2.** Sketch of the region growing strategy

## 2.4 MPI Parallel Environment Layer

The MPI (Message Passing Interface) parallel environment offers convenience for the parallel programming practice, which encapsulates various standard MPI functions and provides specific interfaces to parallel mesh generation.

## 3 Domain Decomposition Strategies

The development of unstructured mesh generation software based on the PadMesh framework involves two aspects. For the client application, software engineers have to implement friendly GUI, efficiently manage the visual data, and accomplish the message processing work. For the server application, they need to deal with the management of distributed mesh data, the implementation of mesh generation business logics, and the efficient transmission of massive visual data. Among them, the management of distributed mesh data can be further subdivided into data structure design, storage mechanism, *domain decomposition strategy*, data synchronization mechanism, etc.

In this research, we focus on the domain decomposition strategy of distributed unstructured mesh, namely how to properly allocate mesh data within the memory space of multiple parallel processes. When performing domain decomposition, load balancing is an issue that must be considered and can be affected by computational cost, communication pattern, and underlying hardware performance. The precise prediction of load balance metrics is a very complicated process. Therefore, we simply takes the number of cells in each subdomain as the basis for measuring load balance performance.

**Algorithm 1:** Region growing strategy**Input:**

The number of subdomains,  $n$ ;  
 The number of cells in each subdomain,  $NP^i (1 \leq i \leq n)$ .

**Output:**  $n$  subdomains.

---

```

1 Set the current subdomain index  $\kappa$  to 0.
2 Push the first cell  $\theta_0$  into the double-ended queue  $\Psi$ .
3 while  $\Psi$  is not empty do
4   Get the head element  $\theta$  from  $\Psi$  and set the finished flag  $\theta_{done}$  to 1.
5   if  $\theta$  belongs to none of the subdomains then
6     Assign  $\theta$  to the subdomain with index  $\kappa$ .
7   Get the adjacent cells set  $\Phi$  of  $\theta$ .
8   for each element  $\phi \in \Phi$  do
9     if the finished flag  $\phi_{done} = 0$  then
10      Push  $\phi$  to the tail of  $\Psi$ .
11      if  $\phi$  belongs to none of the subdomains then
12        Assign  $\phi$  to the subdomain with index  $\kappa$ .
13   Pop out the head element from  $\Psi$ .
14   if the number  $\omega$  of elements in subdomain  $\kappa$  satisfies  $\omega \geq NP^\kappa$  then
15      $\kappa = \kappa + 1$ .
16   if  $\kappa \geq n$  then
17     Algorithm stops.

```

---

### 3.1 Region Growing Strategy

The unstructured surface meshes are constructed with two-dimensional polygons, such as triangles. Unlike structured meshes, they do not have determined topological structure and none implicit relationship exists between different units. Consequently, users need to explicitly assign the relationship between adjacent elements.

Algorithm 1 demonstrates the basic idea of region growing strategy. First, a random cell is chosen as the starting element to traverse all the mesh. Second, its adjacent cells are gradually visited and put each of them into the list to be visited. Finally, all the cells are assigned to certain subdomain.

As shown in Fig. 2, the cell labeled with 0 is selected as the starting element  $\theta_0$  (Line 2) and it has three adjacent cells  $\Phi$  labeled with 1 (Line 3). Each element  $\phi \in \Phi$  has not been treated and consequently it is pushed to  $\Psi$  (Line 12). After all the cells in  $\Phi$  have been visited, the head element is popped out from the top of  $\Psi$  (Line 13) and currently only the cells labeled with 1 reside in  $\Psi$ . The algorithm stops either the queue  $\Psi$  is empty or the condition  $\kappa \geq n$  is satisfied. Figure 2 indicates that the trajectory of visited cells walks from inside to outside. Concretely speaking, the cell 0 first links three cells labelled 1, then six elements

---

**Algorithm 2:** Bounding box strategy

---

**Input:**

- The number of subdomains,  $n$ ;
- The set  $\Omega$  containing  $n$  bounding boxes  $B_i (1 \leq i \leq n)$ .

**Output:**  $n$  subdomains.

---

```

1 for each element  $\theta$  in the mesh do
2   Let the index of targeting subdomain be  $n_t$ .
3   Get the vertexes set  $V = \{v_1, v_2, \dots, v_m\}$ .
4   Initiate the counter array  $inCnt[n] = \{0\}$ .
5   for each  $B_\kappa \in \Omega$  do
6     for each vertex  $v_j \in V$  do
7       if  $v_j$  falls in  $B_\kappa$  then
8          $inCnt[\kappa] = inCnt[\kappa] + 1$ .
9       if  $inCnt[\kappa] == m$  then
10         $n_t = \kappa$ .
11      else
12        Find the maximal value in the array  $inCnt$  whose index is  $\kappa_t$ .
13         $n_t = \kappa_t$ .
14   Assign the cell  $\theta$  to the subdomain surrounded by bounding box  $B_{n_t}$ .

```

---

with label 2 and nine with label 3. The growing pattern motivates the name of the strategy.

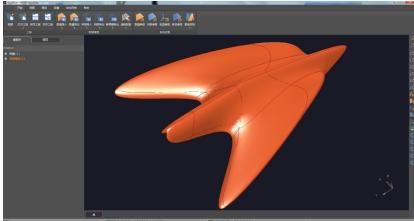
### 3.2 Bounding Box Strategy

Algorithm 2 presents the bounding box strategy, where the bounding box can be given by interactive operation with GUI or parameter file. A traversing operation for all elements in the surface mesh is performed to determine the bounding box  $B_i (1 \leq i \leq n)$  that holds element  $\theta$ .

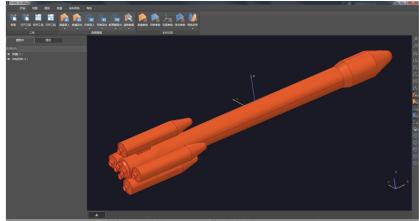
Lines from 3 to 14 presents the calculation of bounding box index for  $\theta$ . The vertexes  $V$  of  $\theta$  are obtained, then the coordinates of vertex  $v_i$  are utilized to achieve the index of the relevant bounding box. In special cases, a vertex may locate on the borders of multiple bounding boxes, therefore we take advantage of an array  $inCnt$  to store the number of vertexes falling in each  $B_\kappa$ . If  $inCnt[\kappa]$  happens to be equal to  $m$ , then all the  $m$  vertexes of cell  $\theta$  are located in the  $\kappa$ th bounding box and the targeting subdomain index  $n_t$  should be  $\kappa$ ; otherwise,  $n_t$  is assigned to the index  $\kappa_t$  of  $inCnt$  with the maximal value. When the algorithm stops, all the cells have been partitioned to various subdomains.

## 4 Experimental Verification and Analysis

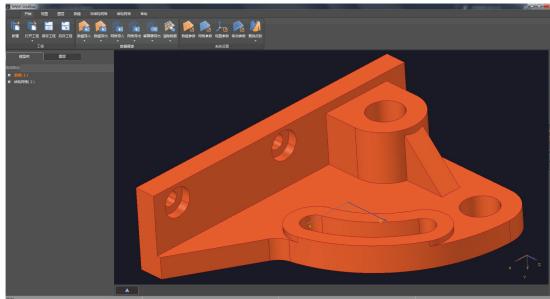
To verify the two proposed domain decomposition strategies, three configurations shown in Fig. 3 are selected for the experimental verification. The experimental settings are introduced, followed by the presentation of domain decomposition results for various surface meshes.



(a) A bionic aircraft configuration



(b) A rocket configuration



(c) A mechanical component configuration

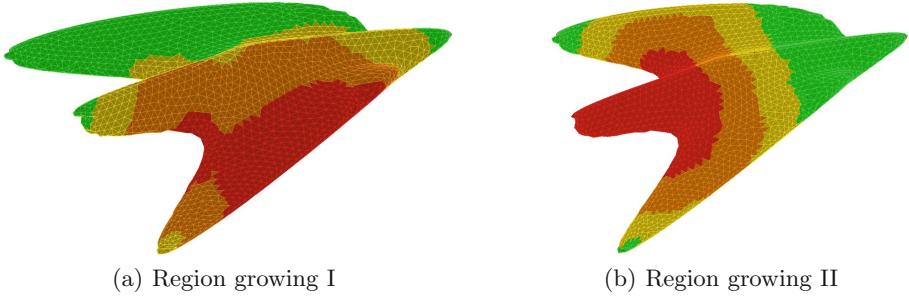
**Fig. 3.** Experimental configurations

**Table 1.** Configuration information

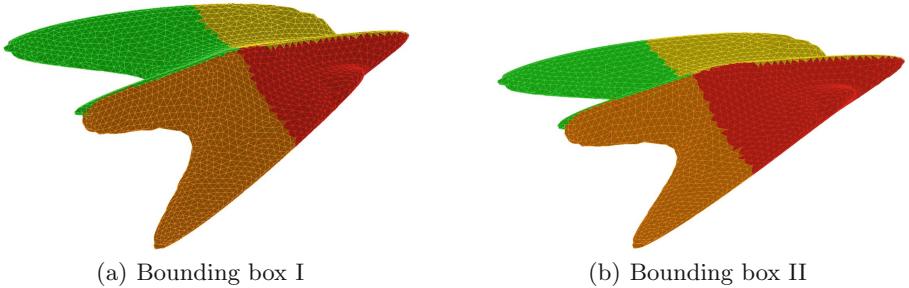
Configuration	Vertex number	Cell number
A bionic aircraft	3221	6438
A rocket	2252	4500
A mechanical component	1452	2920

### 4.1 Experimental Setup

Figure 3 shows the screenshots for experimental configurations in the software NNW-GridStar [13], namely a bionic aircraft (Fig. 3(a)), a rocket (Fig. 3(b)), and a mechanical component (Fig. 3(c)). The region growing and bounding box strategies are utilized to perform the domain decomposition of unstructured surface mesh for all the three configurations. For the sake of clarity, we only decompose the surface mesh into four subdomains.



**Fig. 4.** Domain decomposition based on region growing strategy for surface mesh of the bionic aircraft



**Fig. 5.** Domain decomposition based on bounding box strategy for surface mesh of the bionic aircraft

We employ the following metric  $\eta$  to measure the quality of domain decomposition and

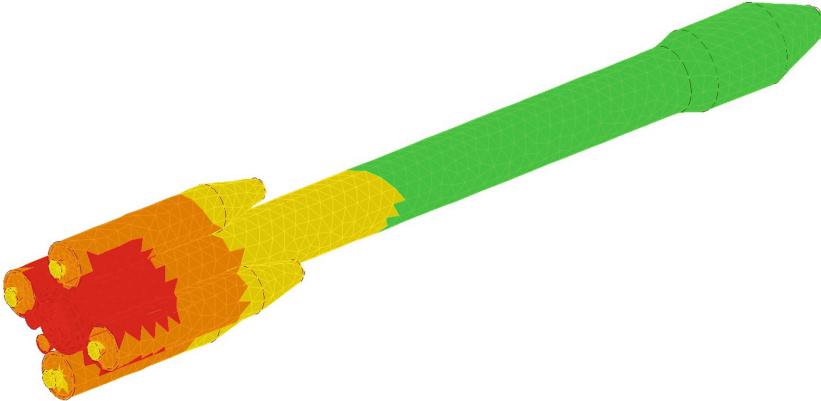
$$\eta = \frac{\min NP^\kappa}{\max NP^\kappa} \times 100\%,$$

where  $NP^\kappa$  denotes the number of cells in the  $\kappa$ th subdomain.

## 4.2 Domain Decomposition Results for the Surface Mesh of a Bionic Aircraft

Figure 4 shows the domain decomposition results for the surface mesh of bionic aircraft based on the region growing strategy. Each color represents a subdomain. Concretely speaking, Fig. 4(a) depicts the result by using the first cell as the starting  $\theta_0$  and Fig. 4(b) is relevant to the last cell as  $\theta_0$ . It can be seen that there are disconnected areas for certain subdomain and the data locality is not good. For instance, the green subdomain in Fig. 4(b) is discontinuous.

Figure 5 demonstrates the subdomains of bionic aircraft surface mesh based on the bounding box strategy. Unlike the results based on the region growing strategy, the generated subdomains show good data locality, which can be validated in Fig. 5(a) and (b).



**Fig. 6.** Domain decomposition based on region growing strategy for surface mesh of the rocket

Table 2 lists the number of cells in different subdomains of the bionic aircraft surface mesh. For the region growing strategy, the number of cells in all subdomains are basically the same, and  $\eta$  is 99.87%. However, the position of the bounding box has a great impact on the load balance metric. For instance, the number of cells for subdomains generated by bounding box I is quite different ( $\eta = 55.21\%$ ), while the bounding box II can obtain load-balanced subdomains with  $\eta = 94.90\%$ .

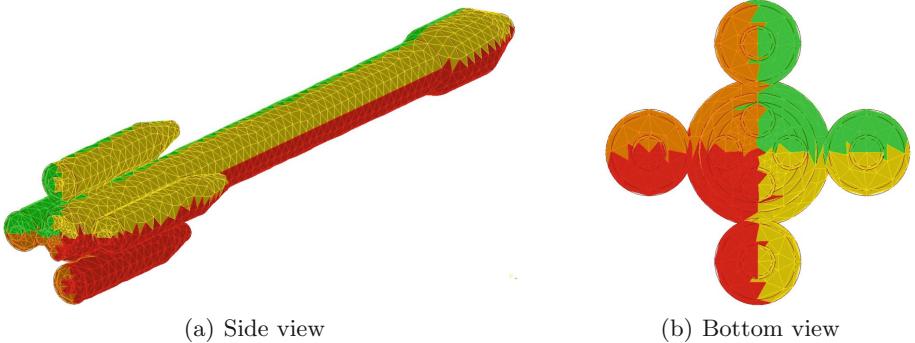
**Table 2.** Subdomain statistics for surface mesh of certain bionic air vehicle

Domain decomposition strategy		subdom1	subdom2	subdom3	subdom4
Region growing	Starting vertex I	1611	1609	1609	1609
	Starting vertex II	1611	1609	1609	1609
Bounding box 1	Position I	1140	2053	1180	2065
	Position II	1563	1632	1596	1647

### 4.3 Domain Decomposition Results for the Surface Mesh of a Rocket

Figure 6 depicts the domain decomposition for a rocket surface mesh based on the region growing strategy, using four colors to separate the subdomains. Statistic information in Table 3 shows that the subdomains have a good load balance metric (up to 100%). However, it does not perform well in terms of data locality. For example, there are discontinuities in the yellow subdomain, which consists of cells covering the racket body and four nozzles at the bottom.

Figure 7 presents the domain decomposition results based on the bounding box strategy, and Fig. 7(b) shows the bottom view of the rocket. It can be seen



**Fig. 7.** Domain decomposition based on bounding box strategy for surface mesh of the rocket

**Table 3.** Subdomain statistics for surface mesh of certain rocket

Domain decomposition strategy	subdom1	subdom2	subdom3	subdom4
Region growing	1125	1125	1125	1125
Bounding box	1150	1105	1130	1115

that all cells in each subdomain are adjacent, which means a better data locality compared to the region growing scenario. Furthermore, Table 3 indicates that the load balancing metric reaches 96.08%.

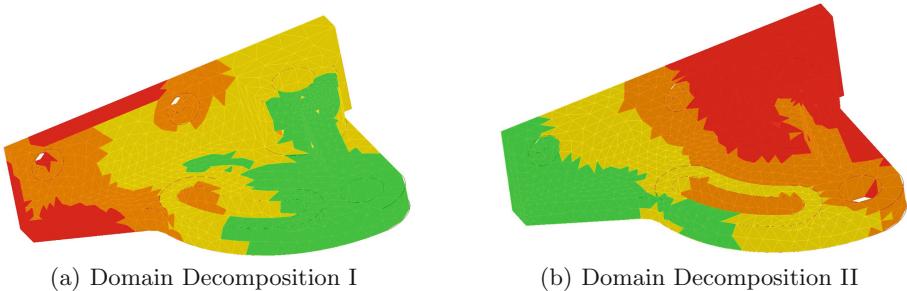
We observe that the bounding box strategy can obtain subdomains with good data locality and high load balance metric for configurations with regular or symmetric shapes. Therefore, it is recommended to utilize the bounding box strategy to perform the domain decomposition for a rocket-like configuration.

#### 4.4 Domain Decomposition Results for the Surface Mesh of a Mechanical Component

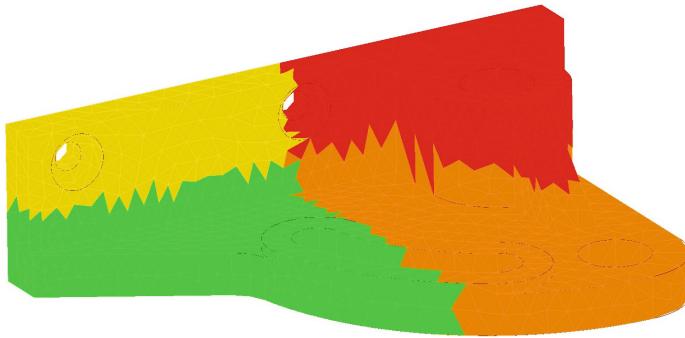
Figure 8 shows the subdomains of a mechanical component based on the region growing strategy, and Table 4 gives the statistical information of each subdomain cells. Similar to the previous two examples, the subdomains obtained according to the region growing strategy have very good load balance property, and the relevant metric  $\eta$  reaches 98.63%. However, the data locality of the subdomains are not well.

Figure 9 demonstrates the domain decomposition based on the bounding box strategy. It can be seen that the strategy can ensure a good data locality for each subdomain, but results in a rather low load balancing metric with  $\eta = 29.85\%$ .

For irregular configurations, we notice that the bounding box strategy can hardly guarantee the load balance performance, and therefore recommend users to adopt the region growing domain decomposition strategy.



**Fig. 8.** Domain decomposition based on region growing strategy for surface mesh of mechanical component



**Fig. 9.** Domain decomposition based on bounding box strategy for surface mesh of mechanical component

**Table 4.** Subdomain statistics for surface mesh of certain mechanical component

Domain decomposition strategy	subdom1	subdom2	subdom3	subdom4
Region growing	730	730	730	720
Bounding box	541	891	342	1146

## 5 Conclusion

Mesh generation is a key pre-processing procedure for CFD. In order to cope with the demand on ultra-large mesh generation, we have proposed a parallel and distributed mesh generation software framework (PadMesh), which can serve as the fundamental infrastructure for developing interactive mesh generation software.

In this paper, we focused on the domain decomposition of unstructured meshes and proposed two strategies. Three configurations have been utilized to verify the proposed strategies. Experimental results indicate two observations. On the one hand, the region growing strategy can produce subdomains with

better load balance; on the other hand, the bounding box strategy is suitable for regular or symmetric configurations, and can obtain subdomains with high load balance metric and good data locality.

**Acknowledgements.** This work was supported by the Pre-Research Generic Technology Project under grant no. 41406030201 and the National Numerical Windtunnel Project.

## References

1. Baker, T.J.: Mesh generation: art or science? *Prog. Aerosp. Sci.* **41**, 29–63 (2005)
2. Chen, J., Zhao, D., Huang, Z., Zheng, Y., Wang, D.: Improvements in the reliability and element quality of parallel tetrahedral mesh generation. *Int. J. Numer. Meth. Eng.* **92**, 671–693 (2012)
3. Loseille, A., Menier, V., Alauzet, F.: Parallel generation of large-size adapted meshes. *Procedia Eng.* **124**, 57–69 (2015)
4. Li, X., Yu, W., Liu, C.: Geometry-aware partitioning of complex domains for parallel quad meshing. *Comput. Aided Des.* **85**, 20–33 (2017)
5. Nishikawa, T., Yamade, Y., Sakuma, M., Kato, C.: Application of fully-resolved large eddy simulation to KVLCC2. *J. Japan Soc. Naval Archit. Ocean Eng.* **16**, 1–9 (2012)
6. Bermejo-Moreno, I., Bodart, J., Larsson, J., Barney, B.M., Nichols, J.W., Jones, S.: Solving the compressible navier-stokes equations on up to 1.97 million cores and 4.1 trillion grid points. In: Proceedings of SC 2013: The International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1–10. IEEE Computer Society, Denver (2013)
7. Pointwise, Inc.: Pointwise Homepage (2019). <https://pointwise.com/>
8. ANSYS, Inc.: ICEM Homepage (2019). <https://www.ansys.com/products/platform/ansys-meshing>
9. Ekelschot, D., Ceze, M., Garai, A., Murman, S.M.: Parallel high-order anisotropic meshing using discrete metric tensors. In: Proceedings of AIAA Scitech 2019 Forum, pp. 1–14. AIAA (2019)
10. Sarkarati, M., Briess, K., Kayal, H.: Design and implementation of a remote, server-client-based telemetry retrieval and monitoring system. In: Proceedings of SpaceOps 2006 Conference, pp. 1–8. AIAA (2006)
11. Lombillo, I., Blanco, H., Pereda, J., Villegas, L., Carrasco, C., Balbas, J.: Structural health monitoring of a damaged church: design of an integrated platform of electronic instrumentation, data acquisition and client/server software. *Struct. Control Health Monitor.* **23**, 69–81 (2016)
12. Pedone, G., Mezgar, I.: Model similarity evidence and interoperability affinity in cloud-ready industry 4.0 technologies. *Comput. Ind.* **100**, 278–286 (2018)
13. Lu, F., Chen, B., Qi, L., Liu, Y., Pang, Y., Zhou, J., Jiang, X.: PaDMesh: a parallel and distributed framework for interactive mesh generation software. *Eng. Comput.* (2020). <https://doi.org/10.1007/s00366-020-01049-0>
14. The Apache Software Foundation: RocketMQ Homepage (2019). <https://rocketmq.apache.org/>
15. SpringSource: RabbitMQ Homepage (2019). <https://www.rabbitmq.com/>
16. Redis Labs: Redis Homepage (2019). <https://redis.io/>
17. The Apache Software Foundation: Kafka Homepage (2019). <http://kafka.apache.org/>



# Dynamic Trajectory Prediction for Continuous Descend Operations Based on Unscented Kalman Filter

Jun Zhang, Guoqing Wang, and Gang Xiao<sup>(✉)</sup>

School of Aeronautics and Astronautics, Shanghai Jiao Tong University,  
Shanghai 200240, China  
[xiaogang@sjtu.edu.cn](mailto:xiaogang@sjtu.edu.cn)

**Abstract.** Aiming to reduce the uncertainty of the trajectory of Continuous Descent Operations (CDO) in high-density airspace, this paper combines the unscented transformation and Kalman filtering, and a dynamic flight trajectory prediction algorithm based on Unscented Kalman Filter (UKF) is proposed. Firstly, according to the real-time ADS-B data, the position and speed of the aircraft are converted. Then, to improve the prediction accuracy of the dynamic flight trajectory, the aircraft state equations are processed by unscented transformation, and a dynamic trajectory prediction model based on UKF is established. Finally, taking the CES2492 flight as an example, a fast-time simulation is carried out, and the results are compared with those obtained by the traditional Kalman filter and the extended Kalman filter. The simulation results show that the method proposed in this paper can improve the filtering effect and effectively enhance the trajectory prediction accuracy.

**Keywords:** Unscented transformation · Unscented Kalman filter · Trajectory prediction · Continuous Descent Operations

## 1 Introduction

To improve flight efficiency and airspace traffic flow, the Civil Aviation Organization (ICAO) released the Continuous Descent Operations (CDO) Manual in 2010, and proposed the concept of Trajectory Based Operation (TBO) in the “Aviation System Block Upgrades (ASBU)” plan [1]. CDO requires the aircraft to descend continuously from the top of descent prior to final approach fix at a constant descent angle. The core technology of TBO is to accurately predict the flight time and trajectory of the aircraft, so as to ensure that the aircraft reaches the designated position at controlled time [2]. The aircraft is affected by nonlinear random interference such as aerodynamic drag, gravity, wind speed and wind direction during descent, which leads to uncertainty of the descent trajectory. Therefore, it is necessary to precisely predict the real-time dynamic trajectory of CDO, which has important practical significance for enhancing flight safety and improving terminal airspace flow.

With the continuous updating of communication, navigation, surveillance technology and airborne equipments, there are three main methods for predicting aircraft trajectory: trajectory prediction algorithm based on data mining, prediction research based on aircraft mass-point model, and hybrid estimation method. Hong et al. [3] analyzed the historical trajectory by clustering method and combined the probability information to predict the trajectory of the aircraft. However, this method requires an ample data storage space and ignores the influence of weather factors on the flight trajectory. Kaneshige J et al. [4] based on the dynamics and kinematics models, built the flight trajectory calculation module by using the performance parameters of the aircraft and the total energy equations. While the aircraft is also affected by ground control during flight, which is ignored by the point-mass model, resulting in low prediction accuracy. Feng et al. [5] and Qiao et al. [6] used the Kalman filter algorithm to predict the three-dimensional trajectory of the aircraft. Wang et al. [7] proposed an improved Kalman filter algorithm for real-time estimation of the system noise in the prediction model to improve the prediction accuracy. Kalman filter is a linear optimal filtering algorithm, which is only suitable for linear Gaussian systems, and it is difficult to predict the trajectory of aircraft precisely. Sun et al. [8] improved the standard Extended Kalman Filter (EKF) algorithm, used the calculation of the innovation covariance to adjust the gain matrix of the Kalman filter, and predicted the target trajectory in the air based on the filtered data. However, the EKF unavoidably introduces linearization errors by Taylor expansion of the nonlinear system equation or observation equation and retaining the first-order approximation term.

In order to improve the effect of filtering nonlinear problems, this paper firstly performs Unscented transformation when processing the aircraft motion state equation. The Unscented transformation determines the sampling points near the estimated points, and uses the Gaussian density represented by these sample points to approximate the probability density function of the state, the accuracy after nonlinear transformation obtained in this way can reach the third-order accuracy. Then, on this basis, the Unscented Kalman Filter (UKF) algorithm is proposed. Since the UKF does not ignore high-order terms, it has higher calculation accuracy for statistics of nonlinear distribution. Finally, a continuous descent trajectory prediction simulation based on UKF is carried out, and compared with the results obtained by applying traditional Kalman filtering and extended Kalman filtering algorithms, it is verified that the method proposed can improve the prediction accuracy.

## 2 ADS-B Data Processing

Real-time flight trajectory information of the aircraft, including aircraft identification code, aircraft type, heading, position data (longitude, latitude and altitude), ground speed and time of passing through the corresponding waypoints, can be obtained through Automatic Dependent Surveillance-Broadcast (ADS-B) receiver.

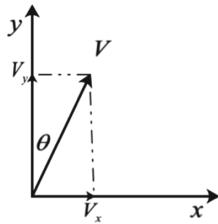
## 2.1 Coordinate Conversion of Trajectory Data

The flight position information acquired in real-time through ADS-B is expressed in latitude and longitude, and it is generally more convenient to use the spatial right-angle coordinate system when calculating and predicting the trajectory. Therefore, it is necessary to transform the real-time longitude and latitude ( $\varphi, \vartheta$ ) into the spatial right-angle coordinate ( $x, y, z$ ) with the airport reference point as the projection reference point through Mercator projection, and then predict the trajectory of aircraft. The calculation expression for converting from  $\varphi$  and  $\vartheta$  to  $x, y$  and  $z$  is as follows:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \cos(\varphi) \sin(\vartheta) \\ \sin(\vartheta) \\ \cos(\varphi) \cos(\vartheta) \end{bmatrix} \quad (1)$$

## 2.2 Aircraft Speed Conversion

The aircraft speed fed back through ADS-B is ground speed, and the heading is the angle from the north end of the meridian where the aircraft is located, clockwise to the direction pointed by the nose. As shown in Fig. 1, assuming the ground speed is  $V$ , the heading angle is  $\theta$ , the speed in direction  $x$  is  $V_x$ , and the speed in direction  $y$  is  $V_y$ .



**Fig. 1.** Schematic diagram of aircraft ground speed decomposition.

It can be seen from Fig. 1 that the calculation expressions of the  $x$ -axis and  $y$ -axis speed components are as follows:

$$V_x = V \sin\left(\frac{\theta \cdot \pi}{180}\right) \quad (2)$$

$$V_y = V \cos\left(\frac{\theta \cdot \pi}{180}\right) \quad (3)$$

### 3 Dynamic Trajectory Prediction Based on UKF

#### 3.1 Unscented Transformation

Unscented transformation is to sample the state value of the nonlinear system, so that the mean and covariance of the selected sigma points equal to the mean and covariance of the original distribution [9]. The selection of sampling points is based on the correlation column of the prior mean and the square root of the prior covariance matrix. Substituting these points into the non-linear function, the corresponding non-linear function value point set is obtained, and the transformed mean and covariance are obtained through these point sets. The mean and covariance accuracy obtained in this way after nonlinear transformation has at least the second-order accuracy. For the Gaussian distribution, the third-order accuracy can be achieved [10]. In this paper, according to the mean value  $\bar{X}$  and variance  $P_X$  of the model state variable  $X$ , the symmetric sampling strategy is used to select sigma points and uses some symmetric sigma points to approximate  $\bar{X}$  and  $P_X$ , which can get the model state sigma points  $\zeta_i, i = 1, 2 \dots 2n + 1$ , where  $n$  represents the dimension of the system state. The following sampling points are available [11]:

$$\begin{cases} \zeta_0(k) = \bar{X}(k) & i = 0 \\ \zeta_i(k) = \bar{X}(k) + \left( \sqrt{(n + \lambda) P_X} \right)_i & i = 1 \dots n \\ \zeta_i(k) = \bar{X}(k) - \left( \sqrt{(n + \lambda) P_X} \right)_{i-n} & i = n + 1 \dots 2n \end{cases} \quad (4)$$

where the parameter  $\lambda = \alpha^2(n + \delta) - n$ , which can adjust higher-order moments and reduce prediction errors;  $\alpha$  is the expansion factor that affects the distribution of the sigma points and  $\alpha \in [0.0001, 1]$ ;  $\delta$  is the scaling parameter, it determines the distance between the sampling point and the mean, the value of  $\delta$  needs to ensure the positive semidefinite of the matrix  $(n + \lambda) P_X$  and usually taken as 0;  $\left( \sqrt{(n + \lambda) P_X} \right)_i$  represents the  $i$ -th column of the square root of the matrix  $(n + \lambda) P_X$ . After selecting sigma points  $\zeta_i$ , the sigma points are brought into the nonlinear transfer function by Eq. (5), and yield the set of transformed sigma points  $Y_i$ :

$$Y_i = f(\zeta_i), i = 0, 1, 2 \dots 2n \quad (5)$$

From Eqs. (6)–(8), the corresponding mean  $\bar{Y}$  and covariance  $P_Y$  are given by the weighted average of the transformed points  $Y_i$ , that is, the update of time and measurement in the filtering process. The specific weight still needs to correspond to the weight of each sigma point sampled for the input variable  $X$ .

$$\bar{Y} = \sum_{i=0}^{2n} w_i^m Y_i \quad (6)$$

$$P_Y = \sum_{i=0}^{2n} w_i^c [Y_i - \bar{Y}] [Y_i - \bar{Y}]^T \quad (7)$$

where,

$$\begin{cases} w_0^m = \frac{\lambda}{n + \lambda} \\ w_0^c = \frac{\lambda}{n + \lambda} + (1 - \alpha^2 + \beta) \\ w_i^m = w_i^c = \frac{1}{2(n + \lambda)}, i = 1 \dots 2n \end{cases} \quad (8)$$

which satisfies  $\sum_{i=0}^{2n} w_i^m = 1$  and  $\sum_{i=0}^{2n} w_i^c = 1$ ;  $\beta$  contains the prior information of the distribution of the  $X$ .

### 3.2 Trajectory Prediction Model Based on UKF

During the continuous descent flight, the state vectors of the aircraft model include  $x$ -axis coordinates,  $y$ -axis coordinates,  $z$ -axis coordinates, and speed and acceleration in corresponding directions. Since the aircraft is flying with a constant descent gradient when implements the CDO procedures and the acceleration in the  $z$ -axis direction is 0, the equations of the state of motion of the aircraft from time step  $k$  to  $k + 1$  are:

$$\left\{ \begin{array}{l} x(k+1) = x(k) + V_x(k) \cdot T + 0.5 \cdot a_x(k) \cdot T^2 \\ y(k+1) = y(k) + V_y(k) \cdot T + 0.5 \cdot a_y(k) \cdot T^2 \\ z(k+1) = z(k) + V_z(k) \cdot T + 0.5 \cdot a_z(k) \cdot T^2 \\ V_x(k+1) = V_x(k) + a_x(k) \cdot T \\ V_y(k+1) = V_y(k) + a_y(k) \cdot T \\ V_z(k+1) = V_z(k) + a_z(k) \cdot T \\ a_x(k+1) = a_x(k) \\ a_y(k+1) = a_y(k) \\ a_z(k+1) = a_z(k) = 0 \end{array} \right. \quad (9)$$

where  $T$  is the interval of the forecast time;  $x(k)$ ,  $y(k)$  and  $z(k)$  are the spatial position data of the aircraft at time step  $k$ ;  $V_x(k)$ ,  $V_y(k)$  and  $V_z(k)$  are corresponding speeds;  $a_x(k)$ ,  $a_y(k)$  and  $a_z(k)$  are corresponding accelerations.

Suppose that the motion state of the aircraft is  $X(k)$  at time step  $k$ , then

$$X(k) = [x(k) \ y(k) \ z(k) \ V_x(k) \ V_y(k) \ V_z(k) \ a_x \ a_y \ a_z]^T \quad (10)$$

According to Eq. (9), the recursive equation can be obtained. The dynamical system described either in the additive form

$$X(k+1) = A(k+1 | k)X(k) + F(k+1 | k)w(k) \quad (11)$$

$$A = \begin{bmatrix} 1 & 0 & 0 & T & 0 & 0 & 0.5T^2 & 0 & 0 \\ 0 & 1 & 0 & 0 & T & 0 & 0 & 0.5T^2 & 0 \\ 0 & 0 & 1 & 0 & 0 & T & 0 & 0 & 0.5T^2 \\ \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (12)$$

$$F = \begin{bmatrix} 0.5T^2 & 0 & 0 & T & 0 & 0 & 1 & 0 & 0 \\ 0 & 0.5T^2 & 0 & 0 & T & 0 & 0 & 1 & 0 \\ 0 & 0 & 0.5T^2 & 0 & 0 & T & 0 & 0 & 1 \end{bmatrix}^T \quad (13)$$

where  $X(k)$  is a 9-dimensional state vector, describing the value of each state of the motion system at the time step  $k$ ;  $X(k+1)$  is the state estimate at the time step  $k+1$ ;  $A$  is the state transition matrix;  $F$  is an interference input matrix;  $w$  represents the system noise of acceleration in different directions, and it is 3-dimensional Gaussian white noise.

The measurement equation can be expressed as

$$Z(k+1) = H(k+1 | k)X(k) + v(k) \quad (14)$$

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}_{3 \times 9} \quad (15)$$

where  $Z$  represents the observation vector;  $H$  is the observation matrix;  $v$  represents the observation noise generated during the flight, also set as the Gaussian white noise. The noise terms  $w$  and  $v$  are assumed to be uncorrelated, with zero means, and covariance matrices  $Q$  and  $R$ , respectively.

The specific calculation steps are as follows:

(1) Filter initialization

$$\hat{X}(0) = E(X(0)) \quad (16)$$

$$P_0(0) = E \left[ (X(0) - \hat{X}(0)) (X(0) - \hat{X}(0))^T \right] \quad (17)$$

(2) Calculate sigma points according to Eq. (4)

## (3) Time update equations

$$Y_i(k | k - 1) = A\zeta_i, i = 0, 1 \dots 2n \quad (18)$$

$$\hat{X}(k | k - 1) = \sum_{i=0}^{2n} w_i^m Y_i(k | k - 1) \quad (19)$$

$$P_X(k | k - 1) = \sum_{i=0}^{2n} w_i^c \left( Y_i(k - 1) - \hat{X}(k | k - 1) \right) \\ \left( Y_i(k - 1) - \hat{X}(k | k - 1) \right)^T + Q \quad (20)$$

$$V_i(k | k - 1) = HY_i(k | k - 1), i = 0, 1 \dots 2n \quad (21)$$

$$\hat{Z}(k) = \sum_{i=0}^{2n} w_i^m V_i(k | k - 1) \quad (22)$$

## (4) Measurement update equations

$$P_{ZZ}(k | k - 1) = \sum_{i=0}^{2n} w_i^c \left( V_i(k | k - 1) - \hat{Z}(k | k - 1) \right) \\ \left( V_i(k | k - 1) - \hat{Z}(k | k - 1) \right)^T + R \quad (23)$$

$$P_{XZ}(k | k - 1) = \sum_{i=0}^{2n} w_i^c \left( Y_i(k | k - 1) - \hat{Y}(k | k - 1) \right) \\ \left( Y_i(k | k - 1) - \hat{Y}(k | k - 1) \right)^T \quad (24)$$

$$L(k) = P_{XZ}(k | k - 1) P_{ZZ}^{-1}(k | k - 1) \quad (25)$$

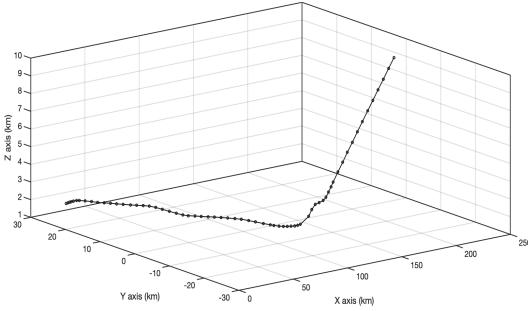
$$\hat{X}(k) = \hat{X}(k | k - 1) + L(k) \left( Z(k) - \hat{Z}(k | k - 1) \right) \quad (26)$$

$$P_X(k | k) = P_X(k | k - 1) - L(k) P_{ZZ}(k | k - 1) L^T(k) \quad (27)$$

## 4 Simulations and Analysis

### 4.1 Simulation Data

In this paper, taking the CES2492 flight as an example, the real-time data of flight CES2492 from Wenzhou to Guangzhou on March 25, 2020. The reference point of Baiyun international airport is selected as the origin of the spatial right-angle coordinate system, and the real flight trajectory data is converted into spatial coordinates, as shown in Fig. 2. At the same time, to verify the effectiveness of the trajectory prediction of the CDO proposed in this paper,



**Fig. 2.** The real flight trajectory of the CES2492.

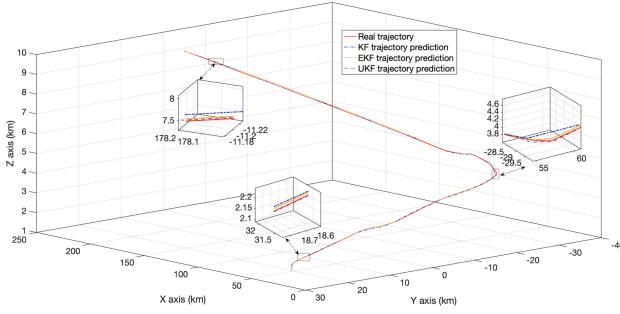
the Kalman filter and extended Kalman filter algorithm are used for trajectory prediction, designed as a comparative experiment.

According to the data obtained by ADS-B, the observation interval is  $T = 24$  s, the number of observations is  $N = 65$ ,  $k = 1 \cdots N$ , the initial state of the aircraft is  $X(0) = [235333 \ -1082 \ 9182 \ -194.4 \ -37.8 \ -4.8 \ 0.5 \ 0.1 \ 0.0]^T$ , the initial value of the filter is  $X(0|0) = X(0)$ , the flight time is 1560 s,  $Q = 0.03^2 I_3$ , and  $R = 0.01^2 I_3$ . In the unscented transformation,  $\alpha$  is the expansion factor, which is generally a small value,  $\delta$  is scale factor, usually set to 0,  $\beta$  is 2 under the Gaussian distribution, so the coefficients are set to  $\alpha = 0.0001$ ,  $\delta = 0$  and  $\beta = 2$ .

## 4.2 Simulation Results and Analysis

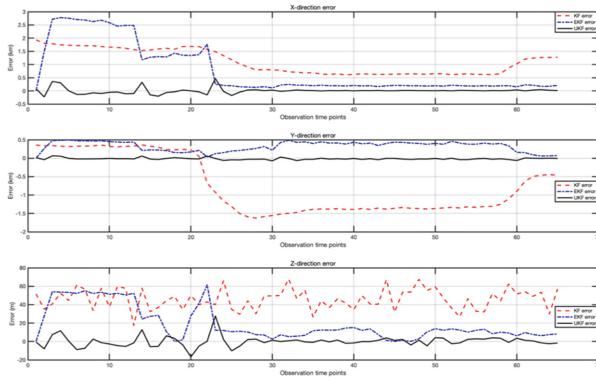
It can be seen from Fig. 3 that the estimated trajectory of the UKF for the aircraft position changes with the change of the real trajectory, especially the prediction errors are tiny when the aircraft turns, the two trajectories almost coincide; The EKF has a poor effect on trajectory prediction, but there is a tendency for the prediction errors to converge on the whole; The trajectory prediction performance of the KF is extremely poor, especially when the aircraft turns, due to changes in speed, heading, and wind speed, the filter estimation errors increase rapidly. From the simulation results, the UKF can accurately estimate the real-time trajectory of the aircraft, and it has better adaptability and stability to the aircraft whose motion parameters change rapidly.

During the actual flight, due to the continuous changes in speed and heading, it is relatively normal for the prediction errors to fluctuate. As shown in Fig. 4: In the X direction, At the beginning of the descent of the aircraft, the UKF has the smallest errors, and the fluctuation amplitude is less than 0.5 km, the fluctuation range of errors of KF is between 1.5 km and 2.0 km, the prediction errors of EKF fluctuates more than 2.5 km. As the prediction time increases, the prediction errors of UKF gradually converge to 0, the prediction errors of EKF do not converge, and the errors of the KF tend to diverge; In the Y direction, the errors of UKF is minimal and is relatively stable. While, the errors of the EKF



**Fig. 3.** The real trajectory and predicted trajectories based on KF, EKF and UKF.

are relatively large, and the errors of KF is far from satisfying the requirements of trajectory prediction. In the Z direction, since the aircraft is in a state of constant speed descent, the three prediction algorithms have relatively small errors. Among them, the UKF has relatively tiny errors in trajectory prediction; While, the prediction errors of KF and EKF are relatively large and continuously oscillate, with a tendency of divergence.



**Fig. 4.** Comparison of prediction errors of KF, EKF, UKF in X, Y, and Z directions.

The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) of KF, EKF and UKF algorithms in the direction of X, Y and Z are shown in Table 1. It can be seen from Table 1 that in the X, Y and Z directions, the prediction errors of KF and EKF are much larger than that of UKF. UKF has the highest accuracy for aircraft trajectory prediction, compared with KF and EKF algorithms, the MAE in the X direction has decreased by 94.39% and 92.14%, and RMSE has decreased by 90.00% and 79.67%, respectively; In the Y direction, MAE decreased by 97.43% and 92.54%, and RMSE decreased by 96.12% and 85.41%, respectively; In the Z direction, MAE decreased by 92.69%

and 81.11%, and RMSE decreased by 62.97% and 77.51%, respectively. The calculation of MAE and RMSE for prediction errors in different directions more strongly proves the superiority of UKF over EKF and KF.

As a result, KF cannot accurately predict the trajectory of nonlinear systems such as aircraft; For EKF, calculating Jacobian matrices for the nonlinear equations of state can be a challenging and error-prone process, so it is also difficult to predict the dynamic trajectory of the aircraft based on the initial state and observation information. The UKF method can not only obtain precise trajectory prediction accuracy but also effectively overcome the problem of filtering divergence, with the characteristics of unbiased, stable and optimal.

**Table 1.** The MAE and RMSE of the predicted errors in X, Y and Z directions.

Method	X direction		Y direction		Z direction	
	MAE (m)	RMSE (m)	MAE (m)	RMSE (m)	MAE (m)	RMSE (m)
KF	1099.7	461.3	942.0	515.8	48.4	11.0
EKF	785.2	938.2	324.5	137.1	18.7	18.1
UKF	61.7	93.8	24.2	20.0	3.5	4.1

## 5 Conclusions

Given the uncertainty of continuous descent trajectory, this paper establishes a trajectory prediction model based on UKF and proposes a dynamic trajectory prediction method. Firstly, the ADS-B data is processed by Mercator projection. Then, the UKF method is used to improve the estimation of aircraft position by combining with the state equation of the unscented transformation processing system. Finally, compared with the results obtained by applying the traditional Kalman filter and extended Kalman filter algorithm, the simulation results verify that the UKF can improve the filtering effect. Therefore, the method proposed in this paper has an excellent adaptability to the aircraft with rapidly changing motion parameters, which can improve the predictability and stability of the continuous descent trajectory and increase the capacity of the terminal area.

**Acknowledgement.** This work was support by National Natural Science Foundation of China (61973212, 61673270) and China Scholarship Council.

## References

1. International Civil Aviation Organization (ICAO). Continuous Descent Operations (CDO) Manual (DOC 9931) International Civil Aviation Organization, Montreal (2010)

2. Sipe, A., Moore, J.: Air traffic functions in the NextGen and SESAR airspace. In: 2009 IEEE/AIAA 28th Digital Avionics Systems Conference, pp. 2.A.6-1–2.A.6-7 (2009)
3. Hong, S., Lee, K.: Trajectory prediction for vectored area navigation arrivals. *J. Aerosp. Inf. Syst.* **12**(7), 490–502 (2015). <https://doi.org/10.2514/1.I010245>
4. Kaneshige J, Benavides J, Sharma S, et al.: Implementation of a trajectory prediction function for trajectory based operations. In: Proceedings of the AIAA Atmospheric Flight Mechanics Conference, AIAA 2019 (2014). <https://doi.org/10.2514/6.2014-2198>
5. Feng, Z.X.: Analysis of track prediction based on Kalman filtering. *Sci. Tech. Inf. Gansu* **048**(003), 33–36 (2019)
6. Qiao, S.J., Han, N., Zhu, X.W.: A dynamic trajectory prediction algorithm based on Kalman filter. *Acta Electronica Sinica* **046**(002), 418–423 (2018)
7. Wang, T.B., Huang, B.J.: 4D flight trajectory prediction model based on improved Kalman filter. *J. Comput. Appl.* **34**(6), 1812–1815 (2014)
8. Sun, J., Hu, R., Yang, H.: Air target tracking based on improved extended Kalman filtering algorithm. *Shipboard Electron. Countermeasure* **36**(06), 40–43 (2013)
9. Zhang, Y.G., Huang, Y.L., Wu, Z.M.: A high order unscented Kalman filtering method. *Acta Automatica Sinica* **40**(5), 838–848 (2014)
10. Julier, S.J., Uhlmann, J.K.: Unscented filtering and nonlinear estimation. *Proc. IEEE* **92**(3), 401–422 (2004). <https://doi.org/10.1109/JPROC.2003.823141>
11. Liu, Y., Li, X.: Study of target tracking based on improved unscented transform Kalman filtering. *Computer Engineering and Design*. **31**(14), 3331–3335 (2010)



# Research on Weighted Multiple Model Adaptive Control Based on U-Model

Jiayi Li and Weicun Zhang<sup>(✉)</sup>

School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China  
lijiaiyiqaq@163.com, weicunzhang@263.net

**Abstract.** This article proposes a method combining the U-model and weighted multiple model adaptive control to solve control problems of nonlinear uncertain plants. The key to apply the mature linear system theory to the nonlinear system is to establish a system model that retains the form of nonlinear characteristics, which stimulates the creation of the U-model. The weighted multiple model adaptive control overrides the controlled object by constructing the model set of multiple fixed models, so as to solve the problem of uncertain parameters and parameters jump. Finally, we use the new control strategy to simulate a nonlinear system through the MATLAB software, and the results verify the convergence and stability of the multiple U-model control strategy.

**Keywords:** Nonlinear system · Uncertain parameters · U-model · Weighted multiple model adaptive control · MATLAB

## 1 Introduction

At present, linear system control problem has been developed to a relatively mature stage, pole placement and other linear system control methods have been proved to apply to the linear system control problems successfully. However, in the actual production, most control systems have strong nonlinear characteristics and uncertain parameters, which brings greater challenges and higher requirements for the design methods of control systems.

The piecewise linearization method [1] is often adopted in the control of nonlinear system. However, for strongly nonlinear systems, this method has problems such as excessive control deviation and complicated calculation. In 2002, the U-model has been raised to represent nonlinear discrete dynamic system [2]. This method has few parameters, simple calculation, and will not lose the nonlinear characteristics of the system, which simplifies the controller design of nonlinear systems to some extent. Reference [2] has studied the design of U-model pole assignment controller. Reference [3] has proposed internal model control under U-model.

The traditional adaptive control is suitable for control object with constant or slowly changing parameters. When there is a large disturbance outside the system or a large change in internal characteristics, the adaptive control is restricted

by the convergence speed of the identification algorithm. To solve this situation, Magill put forward the basic idea of multiple models [4], which mainly uses the estimated values of a group of state estimators for weighted combination to achieve optimal estimation and optimal control. In 1971, Lainiotis [5] formally proposed the multiple model adaptive control (MMAC). The idea of weighted multiple model adaptive control (WMMAC) is to set up many sub models and corresponding sub controllers offline, calculate the weight of each sub model online through a certain algorithm, and control the system effectively by the controller which recombines the sub controllers according to the weight. It is an important method to realize the robust adaptive control. If the parameters of the controlled system jump, the adaptive system can recalculate the weight, so that the whole control process can meet the requirements of transient and static performance indicators and rapidity. The idea of weighted multiple model has a wide range of practical applications, and has been studied in some fields, such as fault diagnosis, medical, target tracking, aviation, process control, etc. With the development of predictive control, robust adaptive control, fuzzy PID control and other controller form, the control capability could be improved through the combination of WMMAC and those control strategies above.

This study combines the U-model with the weighted multiple model control strategy. The combination method could facilitate the controller design process and solve the parameter uncertainty problem of the nonlinear system.

## 2 Pole Placement Controller Design of U-Model

### 2.1 Control-Oriented Nonlinear Plant Models

Consider nonlinear system as:

$$y(t) = f[y(t-1), \dots, y(t-n), u(t-1), \dots, u(t-n), e(t), \dots, e(t-n)] \quad (1)$$

In the equation above, the plant output is expressed in  $y(t)$  and the input is expressed in  $u(t)$ ,  $e(t)$  represents the system error.

U-model is a nonlinear modeling method oriented to control. The U-model is defined as follows

$$y(t) = U(t) \quad (2)$$

where

$$U(t) = \sum_{j=0}^N \alpha_j(t) u^j(t-1) + e(t) \quad (3)$$

In the equation above,  $N$  is the order of the input, the parameter  $\alpha_j(t)$  is the expression made up of the past input signals and output signals and errors.

The definition of U-model indicates that the nonlinear representation does not use the linearization method, so it retains the nonlinear characteristics of the system. The main difficulty of the nonlinear system is the complexity of the model structure. In the U-model, each term is a power series of  $u(t-1)$ , and each

coefficient is a complex time-varying parameter. Some unknown or uncertain error terms can be attributed to the time-varying parameters, which can simplify the model structure to a certain extent. The main advantage of U-model is that it can transform nonlinear system into similar linear system, and connect nonlinear system and linear control theory by calculating analytical solution.  $U(t)$  in the Eq. (3) is a nonlinear equation, which is composed of different orders of the real input. The spurious input  $U(t)$  is designed by some way, and then the true input could be acquired through calculation. There are many ways to calculate the analytical solution.

## 2.2 Pole Placement Controller Design

For linear stationary systems, the stability of the system and the quality index of various performances are largely determined by the pole position of the control system. Therefore, it can effectively perfect the performance and quality of the system to try to make the poles located in a reasonable set of poles with the desired performance and quality indicators on the s-plane. This is a common method in linear control problems called pole placement. However, it is not easy to determine the zeros and poles position of the nonlinear system, so the method of pole placement cannot be applied directly. Generally, the nonlinear system will be approximately linearized, but it will be limited by its local performance. For the nonlinear object expressed by U-model structure, the control quantity is expressed by multinomial, and the variable is controllable. Only solving the equation can get the output of the controller and carry out pole placement, thus completing the pole placement design of the nonlinear system.

The controller could be represented as:

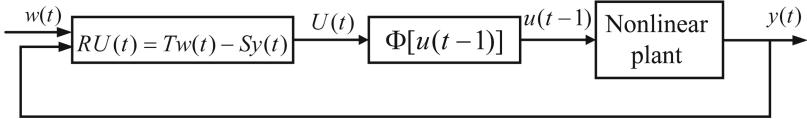
$$RU(t) = Tw(t) - Sy(t) \quad (4)$$

In Eq. (4),  $R$ ,  $T$  and  $S$  are polynomials of the forward shift operator,  $w(t)$  is the reference input of the system. Figure 1 is the block diagram of the controlled plant. The relationship between the plant output and the reference input can be shown in Eq. (5).

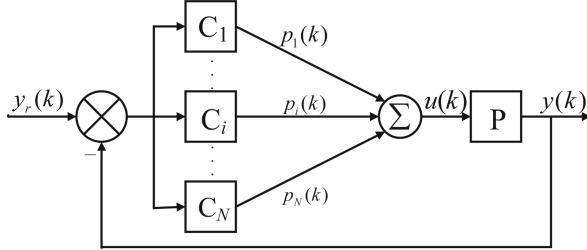
$$y(t) = \frac{T}{R + S}w(t) = \frac{T}{A_c}w(t) \quad (5)$$

## 3 Weighted Multiple Model Adaptive Control Algorithm

The structure of weighted multiple model adaptive control system shown in Fig. 2 is composed of three sections: multiple model set, controller set and weighted algorithm.  $y_r(k)$  is the reference input,  $p_i(k)$  is the corresponding weight of the sub controller  $C_i$ . The calculation method of the weight is given later. The control rate  $u_i(k)$  generated by each sub controller is multiplied by the corresponding weight and then added to obtain a new control rate  $u(k)$ . It acts on the controlled system and generates the system output  $y(k)$ .



**Fig. 1.** The block scheme of closed-loop nonlinear control system



**Fig. 2.** Weighted multiple model adaptive control system block diagram

The way to choose the nearest sub model is according to the error which is the difference value between the actual system output and the sub model output. By increasing the weight of the sub model with smaller error and reducing the weight of the sub model with larger error, the weight of the sub model nearest to the real system approaches to 1 and the weights of other sub models approach to 0. Using the identified sub model controller to control the system can achieve better control effect [6, 7].

Based on the above ideas, the following weighted algorithm is designed:

$$l_i(0) = p_i(0) = \frac{1}{N} \quad (6)$$

$$l'_i(k) = 0.001 + \frac{1}{k} \sum_{p=1}^k e_i^2(p) \quad (7)$$

$$l'_{min}(k) = \min_i l'_i(k) \quad (8)$$

$$l_i(k) = \frac{l'_{min}(k)}{l'_i(k)} l_i(k-1) \quad (9)$$

$$p_i(k) = \frac{l_i(k)}{\sum_{i=1}^N l_i(k)} \quad (10)$$

## 4 Simulation Results

To prove the new control strategy is feasible for uncertain nonlinear systems, lots of nonlinear systems are simulated in this study, and one of them is given below.

The closed-loop characteristic polynomial is expressed as:

$$A_c = q^2 - 1.3205q + 0.4966 \quad (11)$$

Let

$$T = A_c(1) = 1 - 1.3205 + 0.4966 = 0.1761 \quad (12)$$

Let  $R$  and  $S$  are expressed as follows:

$$\begin{aligned} R &= q^2 + r_1q + r_2 \\ S &= s_0q + s_1 \end{aligned} \quad (13)$$

From Eq. (5)

$$R + S = A_c \quad (14)$$

Substituting Eq. (11) and Eq. (13) into Diophantine equation of (14) to obtain

$$\begin{aligned} s_1 + r_2 &= 0.4966 \\ s_0 + r_1 &= -1.3205 \end{aligned} \quad (15)$$

In order to make  $U(t)$  have convergence, let

$$\begin{aligned} r_1 &= -0.9 \\ r_2 &= 0.009 \end{aligned}$$

According to the expression (15), we can get

$$\begin{aligned} s_0 &= -0.4205 \\ s_1 &= 0.4876 \end{aligned}$$

According to the values obtained above, Eq. (4) can be further expressed as:

$$\begin{aligned} U(t+1) &= 0.9U(t) - 0.009U(t-1) + 0.1761w(t-1) \\ &\quad + 0.4205y(t) - 0.4876y(t-1) \end{aligned} \quad (16)$$

By solving Eq. (16), the real control output can be obtained.

The simulation system:

$$\begin{aligned} y(t) &= 0.5y(t-1) + x(t-1) + 0.1x(t-2) \\ x(t) &= 1 + u(t) - u^2(t) + 0.2u^3(t) \end{aligned}$$

The equivalent U-model can be expressed as:

$$y(t) = \alpha_0(t) + \alpha_1(t)u(t-1) + \alpha_2(t)u^2(t-1) + \alpha_3(t)u^3(t-1)$$

where  $\alpha_0(t) = 0.5y(t-1) + 1 + 0.1x(t-2)$ ,  $\alpha_1(t) = 1$ ,  $\alpha_2(t) = -1$ ,  $\alpha_3(t) = 0.2$ . Taking the parameter  $h = 0.5$  in  $\alpha_0(t)$  as the uncertain parameter of nonlinear system, four corresponding models are selected for this parameter:  $h_1 = 0.51$ ,  $h_2 = 0.52$ ,  $h_3 = 0.53$ ,  $h_4 = 0.3$ . For the four models, design the controllers

according to the pole placement method, and add noise  $\text{sqrt}(0.01) \cdot \text{randn}(1)$  to the system. The local controllers of four fixed models are

Model 1: in this sub model, if the uncertain parameter value is 0.51, then

$$0.2u^3(t) - u^2(t) + u(t) + 0.51y(t) + 1 + 0.1x(t-1) = U(t+1)$$

Model 2: in this sub model, if the uncertain parameter value is 0.52, then

$$0.2u^3(t) - u^2(t) + u(t) + 0.52y(t) + 1 + 0.1x(t-1) = U(t+1)$$

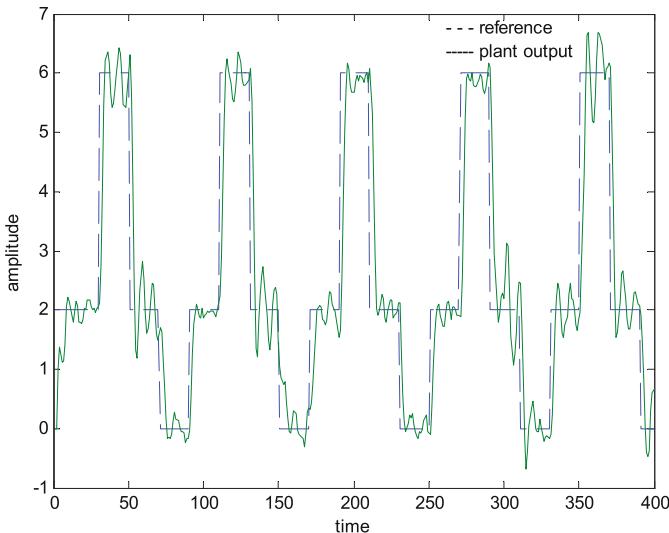
Model 3: in this sub model, if the uncertain parameter value is 0.53, then

$$0.2u^3(t) - u^2(t) + u(t) + 0.53y(t) + 1 + 0.1x(t-1) = U(t+1)$$

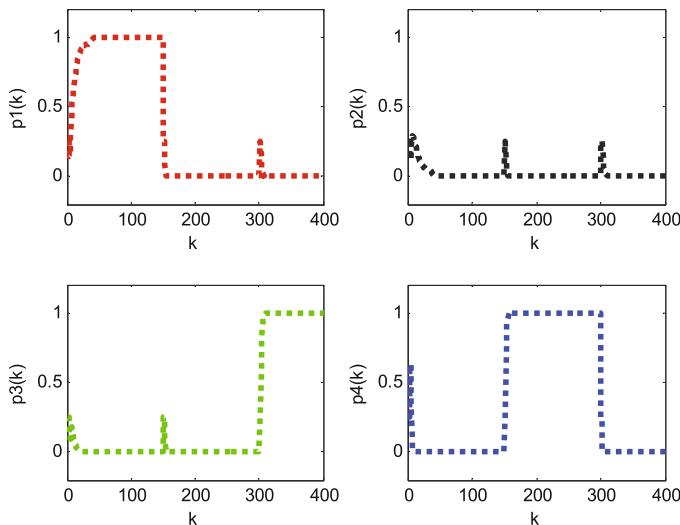
Model 4: in this sub model, if the uncertain parameter value is 0.3, then

$$0.2u^3(t) - u^2(t) + u(t) + 0.3y(t) + 1 + 0.1x(t-1) = U(t+1)$$

Taking the rectangular wave signal as the reference input. To prove the effectiveness of the new control strategy when the controlled system has uncertain parameter and parameter jump, the following experiments are carried out. When  $t = 150$ ,  $h$  jumps to 0.28 and holds. When  $t = 300$ ,  $h$  jumps to 0.55 and holds. The experimental outcomes are shown in Fig. 3 and 4. The plant output can effectively track the reference input in Fig. 3. In Fig. 4, when  $h = 0.5$ , model 1 is closest to the controlled system, so model 1 is selected; when  $h$  jumps to 0.28, model 4 is closest to the controlled system, so “switch” to model 4; finally to model 3. Therefore, the weighted algorithm can quickly and accurately select the corresponding model and effectively control the object when the parameter changes.



**Fig. 3.** Output and input signals of parameter jumps system



**Fig. 4.** Local controller weight of parameter jumps system

## 5 Conclusion

In conclusion, this article puts forward a control algorithm combining U-model and weighted multiple model. The design of the controller adopts the pole assignment method in linear system, and the weighted algorithm is designed by the idea of model output error. The simulation results display clearly that this control method owns a fairly good performance on the nonlinear controlled system with uncertain parameters or parameters jump. In a word, weighted multiple U-model control strategy provides a convenient and efficient method for nonlinear system control. Next, the application of this method in practice can be further studied.

## References

1. Leenaerts, D.M.W., Bokhoven, W.M.G.V.: Piecewise Linear Modeling and Analysis. Kluwer Academic Publishers, Boston (1998)
2. Zhu, Q.M., Guo, L.Z.: A pole placement controller for nonlinear dynamic plant. Proc. Inst. Mech. Eng. Part I J. Syst. Control Eng. **216**, 467–476 (2002). <https://doi.org/10.1177/095965180221600603>
3. Shafiq, M., Butt, N.R.: Real-time adaptive tracking of DC motor speed using U-model based IMC. Automat. Control Comput. Sci. **41**, 31–38 (2007). <https://doi.org/10.3103/s0146411607010051>
4. Magill, D.T.: Optimal adaptive estimation of sampled stochastic processes. IEEE Trans. Automat. Control **10**, 434–439 (1965). <https://doi.org/10.1109/tac.1965.1098191>

5. Lainiotis, D.: Optimal adaptive estimation: structure and parameter adaption. *IEEE Trans. Autom. Control* **16**, 160–170 (1971). <https://doi.org/10.1109/tac.1971.1099684>
6. Zhang, W.C.: Stable weighted multiple model adaptive control: discrete-time stochastic plant. *Int. J. Adapt. Control Signal Process.* **27**, 562–581 (2003). <https://doi.org/10.1002/acs.2328>
7. Zhang, Z.H., Wang, S.Q., Wang, R.D.: Nonlinear multiple model control and simulation. *J. Syst. Simul.* **15**, 919–921 (2003). CNKI:SUN:XTFZ.0.2003-07-003



# Human Action Recognition Method Based on Video-Level Features and Attention Mechanism

Qiang Cai<sup>1,2,3</sup>, Jin Yan<sup>1,2,3(✉)</sup>, Haisheng Li<sup>1,2,3</sup>, and Yibiao Deng<sup>1,2,3</sup>

<sup>1</sup> Beijing Technology and Business University, Beijing 100048, China  
yj9621394910163.com

<sup>2</sup> Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing, China

<sup>3</sup> National Engineering Laboratory for Agri-Product Quality Traceability,  
Beijing, China

**Abstract.** In order to capture the spatiotemporal information in the video and improve the long-term modeling capability of the network, two-stream network usually adopts the method of sparse sampling. However, there are two problems in the process of feature extraction on the sample. One is the unreasonableness of the snippet-level features corresponding to the video labels, and the other is that the salient features are not highlighted. In view of the above two points, we propose an action recognition method based on video-level features and attention mechanism (VFAM), which combines snippet-level features to generate video-level features, and adds attention mechanism to give effective features with greater weight, and good experimental results have been achieved on the dataset HMDB51, reflecting the superiority and robustness of our method.

**Keywords:** Two-stream network · Snippet-level features · Video-level features · Attention mechanism · Action recognition

## 1 Introduction

In this era of data explosion, video as one of the main carriers carries a lot of information, and human action recognition in video is a key technology for video understanding, has a wide range of application scenarios and research significance in human-computer interaction, intelligent video monitoring, intelligent home, consumer action analysis, etc.

With the great progress of convolutional neural network in image classification, the development of deep learning methods in video processing has been promoted. Based on deep learning method can learn iteratively, use a large amount of video data as the driver, and extract the spatiotemporal information in the video. The method based on 3D convolutional network [2,4,12–14,17,18,20] is mainly to achieve the human action recognition by acquiring spatiotemporal information at the same time, but there are great requirements for computer

memory and calculation speed. The human action recognition method based on two-stream network has been greatly developed as a mainstream method. The two-stream network architecture proposed by Simonyan et al. [15] for action recognition is divided into two branches of spatial stream network and temporal stream network. Wang et al. [22] propose Temporal Segment Network to achieve long-term modeling. Feichtenhofer et al. [6] pointed out the lack of pixel-level correspondence between spatial and temporal features in two-stream networks, and inspired by the excellent performance of the residual network, C. Feichtenhofer [5] proposed STResNet on the basis of the two-stream network. In order to make the existing convolutional neural network available for video, Fernando et al. [1, 7, 8] proposed the network structure of Dynamic Image Networks. Due to the large amount of calculation of optical flow images, the method of extracting video temporal information from static images [3, 9, 10, 24] is also a major improvement direction.

The general idea of the two-stream network is to input various modalities of the video, extract the features of the video through the convolutional neural network, and finally classify them according to the features. It can be seen that feature extraction is crucial in human behavior recognition. The model in this paper integrates the features in the Temporal Segmentation Network (TSN) [22] to obtain video-level features with better expressive power, and adds an attention mechanism to improve the quality of the features, thereby improving the accuracy of human action recognition. The main contributions of this article are as follows:

- (1) Extract video-level features;
- (2) Based on video-level features, we propose a two-stream network structure based on attention mechanism;
- (3) Our work has achieved good experimental verification.

## 2 Related Works

The main idea of the human action recognition method based on the two-stream network is to input video frames and optical flow images into the spatial stream network and the temporal stream network to extract the spatial information and temporal information contained in the video, and then combine the two-stream network to obtain video information to get the final human action recognition result.

In order to achieve long-term modeling, Temporal Segment Network (TSN) [22] propose a sparse sampling strategy. First, the video frames and optical flow images in snippets are input to obtain the feature map of this snippet and classification score. Then fuse the video snippets to get the single stream classification score; Finally merge the two-stream classification score with a certain weight to complete the recognition of human action.

As shown in Fig. 1, this is the result of sparse sampling of a person's smoking video, and it is difficult to recognize the human action of smoking from these few frames of snippets. This leads to two problems:



**Fig. 1.** Examples of smoking video frames

- (1) It is difficult to match the human action label of smoking only from the video snippet as shown in Fig. 1. Is it reasonable to train the network with the video-level label corresponding to the sampling video snippet?
- (2) If a segment of the video occurs as shown in Fig. 1, and then combine the N segments of the video at the same equal ratio, will it reduce the accuracy of human action recognition?

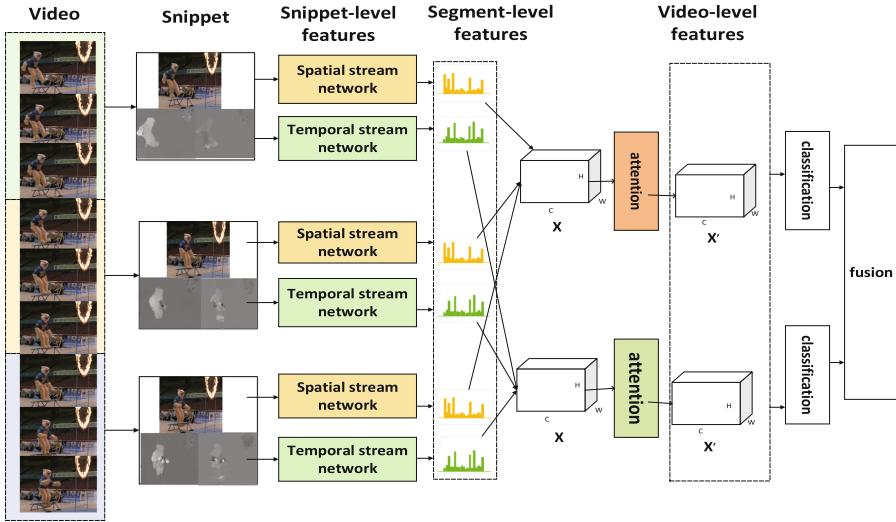
Therefore, we propose a human action recognition method based on video-level features and attention mechanism.

### 3 Human Action Recognition with VFAM

In this part, we will talk about the specific implementation details of the proposed method. First, we introduce the network framework of two-stream network (VFAM) based on video-level features and attention mechanism; then we will talk about the extraction of video-level features; and finally, we will describe the details of the attention mechanism we use.

#### 3.1 Network Architectures

Our method is based on Temporal Segment Network (TSN) implemented by BN-Inception. The overall network structure is shown in Fig. 2. The main improvement is to extract the video snippet-level features through the two-stream network, then fuse them into segment-level features, then fuse them into video-level features, and then use the video labels and video-level features for classification loss, so as to achieve the optimization of the network; Another improvement lies in the introduction of the attention module, which assigns greater weight to features that are effective in human action recognition, and thus improve the accuracy of human action recognition method. This network includes temporal stream network and spatial stream network. The whole idea of two-stream network is the same, but it is different when input. Temporal stream network uses optical flow image as input to extract temporal information in video; spatial stream network uses video frames as input to extract spatial information in video. So take the spatial stream network as an example to describe the overall steps of the algorithm:



**Fig. 2.** Two-stream network structure based on video-level features and attention mechanism

In the first step, the video is sampled sparsely, that is, the video is divided into three segments on average, and then random sampling is carried out in each segment to get some video snippets;

In the second step, the BN-Inception network is used as the feature extractor to extract the snippet-level features;

In the third step, the snippet-level features of each segment in step 2 are fused to get the segment-level features and achieve the partial consistency;

In the fourth step, the attention module is added, so that the more effective feature of human action recognition in step 3 has more weight, and the corresponding weight value of invalid or small effect segments is smaller, and the video-level spatial stream feature is obtained through the attention module;

In the fifth step, the video-level spatial stream feature in step 4 is input into the classifier to get the classification score.

According to the above steps, temporal stream network can also extract video-level features and their classification scores; finally, spatial stream network and temporal stream network are fused with a weight of 1:1.5 to achieve human action recognition.

### 3.2 Video-Level Feature Extraction

In view of the irrationality of the video tags corresponding to the video snippets obtained by sparse sampling  $\{C_1, C_2, \dots, C_k\}$ , we propose video-level features. For a video  $V$ ,  $K$  video snippets can be obtained after random sampling. The based on video-level features and attention mechanism (VFAM) is modeled according

to formula one:

$$VFAM(C_1, C_2, \dots, C_k) = H(V_F, W) \quad (1)$$

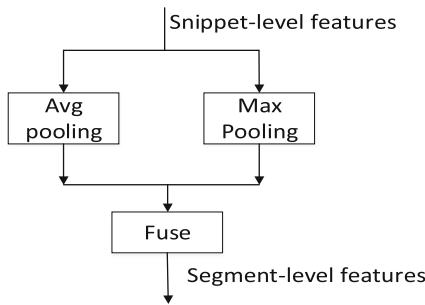
Among them, softmax function  $H$  is used to classify human action; video-level features  $V_F$  and  $W$  are the parameter groups in softmax function.

Video-level features are obtained by aggregating segment-level features and inputting them into the attention module:

$$V_F = A_M(g(F_1, F_2, \dots, F_k)) \quad (2)$$

Among them,  $g$  is the aggregation function of segment-level features and the operation of attention module is  $A_M$ .

In order to achieve partial consistency, we use the combination of average pooling and maximum pooling to fuse the segment-level features. That is the results of average pooling and maximum pooling, to achieve the retention of each video segment information while emphasizing more significant features. The schematic diagram is shown in Fig. 3.



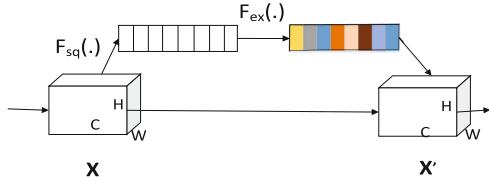
**Fig. 3.** Schematic diagram of feature fusion

Finally, video-level features and video labels are used as loss function parameters to optimize the network through back propagation.

### 3.3 Attention Mechanism

Only by adding the attention mechanism module after merging segment-level features can obtain more effective video-level features. By learning the each channel's weight value of the feature after segment-level feature fusion, the weight value of the part that is effective for human action recognition is larger, and the weight value corresponding to the part that is invalid or less effective is smaller. Multiply the obtained weight value with the original feature to obtain the video-level feature.

In this paper, channel attention [11] is used in our method, and its schematic diagram is shown in Fig. 4. The use of attention mechanism is divided into

**Fig. 4.** Channel attention diagram

two steps. The first step is squeeze operation, which uses global average pooling to generate channel statistics  $\{Z_1, Z_2, \dots, Z_c\}$ , and compresses the information after segment-level feature fusion into a channel descriptor. Formally, weight statistics are generated by reducing feature  $X$  to its spatial dimension  $H \times W'$ , so that element  $c$  of  $Z$  is calculated by the following formula:

$$Z_c = F_{sq}(X_c) = \frac{1}{H \times W'} \sum_{i=1}^H \sum_{j=1}^{W'} x_c(i, j) \quad (3)$$

In the formula (3), it is a squeeze operation function  $F_{sq}$ , which represents the weight statistics generated after the global average pooling of features on channel  $C$ .

The second step, in order to capture the information after squeeze operation, followed by an activation operation, can fully capture the dependency between each channel of the feature, so that the channel with a greater role in human action recognition has a greater weight, so sigmoid activation function is used as a simple gate mechanism, and the formula is as follows:

$$X' = F_{ex}(z, W'') = \sigma(g(z, w'')) = \sigma(W_2 \delta(W_1 z)) \quad (4)$$

In Eq. (4),  $F_{ex}$  is the activation operation function,  $\sigma$  is sigmoid activation function, and  $\delta$  is Relu activation function.

## 4 Experiments

### 4.1 Dataset and Experimental Environment

In this work, the HMDB database released in 2012 contains 6849 clips into 51 action categories, each containing a minimum of 101 clips. HMDB collected from various sources, mostly from movies, and a small proportion from public databases such as the Prelinger archive, YouTube and Google videos. The experimental operating system is CentOS7 and two GPUs (NVIDIA TitanX).

### 4.2 Training Details

The input of our model is two modalities of video frame and optical flow image, so we need to cut the video frame first, and then calculate the optical flow image.

We sparsely sample the video and sample 8, 9, and 8 video clips in each of the three segments of the video. Each video snippets extracts one video frame and five optical flow images as inputs for spatial stream network and temporal stream network. The learning rate of the training network is 0.001. And because of the different input during network training, the parameter settings are also different. The input of the spatial stream network is a video frame, the training epochs is set to 80, and the deactivation rate is set to 0.8; the input of the temporal stream network is optical flow images, the training epochs is set to 340, and the deactivation rate is set to 0.7.

### 4.3 Experimental Results and Analysis

In view of the unreasonableness of snippet-level features corresponding to video-level labels and the lack of prominent features, we have improved and optimized on the basis of TSN, extracted video-level features, and integrated channel attention to excellently achieve the final human action recognition.

**Table 1.** Different methods for extracting partial features in HMDB51.

	Split1	Split2	Split3
Max pooling	70.27	69.02	69.11
Avg pooling	69.31	68.95	69.03
Max + Avg	70.39	69.25	69.19

In order to achieve partial consistency, the method of fusing snippet-level features, we considered three methods, namely maximum pooling, average pooling, a combination of maximum pooling and average pooling, the final results are shown in Table 1. It can be seen from the table that the fusion method using the maximum pooling is better than the average pooling effect to fuse snippet-level features, and it can be found through experiments that the form of combining the maximum pooling and the average pooling achieves the best effect on the hmdb51 dataset, it is also the method used in this paper. The principle is that when fusing snippet-level features, the features extracted from each clip are considered, and the snippet-level features that are more relevant to the label are highlighted.

We compare our model with the classic method as shown in Table 2, and finally our results have achieved more accurate human action recognition results. Among them, the average accuracy rate of the TSN network input as video frames and optical flow images is 68.5%. According to the parameter settings in the original paper, the reproduction result is 66.1%. This reproduction error is related to many factors. In the following, we make the reproduced TSN method be TSN\*. We can see that the method of extracting video-level features (VF) on the basis of TSN, is increased by 3.5% compared with TSN\*, which proves the

effectiveness of the method of extracting video-level features; after extracting video-level features, by adding an attention mechanism the accuracy rate of VFAM based on TSN\* has increased by 4.7%, and the accuracy rate based on VF has increased by 1.2%. And compared with other classic methods, our results have achieved more accurate human action recognition results.

**Table 2.** Results of different action recognition methods.

	Average accuracy
IDT [19]	57.2
Two-stream [15]	59.4
Two-stream Fusion [6]	69.2
TSN [22]	68.5
TDD [21]	65.9
I3D (two-stream) [13]	66.4
Three-Stream TSN [22]	69.4
STP [23]	68.9
L2STM [16]	66.2
TSN*	66.1
VF (ours)	69.6
VFAM (ours)	70.8

## 5 Conclusions

In this paper, we have proposed a human action recognition method based on video-level features and attention mechanism (VFAM), it has extracted video-level features, and made the extracted features more effective for human action recognition by adding attention mechanism during feature fusion. A lot of experiments have proved the feasibility and effectiveness of our method.

**Acknowledgment.** This paper is supported by Special subject of Innovation Method Work of the Ministry of Science and Technology (2018IM020200), Beijing Natural Science Foundation-Traffic Control Technology Funding Project (Railway Transportation Joint Fund Project) (Grant No. L191009).

## References

1. Bilen, H., Fernando, B., Gavves, E., Vedaldi, A.: Action recognition with dynamic image networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(12), 2799–2813 (2018)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset, pp. 4724–4733 (2017)

3. Crasto, N., Weinzaepfel, P., Alahari, K., Schmid, C.: Mars: Motion-augmented RGB stream for action recognition, pp. 7882–7891 (2019)
4. Diba, A., Fayyaz, M., Sharma, V., Karami, A.H., Arzani, M.M., Yousefzadeh, R., Van Gool, L.: Temporal 3D convnets: new architecture and transfer learning for video classification. *arXiv: Computer Vision and Pattern Recognition* (2017)
5. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal residual networks for video action recognition, pp. 3468–3476 (2016)
6. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition, pp. 1933–1941 (2016)
7. Fernando, B., Anderson, P., Hutter, M., Gould, S.: Discriminative hierarchical rank pooling for activity recognition, pp. 1924–1932 (2016)
8. Fernando, B., Gould, S.: Learning end-to-end video classification with rank-pooling, pp. 1187–1196 (2016)
9. Gao, R., Xiong, B., Grauman, K.: Im2flow: motion hallucination from static images for action recognition, pp. 5937–5947 (2018)
10. Garcia, N.C., Morerio, P., Murino, V.: Modality distillation with multiple stream networks for action recognition, pp. 106–121 (2018)
11. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1 (2019)
12. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 221–231 (2013)
13. Li, C., Zhong, Q., Xie, D., Pu, S.: Collaborative spatiotemporal feature learning for video action recognition, pp. 7872–7881 (2019)
14. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks, pp. 5534–5542 (2017)
15. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos, pp. 568–576 (2014)
16. Sun, L., Jia, K., Chen, K., Yeung, D., Shi, B.E., Savarese, S.: Lattice long short-term memory for human action recognition, pp. 2166–2175 (2017)
17. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks, pp. 4489–4497 (2015)
18. Tran, D., Wang, H., Torresani, L., Ray, J., Lecun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition, pp. 6450–6459 (2018)
19. Wang, H., Schmid, C.: Action recognition with improved trajectories, pp. 3551–3558 (2013)
20. Wang, L., Li, W., Van Gool, L.: Appearance-and-relation networks for video classification, pp. 1430–1439 (2018)
21. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors, pp. 4305–4314 (2015)
22. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition pp. 20–36 (2016)
23. Wang, Y., Long, M., Wang, J., Yu, P.S.: Spatiotemporal pyramid network for video action recognition, pp. 2097–2106 (2017)
24. Wang, Y., Zhou, L., Qiao, Y.: Temporal hallucinating for action recognition with few still images, pp. 5314–5322 (2018)



# Application of LSTM in Aeration of Sewage Treatment

Shaobo Zhang and Qinglin Sun<sup>(✉)</sup>

Nankai University, Tianjin 300350, China  
718659238@qq.com

**Abstract.** Long Short-Term Memory (LSTM) network is a form of Recurrent Neural Networks, it is suitable for predicting events relatively long delay in term series. In this study, LSTM is used to build a model for predicting dissolved oxygen concentration in sewage tank. This model can transform the collected data into time series data and predict the next moments data based on the current moments data. The experimental results show that the prediction errors of the proposed method are better than those of BP neural network and Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithms, which can make the sewage treatment plant more accurately adjust the air intake of blower to control the dissolved oxygen in the sewage pool.

**Keywords:** LSTM · Predicting · Sewage treatment · Dissolved oxygen

## 1 Research Background, Significance and Progress

In recent years, China has paid more and more attention to the treatment of environmental pollution, and the treatment of sewage is one of the important aspects. At this stage, China's sewage discharge is increasing year by year. From 2004 to 2018, China's sewage discharge increased from 48.24 billion tons to 79.5 billion tons, which shows the importance of sewage treatment. At the same time, the number of urban sewage treatment plants is increasing year by year. From 2009 to 2015, the number of urban sewage treatment plants increased from 1878 to 3542. By the end of 2016, there were 3991 sewage treatment plants, and the sewage treatment capacity has reached 173 million m<sup>3</sup> per day.

The treatment capacity of the sewage treatment plant is very large, and in the treatment process, the aeration unit accounts for 50% to 70% of the total energy consumption of the whole sewage treatment plant, which is the main energy consumption unit [1]. In the aeration process, the best concentration of dissolved oxygen is maintained by controlling the speed of the blower. According to the literature, the best biological reaction rate is generally maintained when the concentration of dissolved oxygen is 1 to 2 mg/L. At present, the aeration unit of sewage treatment plant in China has reached the level of informatization, but the level of intelligence is not high, and the traditional aeration mode is still generally used.

The traditional aeration process can be divided into two types. One is that multiple blowers operate in power frequency state for a long time at the same time. The other way is to optimize the aeration system by conventional PID control. Due to the strong hysteresis and non-linear characteristics of the aeration system, PID parameters need to be adjusted manually constantly, which requires high manpower demand, inaccurate control and poor immunity [2–4].

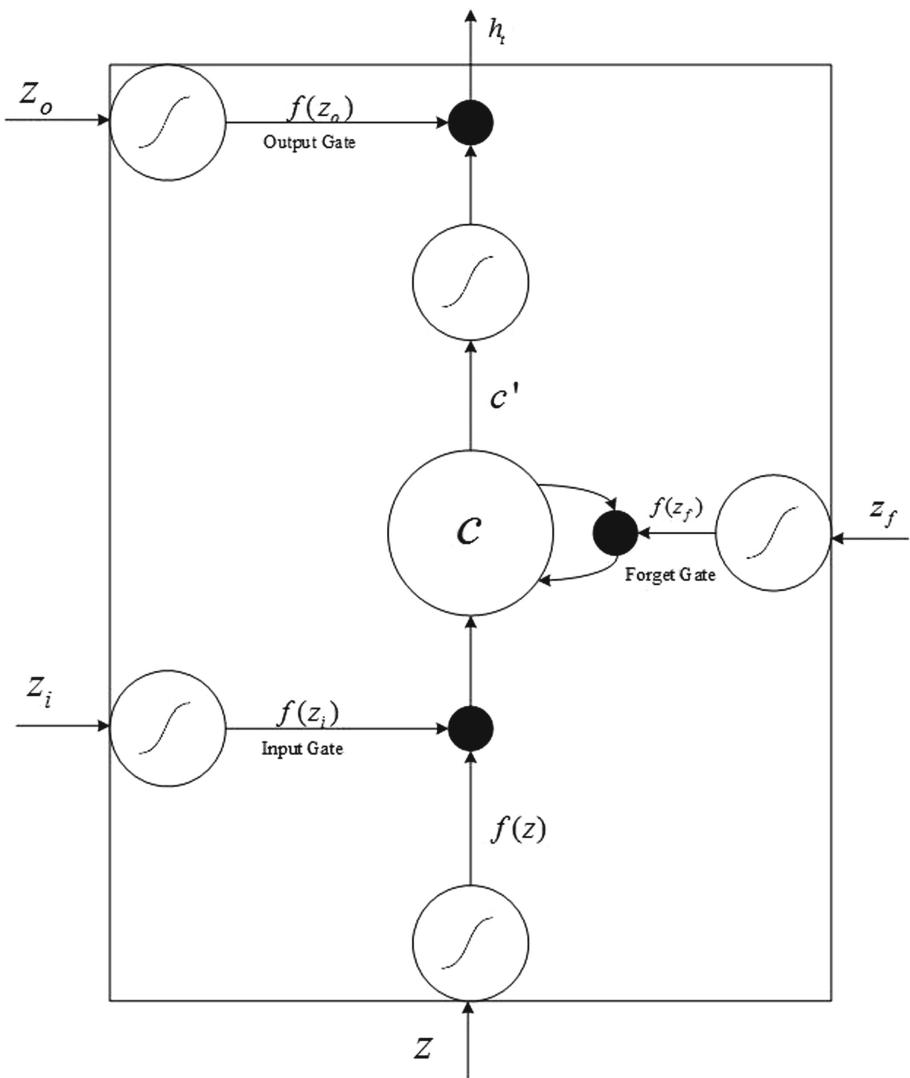
With the rise of artificial intelligence, some deep learning algorithms appear successively [5–7]. In the deep learning algorithm, Recurrent Neural Network (RNN) is a kind of model suitable for sequence data [8], which can extract effective information from time series data, and has been widely used in stock forecasting, speech recognition and other fields [9, 10]. Long Short-Term Memory (LSTM) [11] is a special form of RNN. Compared with RNN, LSTM can solve the problem of gradient explosion and gradient disappearance in the training process of RNN, and is more suitable for predicting relatively long time series information. In recent years, LSTM has been widely used in various industries [12–14]. In this paper, a model based on LSTM is proposed to predict the concentration of dissolved oxygen in the aerobic tank at the next time according to the concentration of each component in the aerobic tank at the current time and the displacement. By predicting the concentration of dissolved oxygen at the next time, the speed of the blower is adjusted, so as to achieve the effect of intelligent control.

## 2 Design of LSTM Model

LSTM is one of RNN neural networks. There are three gates in each LSTM unit, which are forget gate, input gate and output gate. The input gate is used to control the retention of input information and prevent useless information from entering the storage unit. The forget gate is used to selectively discard the information of the last moment. The output gate is used to control the output of information at each time. Compared with the gradient vanishing problem in simple RNN neural network, LSTM can solve this problem well.

LSTM is one of RNN neural networks. There are three gates in each LSTM unit, which are forget gate, input gate and output gate. The input gate is used to control the retention of input information and prevent useless information from entering the storage unit. The forget gate is used to selectively discard the information of the last moment. The output gate is used to control the output of information at each time. Compared with the gradient vanishing problem in simple RNN neural network, LSTM can solve this problem well. For time series  $x_t (t = 1, 2, 3 \dots)$ , the output of LSTM at the current time will be combined with the data at the next time as the input of the next time, and each time step has its output. At the same time, the memory unit generates the state vector of the current time step.

Figure 1 shows the internal structure of an LSTM unit, in which  $C$  represents the LSTM state information stored in the memory unit at the current time, black solid circle represents multiplication, and other hollow circles represent activation functions. In this LSTM unit, there is the following calculation process:

**Fig. 1.** Structure of LSTM unit

$$f(z_i) = \sigma(W^{(i)}H^{(i)} + b_i) \quad (1)$$

$$f(z_f) = \sigma(W^{(f)}H^{(f)} + b_f) \quad (2)$$

$$f(z_o) = \sigma(W^{(o)}H^{(o)} + b_o) \quad (3)$$

$$f(z) = \sigma(W^{(c)}H^{(c)} + b_c) \quad (4)$$

$$c' = f(z)f(z_i) + cf(z_f) \quad (5)$$

$$h_t = \sigma(c')f(z_o) \quad (6)$$

Where  $f(z_i)$ ,  $f(z_f)$ ,  $f(z_o)$  represent the output of input gate, forget gate and output gate and represents the state of cell at the current time.  $h_t$  represents the output of cell.  $W^{(i)}$ ,  $W^{(f)}$ ,  $W^{(o)}$  represent the weight of input gate, forgetting gate and output gate,  $H$  represents the vector formed by superposition of current input vector and output vector at the previous time.

### 3 Factors Affecting the Concentration of Dissolved Oxygen in Sewage Treatment Pools

All kinds of substances in the sewage treatment pool react complicatedly. By collecting samples and analyzing, it is concluded that there are 13 kinds of substances in the water and the discharge  $Q$ . There are 13 substances:

- Dissolved inert organic matter  $S_i$
- Easily biodegradable organic matter  $S_S$
- Particulate inert organic matter  $X_i$
- Slow biodegradation of organics  $X_s$
- Nitrate nitrogen  $S_{NO}$
- Ammonia nitrogen  $S_{NH}$
- Dissolved organic nitrogen  $S_{ND}$
- Granular biodegradable organic nitrogen  $X_{ND}$
- Heterotrophic bacteria  $X_{BH}$
- Autotrophic bacteria  $X_{BA}$
- Particulate products of microbial decay  $X_P$
- Dissolved oxygen  $S_O$
- Total alkalinity  $S_{ALK}$

In this paper, the concentration of each component, the amount of water discharged and the concentration of dissolved oxygen in the sewage treatment pool were measured for 14 days. The interval time of each measurement was the same, and 1344 pieces of data were obtained.

### 4 Establishment of Prediction Model of Sewage Aeration Based on LSTM

The specific operation steps of LSTM are as follows:

- (1) The concentration of each component of sewage and the time series of drainage volume are input into the input layer.
- (2) The LSTM receives the input vector. According to the output of the cell at the last time, when  $t = 1$ , the hidden layer status is set to 0. The input gate and forget gate of LSTM determine the information to be input and the information to be forgotten when entering the storage unit according to the output value of SIGMOD function. The output of the input gate and the forget gate updates the information in the storage unit, and finally determines the output information according to the storage information.

- (3) Take the component concentration and displacement at the next time and the output at the previous time as the input at the next time to enter the LSTM and repeat the above process.
- (4) At last, the output of LSTM is transmitted to the output layer to display the training results.

In this paper, LSTM is used to predict the dissolved oxygen. First, 1344 pieces of data about the concentration of each component and the amount of water discharged in 14 days are extracted. The first 1000 pieces of data are constructed as training sets, and the last 344 pieces of data are constructed as test sets.

The data needs pretreatment before training. Because there are big differences in the value range of component concentration and displacement, direct training will make some data have a great impact on the training results, while other data have no significant impact on the results. Therefore, all data will be normalized, that is, all factors will be reduced to the same range, and the formula is as follows:

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (7)$$

In the above formula,  $x^*$  represents the normalized result of a certain parameter,  $x$  represents the original value of the parameter,  $x_{max}$  and  $x_{min}$  represent the maximum and minimum values of the class in which the parameter is located.

This paper uses mean squared error (MSE) as the objective function, which is defined as follows:

$$MSE = \frac{\sum_{i=1}^N (y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2}{N} \quad (8)$$

MSE is the objective function value,  $N$  is the total number of data,  $i$  is the  $i$ th value of output vector,  $y_{test}^{(i)}$  and  $\hat{y}_{test}^{(i)}$  are the predicted value and actual value of the  $i$ th value of vector respectively.

## 5 Experimental Results

The programming language used in this paper is python, which is programmed with Python 3.6. The platform framework is based on tensorflow1.8.0. In the LSTM model, the hidden layer has 100 neurons, the output layer has one neuron, the input variable is a time step feature, the loss function uses MSE, the optimization algorithm uses Adam, the model uses 500 epochs and the batch size is 50. Using the given data set for training, the loss function curve drawn is shown in Fig. 2.

Figure 3 shows the comparison between the predicted value of the system and the real value. The blue solid line is the original data, the orange dotted line is the training result of the training set, and the green dotted line is the prediction result of the test set. In this paper, Root Mean Square Error (RMSE) is selected

as the index to evaluate the prediction effect of the model. The calculation formula of RMSE is as follows:

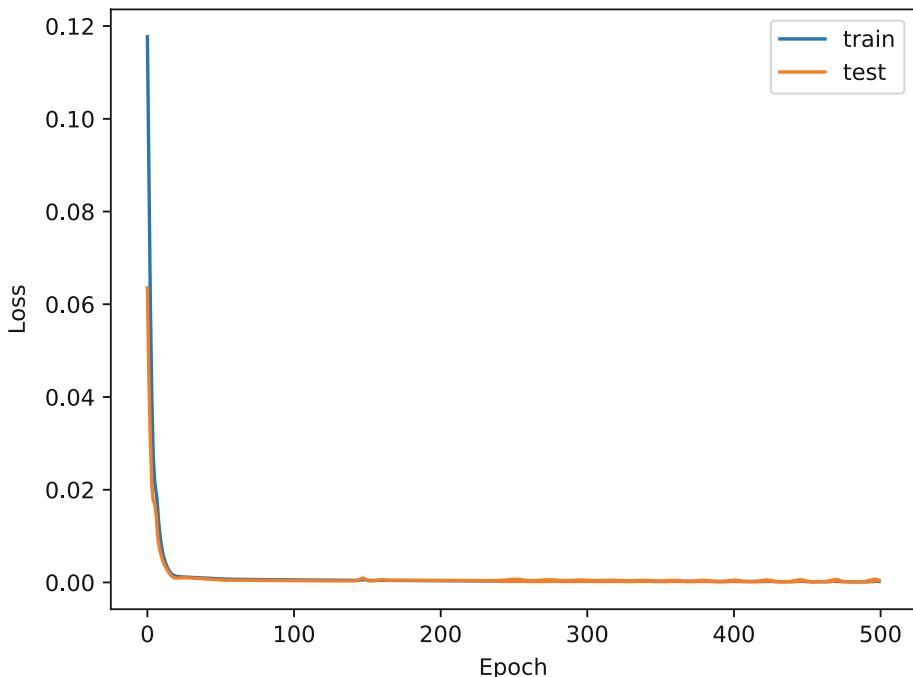
$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2}{N}} \quad (9)$$

At last, the RMSE of the training set is 0.041, and the RMSE of the test set is 0.040.

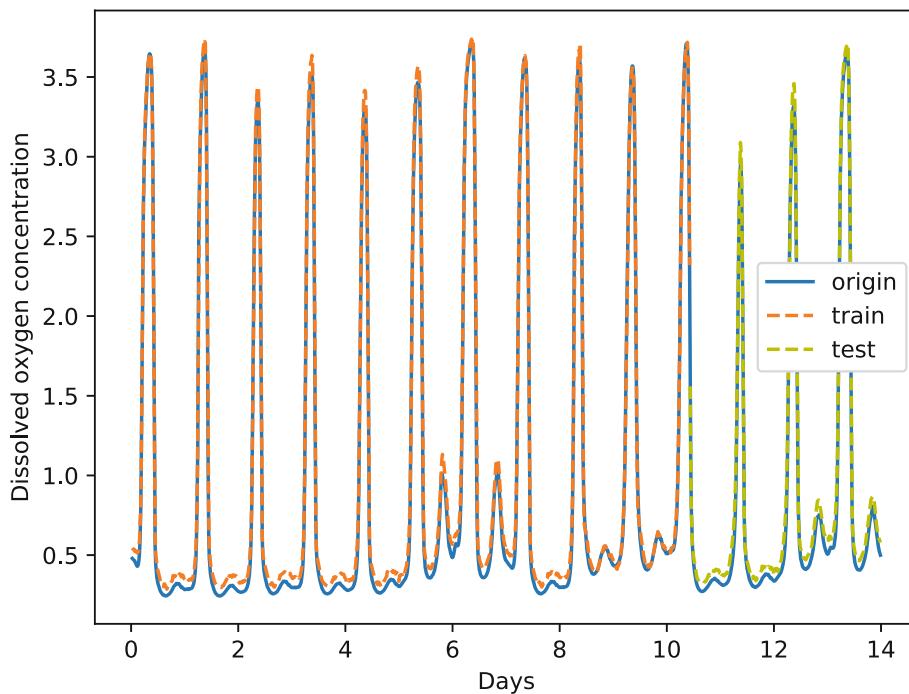
With the same data, BP neural network, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) are used to predict the dissolved oxygen content. Table 1 shows the RMSE of training set and test set obtained by various methods. It can be seen that the prediction accuracy of LSTM method is higher.

**Table 1.** Comparison of training results with different methods

RMSE	LSTM	BPNN	SVM	KNN
Training set	0.041	0.117	0.097	0.101
Test set	0.040	0.083	0.089	0.096



**Fig. 2.** Loss function curve of the model



**Fig. 3.** Comparison between predicted value and real value of the system

## 6 Conclusion

Based on LSTM, a model for the prediction of dissolved oxygen concentration in waste water treatment pool is established in this paper. This model can predict the concentration of dissolved oxygen at the next time according to the concentration of each component in the water at each time and the amount of water discharged. Through comparison, it can be found that the prediction results of LSTM method are better than those of BP neural network, SVM and k-nearest neighbor method, and more accurate prediction values can be obtained. The water treatment plant can control the blower according to the predicted value so as to maintain the dissolved oxygen in the tank at the optimal concentration. In the future, we will try to use more intelligent and accurate methods to predict dissolved oxygen.

## References

1. Jin, C.Q., Wang, C.W.: Analysis of energy consumption characteristics of sewage treatment plant and establishment of energy consumption index. *Constr. Technol.* **3**, 54–55 (2009)
2. Wang, X.W.: Water Pollution Control Project. Coal Industry Publisher, Beijing (2002)

3. He, S.J., Wang, H.X., Yang, L.G.: Deliquescent oxygen control of sewage disposal system of city. *Control Instr. Chem. Industry* **30**(1), 36–38 (2003)
4. Zhao, D.Q., Tong, Q.Y., Li, N.: Intelligent control of dissolved oxygen concentration in energy saving-based blast aeration system. *Water Wastewater Eng.* **34**(7), 116–119 (2008). <https://doi.org/10.13789/j.cnki.wwe1964.2008.07.002>
5. Cui, G.X., Li, D.K.: Overview on deep learning based on automatic encoder algorithms. *Comput. Syst. Appl.* **27**(9), 47–51 (2018). <https://doi.org/10.15888/j.cnki.csa.006542>
6. Tong, J.Y., Chang, X.L., Zhao, Y.J.: Real-time detection and positioning of moving target based on deep learning. *Comput. Syst. Appl.* **27**(8), 28–34 (2018). <https://doi.org/10.15888/j.cnki.csa.006525>
7. Zhang, J., Yu, J., Wang, J.L.: Image recognition technology for transmission line external damage based on depth learning. *Comput. Syst. Appl.* **27**(8), 176–179 (2018). <https://doi.org/10.15888/j.cnki.csa.006458>
8. Graves, A.: Generating sequences with recurrent neural networks (2013). [arXiv:1308.0850](https://arxiv.org/abs/1308.0850)
9. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the 31st International Conference on Machine Learning, Beijing, China pp. 1764–1772 (2014).
10. Sundermeyer, M., Ney, H., Schlter, R.: From feedforward to recurrent LSTM neural networks for language modeling. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(3), 517–529 (2015). <https://doi.org/10.1109/TASLP.2015.2400218>
11. Graves, A.: Long short-term memory. In: Graves, A. (ed.) *Supervised Sequence Labelling with Recurrent Neural Networks*, p. 37–45. Springer, Heidelberg (2012)
12. Yu, W., Zhou, W.N.: Sentiment analysis of commodity reviews based on LSTM. *Comput. Syst. Appl.* **27**(8), 159–163 (2018). <https://doi.org/10.15888/j.cnki.csa.006483>
13. Shi, M.F., Yang, Y., He, L., Chen, C.C.: Question categorization of community question answering by combining bi-LSTM and CNN with attention mechanism. *Comput. Syst. Appl.* **27**(9), 157–162 (2018). <https://doi.org/10.15888/j.cnki.csa.006536>
14. Cao, G.Q., Zhang, X.M., Chen, Y.F.: Early warning of landslide in mined mine dumping site based on PCA-LSTM. *Comput. Syst. Appl.* **27**(11), 252–258 (2018). <https://doi.org/10.15888/j.cnki.csa.006646>



# Sliding Mode Control on Coordination of Master-Slave Manipulator

Lijun Wang<sup>1,2(✉)</sup>, Ningxi Liu<sup>1,2</sup>, Jinkun Liu<sup>3</sup>, Tianyu Cao<sup>1,2</sup>, and Jiaxuan Yan<sup>1,2</sup>

<sup>1</sup> School of Automation and Electrical Engineering,  
University of Science and Technology Beijing, Beijing 100083, China  
[wanglj@ustb.edu.cn](mailto:wanglj@ustb.edu.cn)

<sup>2</sup> Key Laboratory of Knowledge Automation for Industrial Processes,  
Ministry of Education, Beijing 100083, China

<sup>3</sup> School of Automation Science and Electrical Engineering, Beihang University,  
Beijing 100191, China

**Abstract.** In this paper, two cooperative control targets of a master-slave manipulator system are considered. The first target is the mutual alignment control of the two manipulators, and the second target is the tracking command signal control of the manipulator system. These two controllers are based on sliding mode control method. The stability of the manipulator system is guaranteed by Lyapunov method. The effectiveness of the proposed methods in the presence of external disturbances is verified by simulation results.

**Keywords:** Master-slave manipulator · Sliding mode control · Lyapunov function · Coordinative control

## 1 Introduction

With the development of industrial technology, robotic manipulators have been widely studied in order to realize dangerous and repetitive work. With the further refinement of the mechanical manipulator in the industrial division of labor, people have created many derivative models of the manipulator, and the master-slave manipulator is one of them. The master-slave manipulator system can perform tasks that human operators cannot directly accomplish, such as handling toxic substances and hazardous areas [1], space operation [2] and so on.

In 1947, Argonne national atomic energy laboratory developed the world's first master-slave manipulator [3], since then, the master-slave manipulator has attracted a great deal of researches [4, 5]. In [4], Intuitive Surgical company successfully developed the Da Vinci master-slave remotely operated surgical robotic device for minimally invasive surgery. Nowadays, master-slave manipulator is generally composed of human operator, master manipulator, slave manipulator, communication network and environment interaction with slave manipulator [6]. In addition, the control methods of master-slave manipulator have been studied

extensively [7,8], such as PID control [7], adaptive control [8,9], bilateral control [10] and so on. In [7], a fuzzy self-tuning PID controller for master-slave manipulator force sensing system was proposed to solve large disturbances, delays and nonlinear problems. In [8], a novel adaptive bilateral control scheme for obtaining ideal responses for master-slave manipulator systems with uncertainties was proposed. In [9], a concept of virtual master manipulator was introduced to design the nonlinear adaptive controller for achieving stability and transparency in the sense of motion/force tracking. However, these control methods do not consider the flexible manipulator as part of the master-slave manipulator and coordinate the realization of multiple functions. In fact, for the operation of master-slave manipulators, it is often necessary for different types of manipulators to have good coordination ability to achieve multiple functions.

The concept of flexible manipulator is relative to the rigid manipulator, flexible manipulators have lighter-weight, lower-cost, higher-safety characteristics [11]. These characteristics have made flexible manipulators a major research direction for future manipulators. Based on these studies, the coordination among various manipulators can perform a task that is impossible for a single manipulator [12]. In terms of structure, coordinative control approaches can be divided into two types, centralized control and distributed control [13]. In terms of the control law, the existing coordinative control method include PID control [14], neural network control [15], predictive control [16], bilateral teleoperation control [17] and so on. In [14], a PID controller was designed to enforce motion tracking and formation control of master and slave vehicles. In [15], a neural network synchronization control method was proposed and applied to a multiple manipulator systems based on leader-follower network communication topology. These researches provide us with good ideas and the sliding mode control method is applied to achieve alignment and trajectory tracking in master-slave manipulator with external disturbances. Compared with the existing control methods, the tracking effect of the method proposed in this article will be faster and more accurate.

In this paper, the alignment and tracking control problems of master-slave manipulator are considered under the condition of external disturbance. The challenge of the proposed methods design is that control parameters should be carefully selected to achieve the optimal control effect in the process of coordinative control.

The rest of this paper is organized as follows. The model structures of two manipulator system are described in Sect. 2. The designs of manipulator alignment and manipulator tracking controllers are shown in Sect. 3 and Sect. 4. Simulation results and analysis are shown in Sect. 5, and eventually Sect. 6 is the conclusion.

## 2 System Model

In this paper, we consider a typical single-link flexible joint manipulator (Fig. 1) and a typical rigid manipulator (Fig. 2), the equations of the flexible joint

manipulator and rigid manipulator can be written as [18]:

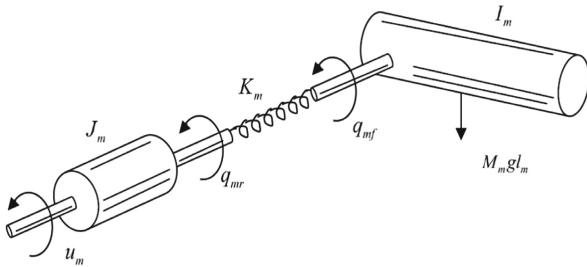
$$\begin{cases} I_m \ddot{q}_{mf} + K_m(q_{mf} - q_{mr}) + M_m g l_m \sin q_{mf} = 0 \\ J_m \ddot{q}_{mr} - K_m(q_{mf} - q_{mr}) = u_m + d_m \end{cases} \quad (1)$$

$$M(q_s) \ddot{q}_s + C(q_s, \dot{q}_s) \dot{q}_s + G(q_s) = u_s + d_s \quad (2)$$

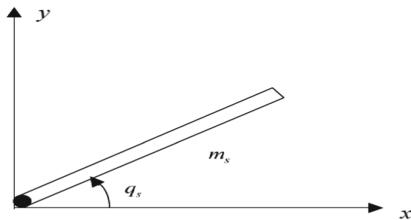
where, for the flexible manipulator model,  $q_{mf}$  and  $q_{mr}$  are the angular positions of the connecting rod and rotor of the flexible manipulator respectively,  $K_m$  represents the elastic stiffness of the flexible joint manipulator.  $M_m$ ,  $g$ ,  $l_m$  are the link mass, gravity acceleration and the length from the center of gravity of the link to the joint.  $u_m$  represents the motor torque input of the flexible manipulator. For the rigid manipulator model,  $q_s$  is the angular position of the rigid joint manipulator.  $M(q_s)$ ,  $C(q_s, \dot{q}_s)$ ,  $G(q_s)$  are the  $n \times n$ ,  $n \times n$ ,  $n \times 1$  order function matrix determined by the specific structure of the manipulator.  $d_m$  and  $d_s$  represent the external input disturbances.

Let the state variable  $x_1 = q_{mf}$  and  $x_3 = q_{mr}$ , the flexible manipulator dynamics can be rewritten as

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -\frac{1}{I_m} (M_m g l_m \sin x_1 + K_m(x_1 - x_3)) \\ \dot{x}_3 = x_4 \\ \dot{x}_4 = \frac{1}{J_m} (u_m + d_m - K_m(x_3 - x_1)) \end{cases} \quad (3)$$



**Fig. 1.** Single link flexible joint manipulator



**Fig. 2.** Rigid manipulator

The above equation can be written in the form of an underactuated system as follows

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = a_1 x_3 + f_1(x_1) \\ \dot{x}_3 = x_4 \\ \dot{x}_4 = a_2 u_m + f_2(x_1, x_3) + d_m \end{cases} \quad (4)$$

where  $a_1 = K_m/I_m$ ,  $f_1(x_1) = -M_m g l_m/I_m \cdot \sin x_1 - K_m/I_m \cdot x_1$ ,  $a_2 = 1/J_m$ ,  $f_2(x_1, x_3) = K_m/J_m \cdot (x_1 - x_3)$ .

### 3 Master-Slave Manipulator Alignment Controller Design

In this section, we introduce a sliding mode function method to design the angle alignment of the two manipulators. The specific control method design and stability proof will be shown below.

For the convenience of expression, we use  $f_1$  to replace  $f_1(x_1)$ ,  $f_2$  to replace  $f_2(x_1, x_3)$ . Therefore, the control target of the flexible manipulator changes to  $q_{mf} - q_s \rightarrow 0$ ,  $x_2 \rightarrow 0$ , the control target of the rigid manipulator changes to  $q_s - q_{mf} \rightarrow 0$ ,  $\dot{q}_s \rightarrow 0$ .

The error equations are designed as

$$\begin{cases} e_1 = x_1 - q_s \\ e_2 = x_2 \\ e_3 = \dot{e}_2 = \dot{x}_2 = a_1 x_3 + f_1 \\ e_4 = \ddot{e}_2 = a_1 \dot{x}_3 + \dot{f}_1 = a_1 x_4 + \dot{f}_1 \end{cases} \quad \text{and} \quad \begin{cases} e_s = q_s - q_{mf} \\ \dot{e}_s = \dot{q}_s \end{cases} \quad (5)$$

Then we design the sliding mode functions as

$$s_1 = c_1 e_1 + c_2 e_2 + c_3 e_3 + e_4 \quad (6)$$

$$s_2 = \dot{e}_s + \lambda e_s \quad (7)$$

where  $c_i (i=1-3)$ ,  $\lambda$  is positive constant.

The control methods are designed as

$$u_m = -\frac{1}{a_1 a_2} \left[ c_1 x_2 + c_2 (a_1 x_3 + f_1) + c_3 (a_1 x_4 + \dot{f}_1) + a_1 f_2 + \ddot{f}_1 + \eta_1 \operatorname{sgn}(s_1) \right] \quad (8)$$

$$u_s = G_s + M_s \ddot{q}_{mf} - M_s \lambda \dot{e}_s + C_s \dot{q}_{mf} - C_s \lambda e_s - \eta_2 \operatorname{sgn}(s_2) \quad (9)$$

Take the derivative of Eq. (6), and substitute Eq. (8) into it

$$\begin{aligned} \dot{s}_1 &= c_1 \dot{e}_1 + c_2 \dot{e}_2 + c_3 \dot{e}_3 + \dot{e}_4 \\ &= c_1 x_2 + c_2 (a_1 x_3 + f_1) + c_3 (a_1 x_4 + \dot{f}_1) + a_1 (a_2 u_m + f_2 + d_m) + \ddot{f}_1 \\ &= -\eta_1 \operatorname{sgn}(s_1) + a_1 d_m \end{aligned} \quad (10)$$

In order to guarantee that the manipulator system meets the requirements, we use Lyapunov's method to judge the stability of the system. The Lyapunov function is designed as

$$V = \frac{1}{2} s_1^2 + \frac{1}{2} s_2^T M_s s_2 \quad (11)$$

The derivative of above equation is

$$\begin{aligned}
\dot{V} &= s_1 \dot{s}_1 + s_2^T M_s \dot{s}_2 + \frac{1}{2} s_2^T \dot{M}_s s_2 \\
&= s_1 [c_1 x_2 + c_2(a_1 x_3 + f_1) + c_3(a_1 x_4 + \dot{f}_1) + a_1(a_2 u_m + f_2) + \ddot{f}_1] \\
&\quad + s_2^T M_s \dot{s}_2 + s_2^T C_s s_2 \\
&= -\eta_1 |s_1| + a_1 d_m s_1 + s_2^T (M_s \dot{s}_2 + C_s s_2) \\
&= -\eta_1 |s_1| + a_1 d_m s_1 + s_2^T (M_s (\ddot{q}_s - \ddot{q}_{mf} + \lambda \dot{e}_s) + C_s (\dot{e}_s + \lambda e_s)) \\
&= -\eta_1 |s_1| + a_1 d_m s_1 \\
&\quad + s_2^T [u_s + d_s - G_s - M_s \ddot{q}_{mf} + M_s \lambda \dot{e}_s - C_s \dot{q}_{mf} + C_s \lambda e_s] \\
&= -\eta_1 |s_1| + a_1 d_m s_1 - \eta_2 |s_2| + d_s s_2 \leq -\theta_1 |s_1| - \theta_2 |s_2| \leq 0
\end{aligned} \tag{12}$$

where  $\theta_1 = \eta_1 - a_1 |d_m|_{max} \geq 0$ ,  $\theta_2 = \eta_2 - |d_s|_{max} \geq 0$ .

According to Eq. (6), when  $s_1 = 0$ , then  $e_4 = -c_1 e_1 - c_2 e_2 - c_3 e_3$ , we take

$E_1 = [e_1 \ e_2 \ e_3]^T$ ,  $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -c_1 & -c_2 & -c_3 \end{bmatrix}$ , thus  $\dot{E}_1 = AE_1$ , by selecting the appropriate values of  $c_i$  ( $i=1-3$ ),  $A$  can be Hurwitz, then we can easily get that when  $t \rightarrow \infty$ ,  $\dot{E}_1 = [e_1 \ e_2 \ e_3]^T \rightarrow 0$ .

In order to make  $A$  become Hurwitz,  $A$  is designed as

$$\begin{aligned}
|A - \lambda I| &= \begin{vmatrix} -\lambda & 1 & 0 \\ 0 & -\lambda & 1 \\ -c_1 & -c_2 & -c_3 - \lambda \end{vmatrix} = \lambda^2(-c_3 - \lambda) - c_1 - c_2 \lambda \\
&= -\lambda^3 - c_3 \lambda^2 - c_2 \lambda - c_1 = 0
\end{aligned} \tag{13}$$

The real part of the root is taken as a negative number to meet the requirements. Because  $V \leq 0$ , so  $e_1 \rightarrow 0$ ,  $e_2 \rightarrow 0$ ,  $e_3 \rightarrow 0$ , and  $s_1 = c_1 e_1 + c_2 e_2 + c_3 e_3 + e_4$ , then  $e_4 \rightarrow 0$ . According to Lasalle invariance principle, the closed-loop system is asymptotically stable. We can draw a conclusion that when  $t \rightarrow \infty$ , then  $q_{mf} - q_s \rightarrow 0$ ,  $\dot{q}_{mf} \rightarrow 0$ ,  $\dot{q}_s \rightarrow 0$ . The flexible and rigid manipulators will eventually reach stable state at the same angle.

## 4 Master-Slave Manipulator Trajectory Tracking Controller Design

In this section, the control purposes of the master-slave manipulator system are that the flexible manipulator tracks the required trajectory, and the rigid manipulator tracks the trajectory of the flexible manipulator, thereby achieving the result of indirectly controlling the rigid manipulator. The ideal trajectory is given as  $x_d$ , so the control target of the flexible manipulator changes to  $x_1 \rightarrow x_d$ ,  $x_2 \rightarrow \dot{x}_d$ , the control target of the rigid manipulator changes to  $q_s \rightarrow x_1$ ,  $\dot{q}_s \rightarrow x_2$ .

The error equations are designed as

$$\begin{cases} e_1 = x_1 - x_d \\ e_2 = \dot{e}_1 = x_2 - \dot{x}_d \\ e_3 = \ddot{e}_1 = \dot{x}_2 - \ddot{x}_d = a_1 x_3 + f_1 - \ddot{x}_d \\ e_4 = e_1^{(3)} = a_1 \dot{x}_3 + \dot{f}_1 - x_d^{(3)} = a_1 x_4 + \dot{f}_1 - x_d^{(3)} \end{cases} \quad \text{and} \quad \begin{cases} e_s = q_s - q_{mf} \\ \dot{e}_s = \dot{q}_s - \dot{q}_{mf} \end{cases} \tag{14}$$

Then we design the sliding mode functions as

$$r_1 = b_1 e_1 + b_2 e_2 + b_3 e_3 + e_4 \quad (15)$$

$$r_2 = \dot{e}_s + \Lambda e_s \quad (16)$$

where  $b_i (i=1-3)$ ,  $\Lambda$  is positive constant.

The control laws are designed as

$$\begin{aligned} u_m = & -\frac{1}{a_1 a_2} \left[ b_1(x_2 - \dot{x}_d) + b_2(b_1 x_3 + f_1 - \ddot{x}_d) + b_3(a_1 x_4 + \dot{f}_1 - x_d^{(3)}) \right] \\ & - \frac{1}{a_1 a_2} \left[ a_1 f_2 + \ddot{f}_1 - x_d^{(4)} + \eta_1 \operatorname{sgn}(r_2) \right] \end{aligned} \quad (17)$$

$$u_s = G_s + M_s \ddot{q}_{mf} - M_s \Lambda \dot{e}_s + C_s \Lambda e_s - \eta_2 \operatorname{sgn}(r_2) \quad (18)$$

Substitute Eq. (17) into the derivative of Eq. (15), then

$$\begin{aligned} \dot{r}_1 &= b_1 \dot{e}_1 + b_2 \dot{e}_2 + b_3 \dot{e}_3 + \dot{e}_4 \\ &= b_1(x_2 - \dot{x}_d) + b_2(a_1 x_3 + f_1 - \ddot{x}_d) + b_3(a_1 x_4 + \dot{f}_1 - x_d^{(3)}) \\ &\quad + a_1(a_2 u_m + f_2 + d_m) + \ddot{f}_1 - x_d^{(4)} \\ &= -\eta_1 \operatorname{sgn}(r_1) + a_1 d_m \end{aligned} \quad (19)$$

Similarly, the stability is guaranteed by Lyapunov method, we designed Lyapunov function as

$$V = \frac{1}{2} r_1^2 + \frac{1}{2} r_2^T M_s r_2 \quad (20)$$

The derivative of above equation is

$$\begin{aligned} \dot{V} &= r_1 \dot{r}_1 + r_2^T M_s \dot{r}_2 + \frac{1}{2} r_2^T \dot{M}_s r_2 \\ &= -\eta_1 |r_1| + a_1 d_m r_1 + r_2^T (M_s \dot{r}_2 + C_s r_2) \\ &= -\eta_1 |r_1| + a_1 d_m r_1 + r_2^T (M_s (\ddot{q}_s - \ddot{q}_{mf} + \Lambda \dot{e}_s) + C_s (\dot{e}_s + \Lambda e_s)) \\ &= -\eta_1 |r_1| + a_1 d_m r_1 \\ &\quad + r_2^T [u_s + d_s - G_s - M_s \ddot{q}_{mf} + M_s \Lambda \dot{e}_s - C_s \dot{q}_{mf} + C_s \Lambda e_s] \\ &= -\eta_1 |r_1| + a_1 d_m r_1 - \eta_2 |r_2| + d_s r_2 \leq -\theta_1 |r_1| - \theta_2 |r_2| \leq 0 \end{aligned} \quad (21)$$

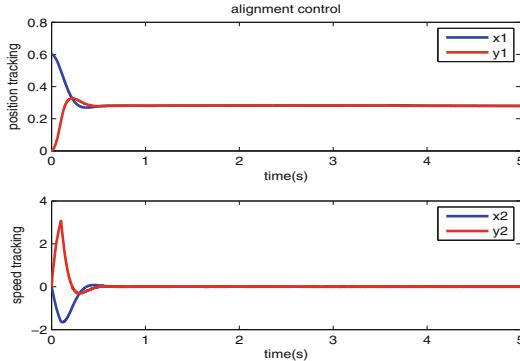
where  $\theta_1 = \eta_1 - a_1 |d_m|_{max} \geq 0$ ,  $\theta_2 = \eta_2 - |d_s|_{max} \geq 0$ .

Same as the content in the previous section, we can easily obtain  $V \leq 0$ , then  $e_1 \rightarrow 0$ ,  $e_2 \rightarrow 0$ ,  $e_3 \rightarrow 0$ ,  $e_4 \rightarrow 0$ . The value of the designed parameters  $b_i (i=1-3)$  can be used to satisfy the structure of the matrix for Hurwitz. When  $t \rightarrow \infty$ , then  $x_1 \rightarrow x_d$ ,  $x_2 \rightarrow \dot{x}_d$ ,  $q_s \rightarrow x_1$ ,  $\dot{q}_s \rightarrow x_2$  are satisfied. The stability of the closed-loop system can be guaranteed by LaSalle invariance principle.

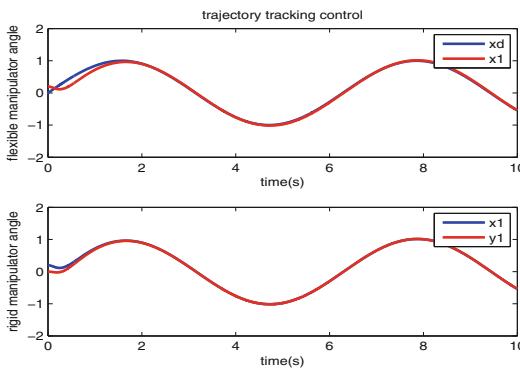
## 5 Simulation Results

In this section, the effectiveness of the trajectory tracking and alignment controllers are illustrated by numerical simulations. The simulation objective is a two-link rigid-flexible manipulator and a typical rigid manipulator, which is expressed by Eqs. (1) and (2). In this paper, the control methods are designed

for two kinds of motion problems of the manipulator system. For the sake of convenience, the manipulator models will remain unchanged. The parameters of manipulator system are as follows: the parameters of the flexible manipulator  $I_m = J_m = 1 \text{ kg} \cdot \text{m}^2$ ,  $M_m g l_m = 5 \text{ N} \cdot \text{m}$ ,  $K_m = 40 \text{ N} \cdot \text{m/rad}$ , the parameters of the rigid manipulator  $M_s(q_s) = 0.1 + 0.06\sin(q_s)$ ,  $C_s(q_s, \dot{q}_s) = 0.03\cos(q_s)$ ,  $G_s(q_s) = m_s g l_s \cos(q_s)$ ,  $m_s = 0.02 \text{ kg}$ .



**Fig. 3.** Alignment control responses



**Fig. 4.** Trajectory tracking control responses

The parameters of control laws for the two control targets of manipulator system are as follows:

- (1) alignment control design. The controller parameters of flexible manipulator are set as  $c_1 = 1000$ ,  $c_2 = 300$ ,  $c_3 = 30$ ,  $\eta_1 = 50$ , initial positions  $q_m(0) = [0.6 \ 0 \ 0 \ 0]^T$ , external disturbance  $d_m = 0.1\sin(t)$ ; the controller parameters

- of rigid manipulator are set as  $\eta_2 = 5$ ,  $\lambda = 10$ , initial positions  $q_s(0) = [0 \ 0]^T$ , external disturbance  $d_s = 0.1\sin(t)$ ;
- (2) tracking control design. The controller parameters of flexible manipulator are set as  $c_1 = 1000$ ,  $c_2 = 300$ ,  $c_3 = 30$ ,  $\eta_1 = 1.5$ , initial positions  $q_m(0) = [0.2 \ 0 \ 0 \ 0]^T$ ,  $d_m = 0.1\sin(t)$ ; the controller parameters of rigid manipulator are set as  $\eta_1 = 0.5$ ,  $A = 2$ , initial positions  $q_s(0) = [0 \ 0]^T$ , external disturbance  $d_s = 0.1\sin(t)$ ; the ideal signal is set as  $x_d = \sin(t)$ .

The simulation results are shown in Fig. 3 and Fig. 4. In Fig. 3, the upper waveform is the result of angles of the flexible and rigid manipulators, and the lower waveform is the result of the speed change of the two manipulators. It can be seen that the angles of flexible and rigid manipulators are finally in the same position, and the speed results eventually change to 0. In Fig. 4, the upper waveform is the result of the tracking ideal signal of the flexible manipulator, and the lower waveform is the result of the tracking of the flexible manipulator by the rigid manipulator. It can be seen that both two manipulator can track the target curve quickly and accurately. In connection with the conclusion of the previous section, it can be proved that the coordinated control method in this paper has a good control effect.

## 6 Conclusions

In this paper, a sliding mode control method is proposed to control a master-slave manipulator system of a rigid and flexible manipulator with external disturbances. The functions of master-slave manipulator alignment and trajectory tracking are realized, the external disturbances are solved by adding robust terms. System has the ability of anti-disturbance, and can complete the coordinative control targets. Lyapunov function is designed to guarantee the stability of the closed-loop system. The effectiveness of proposed methods are verified by simulation, simulation results show that the alignment and trajectory tracking functions of manipulator system can be achieved quickly and accurately.

## References

- Yoon, W.K., Goshozono, T., Kawabe, H., Kinami, M., Tsumaki, Y., Uchiyama, M., Oda, M., Doi, T.: Model-based space robot teleoperation of ETS-VII manipulator. *IEEE Trans. Robot. Autom.* **20**(3), 602–612 (2004)
- Wei, W., Kui, Y.: Teleoperated manipulator for leak detection of sealed radioactive sources. In: 2004 Proceedings of IEEE International Conference on Robotics and Automation, ICRA 2004, vol. 2, pp. 1682–1687. IEEE (2004)
- Goertz, R.C.: Fundamentals of general-purpose remote manipulators. *Nucleonics* **10**(11), 36–42 (1952)
- Ding, J., Xu, K., Goldman, R., Allen, P., Fowler, D., Simaan, N.: Design, simulation and evaluation of kinematic alternatives for insertable robotic effectors platforms in single port access surgery. In: 2010 IEEE International Conference on Robotics and Automation, pp. 1053–1058. IEEE (2010)

5. Swinnen, K., Politis, C., Willems, G., De Bruyne, I., Fieuws, S., Heidbuchel, K., Erum, R., Verdonck, A., Carels, C.: Skeletal and dento-alveolar stability after surgical-orthodontic treatment of anterior open bite: a retrospective study. *Eur. J. Orthod.* **23**(5), 547–557 (2001)
6. Hokayem, P.F., Spong, M.W.: Bilateral teleoperation: an historical survey. *Automatica* **42**(12), 2035–2057 (2006)
7. Liang, X., Chen, Z., Sun, Y., Liao, H.: Design of fuzzy self-tuning PID controller for master-slave manipulator force sensing system. In: 2017 6th International Conference on Measurement, Instrumentation and Automation (ICMIA 2017), vol. 154, pp. 556–562. Atlantis Press (2017)
8. Ryu, J.H., Kwon, D.S.: A novel adaptive bilateral control scheme using similar closed-loop dynamic characteristics of master/slave manipulators. *J. Robotic Syst.* **18**(9), 533–543 (2001)
9. Hung, N.V.Q., Narikiyo, T., Tuan, H.D.: Nonlinear adaptive control of master-slave system in teleoperation. *Control Eng. Pract.* **11**(1), 1–10 (2003)
10. Wang, J., Dad, K., Lee, M. C.: Bilateral control of hydraulic servo system based on SMCSPO for 1DOF master slave manipulator. In: Proceedings of the 2017 International Conference on Artificial Life and Robotics (ICAROB 2017), Miyazaki, Japan, pp. 19–22 (2017)
11. Wang, L., Shi, Q., Liu, J., Zhang, D.: Backstepping control of flexible joint manipulator based on hyperbolic tangent function with control input and rate constraints. *Asian J. Control* (2018). <https://doi.org/10.1002/asjc.2006>
12. Zhai, D.H., Xia, Y.: Adaptive fuzzy control of multilateral asymmetric teleoperation for coordinated multiple mobile manipulators. *IEEE Trans. Fuzzy Syst.* **24**(1), 57–70 (2016)
13. Wang, X., Zeng, Z., Cong, Y.: Multi-agent distributed coordination control: Developments and directions via graph viewpoint. *Neurocomputing* **199**, 204–218 (2016)
14. Rodríguez-Seda, E.J., Troy, J.J., Erignac, C.A., Murray, P., Stipanovic, D.M., Spong, M.W.: Bilateral teleoperation of multiple mobile agents: coordinated motion and collision avoidance. *IEEE Trans. Control Syst. Technol.* **18**(4), 984–992 (2009)
15. Zhao, D., Zhu, Q., Li, N., Li, S.: Synchronized control with neuro-agents for leader-follower based multiple robotic manipulators. *Neurocomputing* **124**, 149–161 (2014)
16. Iqbal, K., Zheng, Y.F.: Predictive control application in arm manipulator coordination. In: Proceedings of 12th IEEE International Symposium on Intelligent Control, pp. 409–414. IEEE (1997)
17. Deka, S.A., Stipanović, D.M., Kesavadas, T.: Stable bilateral teleoperation with bounded control. *IEEE Trans. Control Syst. Technol.* **27**(6), 2351–2360 (2019)
18. Spong, M.W.: Modeling and control of elastic joint robots. *J. Dyn. Syst. Meas. Contr.* **109**(4), 310–319 (1987)



# Pose Ambiguity Elimination Algorithm for 3C Components Assembly Pose Estimation in Point Cloud

Weikun Gu, Xiansheng Yang, Dengwei Dong, and Yunjiang Lou<sup>(✉)</sup>

Harbin Institute of Technology Shenzhen, Shenzhen 518000, China  
louyj@hit.edu.cn

**Abstract.** Pose ambiguity is an inevitable problem in object pose estimation which means it is difficult or impossible to distinguish between several possible poses based on known observations and it is often caused by object symmetry, occlusion, less or repetitive textures. In 3C assembly, components often have plane symmetrical structure. When using traditional point cloud registration algorithms to estimate 6D pose of 3C components, an incorrect ambiguous pose is often obtained, which will cause assembly failure. In order to solve the pose ambiguity problem in 3C assembly, a pose ambiguity elimination algorithm based on PCA (Principal Component Analysis) and 2D image template matching is proposed. In elimination of ambiguity, Our method has higher accuracy than traditional methods and the efficiency meets our need. The reliability and effectiveness are verified by simulation and experiments.

**Keywords:** Pose ambiguity elimination · Pose estimation · Point cloud · Template matching

## 1 Introduction

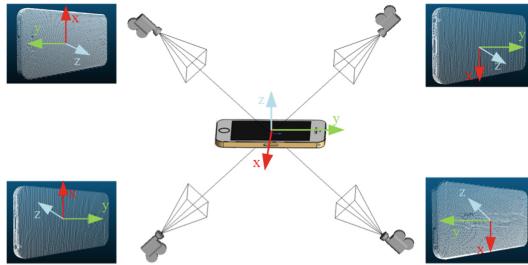
3C is the collective name for computers, communications and consumer electronics. The development of robotics and vision technology has improved the automation degree of 3C assembly. In recent years, 3D vision technology is becoming more and more mature and is gradually being used in the fields of object 6D pose estimation and robot guidance, such as Bin-picking and 3C assembly automation. 3D sensor can acquire depth data or point cloud data and point cloud registration is the common method for object pose estimation. However, pose ambiguity becomes a big problem when point cloud registration is used for pose estimation in 3C assembly.

For object pose estimation, two point clouds are used in point cloud registration; one is part point cloud captured by 3D sensor, the other is complete point

---

Gu, Yang, Dong, Lou (the corresponding author) is with the School of Mechanical Engineering and Automation, Harbin Institute of Technology Shenzhen, HIT Campus, University Town of Shenzhen, Xili, Nanshan, Shenzhen, China.

Y. Lou—IEEE Senior Member.



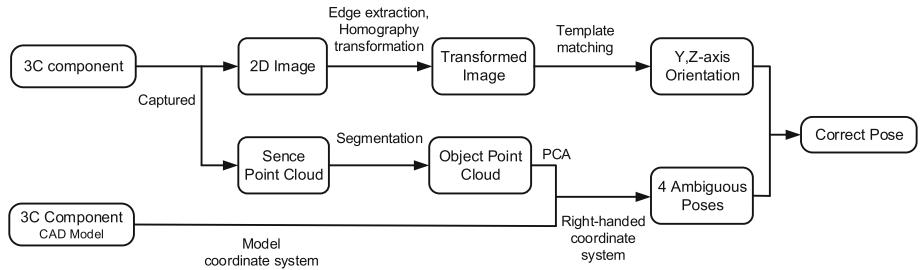
**Fig. 1.** Pose ambiguity: 3D sensors in 4 different viewpoints and the 4 point clouds captured by corresponding 3D sensors. The color information is unavailable and the symmetrical structure cause pose ambiguity. The object local coordinate system is shown, in this paper, we define phone head direction as positive Y-axis direction and screen side as positive Z-axis direction.

cloud obtained by sampling the object CAD model. The color information in captured point cloud is unavailable in point cloud registration because the CAD model of 3C components usually have no color. Geometrically, 3C components often have a plane symmetrical structure, such as screen, battery and back plane of mobile phone, which makes it difficult to determine the right pose from the part point cloud captured. Taking mobile phone model (including screen, front cover and back plane) as an example, using 4 3D sensors to capture the object in four different viewpoint in object coordinate system (shown as Fig. 1), we find that the 4 point clouds captured are very similar in geometry. But in fact they correspond to four different poses whose axis direction are partly opposite to each other. In this paper, we define this situation as the pose ambiguity problem in the point cloud. In that case, we can't distinguish the correct pose from 4 ambiguous pose, which can make assembly failed but still success in bin-picking because robots can grab the object no matter which pose is correct.

According to Manhardt's research [1], pose estimation from a single image is a inherently ambiguous problem, which caused by shape symmetries, occlusion and repetitive textures. One image may correspond to several indistinguishable poses which should be treated as equivalent [2]. Most 6d pose estimation neural networks use ADD-S(Average Distance of Model Points-Symmetric) for symmetric objects such as DenseFusion-6D [3] and PVNet [4], which treats ambiguous poses as equivalent.

However, our problem is different; the pose ambiguity is caused by point clouds geometrically similarity and 3C components have patterns and textures on their surface, which makes the problem solvable. Point cloud registration is the common method for object pose estimation and the most classic point cloud registration algorithm is ICP(Iterative closest point) [5,6]. However, ICP is an accurate registration algorithm and it needs initial pose. Initial registration methods are usually based on representative exhaustive search [7] and feature correspondence [8], such as RANSAC(Random Sample Consensus) [9], FPFH [10] (Fast Point Feature Histogram), SAC-IA (Sample Consensus Initial

Alignment) and 4PCS(4-Points Congruent Sets), etc. However, due to random selection, highly symmetric point cloud and the similarity of feature vectors proposed by point cloud descriptors, all the initial registration methods obtain ambiguous poses. In addition, PCA is also used to calculate the initial pose, but the axis direction is uncertain, which means there are 8 possible ambiguous poses. Chen et al. [11] chose the pose with minimum distance (alignment error) in these 8 cases. Unfortunately, it may generate ambiguous pose due to object symmetry and pose accuracy from PCA.



**Fig. 2.** Algorithm framework

In summary, the traditional initial pose estimation methods cannot solve the pose ambiguity problem for highly symmetrical 3C components. In previous research, our group [12] proposed an algorithm based on Linemod to eliminate pose ambiguity. However it is sensitive to working distance and requires lots of templates to ensure accuracy which cost much time and RAM. Therefore, we further propose the pose ambiguity elimination algorithm based on PCA and shape matching.

The remainder of this paper is organized as follows: In Sect. 2, we detail our method. In Sect. 3, we introduce the specific process and results of simulation and experiment. In Sect. 4, we conclude the paper with a summary of key points and mention our future work.

## 2 Pose Ambiguity Elimination Algorithm

In this paper, we propose a pose ambiguity elimination algorithm for highly symmetrical 3C components in point cloud. The algorithm framework is shown as Fig. 2. First, the 3D sensor captures the scene point cloud and 2D image. The scene point cloud is segmented to obtain the object point cloud of the 3C components. Through PCA, CAD model coordinate system and right-hand coordinate system constraint, we can obtain 4 ambiguity poses. As for captured 2D image, we transfer it by homography after image pre-processing to obtain the top view of the 3C component and template matching algorithm is used to determine the Z-axis and Y-axis positive direction of 3C component local coordinate system (consistent with the CAD model coordinate system). Finally, we can find the correct pose according to the axes direction.

## 2.1 Point Cloud Registration Based on PCA

We use PCA for initial registration due to the high efficiency compared to RANSAC, SAC-IA, 4PCS and descriptors-based methods. In PCA process, SVD is performed on captured point cloud's covariance matrix. The eigenvectors corresponding to the three eigenvalues form the local coordinate system of 3C component, which is consistent with the CAD model coordinate system. The covariance matrix be calculated by formula as follows:

$$J = \frac{1}{n} \sum_{i=1}^n \left\{ \left( p_i - \frac{1}{n} \sum_{i=1}^n p_n \right) \cdot \left( p_i - \frac{1}{n} \sum_{i=1}^n p_n \right)^T \right\} \quad (1)$$

The covariance matrix is decomposed by SVD to obtain the feature vector and as follows:

$$N = \begin{bmatrix} N_1 \\ N_2 \\ N_3 \end{bmatrix} = \begin{bmatrix} n_{11} & n_{12} & n_{13} \\ n_{21} & n_{22} & n_{23} \\ n_{31} & n_{32} & n_{33} \end{bmatrix} \quad (2)$$

The eigenvalues v1, v2, v3 corresponding to eigenvector  $N_1, N_2, N_3$  decreases means the point cloud component along vector  $N_1, N_2, N_3$  decreases, which we can use to determine X/Y/Z axis. In our estimation scheme, vector  $N_1, N_2, N_3$  correspond to Y, X, Z axis respectively. Since the positive direction of the X/Y/Z axis is uncertain, there may be 4 ambiguous poses that satisfy the right-hand coordinate system constraints as inequality (4), and one of ambiguous rotation matrix can be expressed as R:

$$R = [Axis_x \ Axis_y \ Axis_z] = \begin{bmatrix} n_{21} & n_{11} & n_{31} \\ n_{22} & n_{12} & n_{32} \\ n_{23} & n_{13} & n_{33} \end{bmatrix} \quad (3)$$

$$(Axis_x \times Axis_y) \cdot Axis_z > 0. \quad (4)$$

Translation vector of point cloud coordinate system relative to camera coordinate system can be obtained by calculating the point cloud centroid.

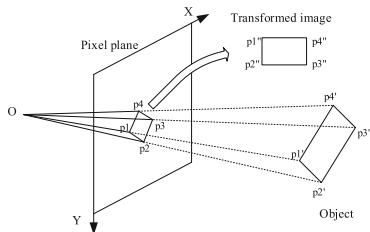
$$t = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} = \begin{bmatrix} p_{1x} + p_{2x} + \dots + p_{nx} \\ p_{1y} + p_{2y} + \dots + p_{ny} \\ p_{1z} + p_{2z} + \dots + p_{nz} \end{bmatrix} / n \quad (5)$$

## 2.2 Image Transformation Using Homography

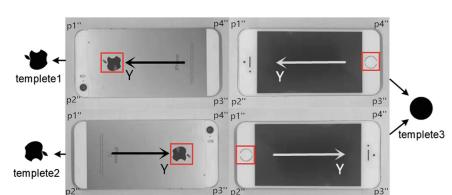
In order to find the correct pose from the 4 ambiguous poses, we use template matching algorithm based on 2D shape to identify the front and back side to determine the positive z-axis direction and the phone head direction to determine the positive y-axis direction. Considering perspective projection in different viewpoints, we first transform the image to the top view of the 3C component to avoid the image distortion (Fig. 6).

Preprocessing for input images mainly includes edge extraction and morphological closed operation and quadrilateral fitting. Then, we get four vertex coordinates of the quadrilateral in pixel coordinate system and then sort the vertices according to certain rules so that they correspond to the order of the pixel coordinates we want to transform to. In this paper, we take the starting point of the short side with the smallest x coordinate in pixel coordinate system as the first point, and sort four vertices in counterclockwise order.

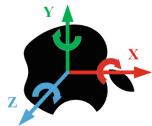
Finally, we calculate the homography matrix and transform the image to top view (shown in Fig. 3). By image transformation, the scale and angle of the object are controllable, which makes template matching faster and more accurate.



**Fig. 3.** Image transformation



**Fig. 4.** Transformed images and corresponding templates



**Fig. 5.** Template transformation



**Fig. 6.** Templates feature points

### 2.3 Shape-Based Template Matching

After image transformation, there are 4 cases of the transformed images. We selected 3 easily distinguishable shapes on images as templates to match transformed image. The corresponding relationship between the transformed images and the templates is shown in Fig. 4.

According to Hinterstoisser [13], Linemod algorithm uses RGB images and depth images to perform 6D pose estimation for textureless objects. However, we found that Linemod algorithm does not perform well for the 3C component with symmetrical planar structures through experiments because surface normal features are similar for planar object. We deprecated depth images and only used gray images to calculate gradient features.

**Template Feature Extraction.** First, we perform a small-angle perspective transformation on the template picture to improve the robustness of template matching. In this paper, the angle is between  $-5^\circ$  to  $5^\circ$  with step =  $1^\circ$  along X/Y/Z axis as Fig. 5. Second, the template pictures are used to calculate the gradient direction which is specified within  $0\text{--}180^\circ$  and quantized into 8 directions [14]. We use the same rules as linemod algorithm to filter out the feature points of the template. All the template information including template id, template features and template image sizes is written into template file for shape matching.

**Template Matching.** Template matching process mainly includes computing Response Maps and Similarity Maps [14]. The computation of the Response Maps is almost similar to the gradient calculation of the template images, except that Spreading the Gradient Orientations is used to increase the robustness of feature matching. For each test images, a sliding window method is used to obtain a region of test images and the similarity between the region and the template is computed. Each match contains the template id, similarity score, and matching coordinates  $(x, y)$ . The match with the highest similarity score is used to eliminate ambiguous poses.

## 2.4 Determination of Real Pose

**Determination of Z-axis Direction:** Template1 and template2 correspond to back side and template3 corresponds to front side, which is used to determine positive Z-axis direction. We use  $zflag = 1$  to represent the front side and  $zflag = -1$  to represent the back side. Thus, we have following formula in which  $Axis_z$  is vector of Z-axis,  $n_{23}$  is z component of  $Axis_z$ .

$$Axis_z = \begin{cases} Axis_z & \text{if } zflag \times n_{23} < 0 \\ -Axis_z & \text{otherwise} \end{cases} \quad (6)$$

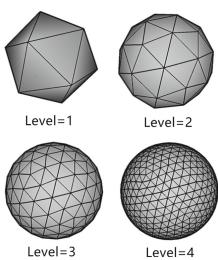
**Determination of Y-axis direction:** The corresponding relationship between the Y-axis direction and the template is shown in Fig. 4. Y-axis direction is represented by a vector calculated from vertex coordinates. For example, if template 1 matched, the Y-axis direction is represented as the vector from point  $p4'$  to point  $p1'$ ; if template 2 matched, the Y-axis direction is represented as the vector from point  $p1'$  to point  $p4'$ . Specially, when template3 matched, the coordinates matched are used for further judgment. If coordinates matched is close to  $p4''$ , the Y-axis direction is represented as the vector from point  $p4'$  to point  $p1'$ ; otherwise, the Y-axis direction is represented as the vector from point  $p1'$  to point  $p4'$ . The Y-axis direction is determined by the following formula, in which  $Axis_{ytrue}$  is the real Y-axis direction. Finally, we find the correct pose.

$$Axis_y = \begin{cases} Axis_y & \text{if } Axis_y \cdot Axis_{ytrue} > 0 \\ -Axis_y & \text{otherwise} \end{cases} \quad (7)$$

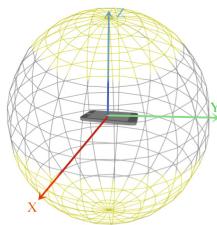
### 3 Simulation and Experiments

#### 3.1 Simulation and Experiments Setup

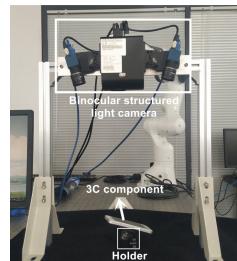
In simulation, first step is to render the point clouds and images data from different viewpoints for algorithm testing. We use icosahedron [15] to generate different viewpoints. Every viewpoint is set in the center of the every polygon. As shown in Fig. 7, the higher the subdivision level of the polygon, the closer the icosahedron is to the spherical surface. In order to keep enough test samples, we set the subdivision level to 5 and get 5120 viewpoints. The elevation angle of the viewpoints relative to the model coordinate system is set to 40–90° (Fig. 8) to eliminate extreme view according to our actual assembly requirement. Finally, 1827 different viewpoints are used to render the point clouds and images for test.



**Fig. 7.** Icosahedron subdivision level



**Fig. 8.** Reasonable area of viewpoints



**Fig. 9.** Experiment platform

Our experiment platform is shown as Fig. 9. We use mobile phone model as our pose estimation target; a phone holder is used to change the pose of the 3C component; a binocular structured light camera is used to capture point clouds and gray images. In order to simplify background processing, we use a black flannel as the background.

#### 3.2 Simulation Design and Result

**Simulation 1: Accuracy Test.** In order to compare our method with the SAC-IA and Chen's method [11], we designed the contrast experiment. We sample the CAD model into a point cloud, which contains about 20000 points, and sample the rendered point cloud to 8000–12000 points. Then the above three methods are used to estimate the initial pose, and the accuracy and runtime is shown in Table 1. As shown in Table 1, the accuracy of the SAC-IA method is very low because the FPFH descriptor causes a lot of mismatches, and the verification stage takes a lot of time. Due to the high symmetry of 3C components, the accuracy of Chen's method [11] is also low. By using 2D shape information, our method achieves higher accuracy and reasonable runtime.

**Table 1.** Simulation results of different algorithms

Algorithm	Accuracy(%)	Runtime(s)
SAC-IA	12.1	12.60
Chen's method [11]	24.8	0.42
Our method	100.0	0.45

**Simulation 2: Robustness Test for Working Distance.** This simulation is to compare the performance of our shape-based template matching algorithm and linemod algorithm to eliminate ambiguous pose. First, we render 1827 images for each rendering radius  $R$  between 300 mm to 305 mm with step = 1 mm. For the linemod algorithm, we use the rendered images ( $R = 300$  mm) to make templates, and for our method, we use the template images mentioned above. Then, we randomly selected 10% of images each whose  $R$  between 301 mm to 305 mm for testing. The simulation results are shown in the Table 2; as we can see, linemod template matching algorithm is sensitive to distance changes and its accuracy decreases as the rendering radius error between the test images and the template images increases. When the rendering radius error is bigger than 5 mm, the accuracy rate is only about 60%. However, Our method improves the robustness of template matching by perspective transformation of images.

**Table 2.** Comparison of template matching algorithms

Rendering radius	Accuracy(%)		Times per match(s)	
	Linemod	Our method	Linemod	Our method
301 mm	95.6	100.0	0.76	0.39
302 mm	91.6	100.0	0.75	0.38
303 mm	74.3	100.0	0.75	0.39
304 mm	65.9	100.0	0.73	0.38
305 mm	64.7	100.0	0.72	0.38

### 3.3 Experiments Design and Result

In the experiment, we use a mobile phone holder to change the pose of the mobile phone cover 30 times; the elevation angle is between  $40^\circ$  to  $80^\circ$ . Then, we artificially judge the correctness based on the captured point cloud and the calculated pose. The number of point clouds captured is about 3 million and we downsample to 8000–12000 points as simulation, which cost extra time. The experimental results are shown as Table 3 which is similar to simulations and proves the correctness of our algorithm. The experiment process has been shown in the supplementary video.

**Table 3.** Experiments results of different algorithms

Algorithm	Accuracy(%)	Runtime(s)
SAC-IA	16.7	13.20
Chen's method [11]	26.7	0.81
Our method	100.0	0.82

## 4 Conclusion

In this paper, we first define the pose ambiguity problem in 6D pose estimation of 3C components based on point clouds. Then, we propose an algorithm to solve the pose ambiguity problem, which combines PCA and shape-based template matching. Our algorithm uses 2D image information and improve the accuracy of 3C components pose estimation to 100%, which is much higher than other traditional algorithms. The results of simulation and experiment evidence the correctness of our algorithm. In the future, we will improve the adaptability of the algorithm and apply our algorithm on other 3C components such as PCB and battery of cellphone.

**Acknowledgment.** This work was supported partially by the NSFC-Shenzhen Robotics Basic Research Center Program (No. U1713202) and partially by the Shenzhen Science and Technology Program (No. JCYJ20180508152226630).

## References

1. Manhardt, F., Arroyo, D.M., Rupprecht, C., Busam, B., Birdal, T., Navab, N., Tombari, F.: Explaining the ambiguity of object detection and 6d pose from visual data. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6841–6850 (2019)
2. Hodaň, T., Matas, J., Obdržálek, Š.: On evaluation of 6D object pose estimation. In: European Conference on Computer Vision, pp. 606–619. Springer, Heidelberg (2016)
3. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: Densefusion: 6D object pose estimation by iterative dense fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3343–3352 (2019)
4. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: PVNet: pixel-wise voting network for 6D of pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4561–4570 (2019)
5. Chen, Y., Medioni, G.: Object modelling by registration of multiple range images. *Image Vis. Comput.* **10**(3), 145–155 (1992)
6. Besl, P.J., McKay, N.D.: Method for registration of 3-D shapes. In: Sensor Fusion IV: Control Paradigms and Data Structures, vol. 1611. International Society for Optics and Photonics, pp. 586–606 (1992)
7. Rusu, R.B., Bradski, G., Thibaux, R., Hsu, J.: Fast 3D recognition and pose using the viewpoint feature histogram. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2155–2162. IEEE (2010)

8. Yang, J., Li, H., Campbell, D., Jia, Y.: Go-ICP: a globally optimal solution to 3D ICP point-set registration. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(11), 2241–2254 (2015)
9. Yoo, H.W., Kim, W.H., Park, J.W., Lee, W.H., Chung, M.J.: Real-time plane detection based on depth map from kinect. In: *IEEE ISR 2013*, pp. 1–4, October 2013
10. Rusu, R.B., Blodow, N., Beetz, M.: Fast point feature histograms (FPFH) for 3D registration. In: *2009 IEEE International Conference on Robotics and Automation*, pp. 3212–3217. IEEE (2009)
11. Chen, X., Chen, C.-H.: Model-based point cloud alignment with principle component analysis for robot welding. In: *2017 International Conference on Advanced Robotics and Intelligent Systems (ARIS)*, pp. 83–87. IEEE (2017)
12. Dong, D., Yang, X., Hu, H., Lou, Y.: Pose estimation of components in 3c products based on point cloud registration. In: *IEEE International Conference on Robotics and Biomimetics*, pp. 339–344 (2019)
13. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In: *Asian Conference on Computer Vision*, pp. 548–562. Springer, Heidelberg (2012)
14. Hinterstoisser, S., Cagniart, C., Ilic, S., Sturm, P., Navab, N., Fua, P., Lepetit, V.: Gradient response maps for real-time detection of textureless objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(5), 876–888 (2011)
15. Hinterstoisser, S., Benhimane, S., Lepetit, V., Fua, P., Navab, N.: Simultaneous recognition and homography extraction of local patches with a simple linear classifier. In: *BMVC*, pp. 1–10 (2008)



# Manipulator Control Law Design Based on Backstepping and ADRC Methods

Lijun Wang<sup>1,2(✉)</sup>, Jiaxuan Yan<sup>1,2</sup>, Tianyu Cao<sup>1,2</sup>, and Ningxi Liu<sup>1,2</sup>

<sup>1</sup> School of Automation and Electrical Engineering,  
University of Science and Technology Beijing, Beijing 100083, China  
[wanglj@ustb.edu.cn](mailto:wanglj@ustb.edu.cn)

<sup>2</sup> Key Laboratory of Knowledge Automation for Industrial Processes,  
Ministry of Education, Beijing 100083, China

**Abstract.** A backstepping-ADRC control law is designed for one-degree of freedom (DOF) link manipulator system with internal and external factor uncertainty. First, the dynamic model of the manipulator is established, and the extended state observer of ADRC is used to estimate the total disturbance. Second, in order to improve the tracking accuracy of position signal, the nonlinear error feedback control law is improved by combining backstepping control method. Finally, the control law designed is simulated and compared. The effectiveness of the proposed method is varied by the simulation.

**Keywords:** Manipulator · ADRC · Tracking · Position signal · Backstepping

## 1 Introduction

In recent years, manipulator is increasingly used for carrying and repetitive tasks [1]. The dynamic model of the manipulator is a complex time-varying system. And it has the characteristics of strong coupling. In addition, the established system model is imprecise and affected by the surrounding environment and self friction [2]. For this highly nonlinear system, it needs to meet the requirements of protecting itself from damage, ensuring human safety and avoiding collision with other objects [3]. It is very complicated to design an optimal control law. How to eliminate the internal and external factors of the manipulator system is the key to the control law design [4].

In order to obtain accurate trajectory tracking and good performance, the current control methods mainly include PID control [5], sliding mode control [6,7], singular perturbation control [8], adaptive control [9,10], iterative learning control [11,12] and so on. In [3], PID is used to increase the stability of the manipulator by reducing the input error. Because its control method is simple and practical, it is widely used in the early stage. However, with the upgrading of the system, more and more control accuracy is sought. However, due to the limitation of PID control's own characteristics, the problem of over harmonic

system's slow response often occurs. In [5], sliding mode control is used to control the manipulator. When controlling the position tracking of the manipulator, the anti-interference ability is very strong. But at the same time, the problem is that the chattering phenomenon often occurs when the system switches at high speed and works for a long time. In [8], an adaptive control manipulator is designed. However, it is difficult to guarantee the stability of the system for the manipulator which is an imprecise model. In [11], the iterative learning method is that the machine can achieve effective control through its own learning. It is very effective for the complicated model to determine the manipulator of the object. However, iterative learning only pays attention to convergence, and the researches on the quality of system convergence speed are not enough.

The technology of ADRC is a new control method of Engineering nonlinear system proposed by researcher Han Jingqing [13], its advantage is that the precise model of the controlled object can not be needed, and it is more suitable for the manipulator which is difficult to establish the model. The backstepping method was proposed by kanella kopoulos and kokotovic in 1990 [14], Melhem and Wang obtained the global asymptotic stability of the flexible joint robot using the backstepping method [15].

This paper aims at the trajectory tracking control of manipulator under the influence of uncertain factors such as unmodeled dynamics, friction moment and external interference. The internal uncertainty and external disturbance of the system are regarded as the total disturbance by using the ADRC method, and the total disturbance that affects the tracking accuracy of the system position signal can be cancelled by ADRC [16]. First, the basic structure of ADRC is introduced. Then, based on backstepping and ADRC methods, backstepping-ADRC control law is proposed. Finally, the effectiveness of the control method is verified by simulation results.

## 2 System Model

The nonlinear mathematical model of the one-degree of freedom (DOF) link manipulator with the internal and external uncertainty parts is expressed as

$$M_0 \ddot{\theta} + V_0 \dot{\theta} + G_0 = \tau + d_t \quad (1)$$

where  $\theta \in R^n$  is the output variable of the single joint manipulator model,  $M_0 \in R^n$  is the moment of inertia of the manipulator,  $V_0 \in R^n$  is the friction coefficient of the manipulator,  $G_0 \in R^n$  is the term of gravity acting on the arm,  $d_t \in R^n$  is uncertain factors such as external interference and friction factor of internal manipulator, and  $\tau \in R^n$  is the input vector of the manipulator.

According to Eq. (1), the expressions for  $M_0$  and  $V_0$  are

$$M_0 = 4ml^2 \quad (2)$$

where the  $m$  is the mass of the manipulator,  $l$  is the distance from the mass center of the manipulator to the rotation center of the manipulator.

$$V_0 = mgl \cos \theta \quad (3)$$

where the  $g$  is the acceleration of gravity.

Define  $x_1 = \theta, x_2 = \dot{\theta}, u = \tau$ , and Eq. (1) can be rewritten as

$$f_1(x_1, x_2) = x_2 \quad (4)$$

$$f_2(\bar{x}_2, u) = -\frac{3C_0}{4ml^2}x_2 - \frac{3g}{4l} \cos x_1 + \frac{3}{ml^2}u + \frac{3}{ml^2}d_t \quad (5)$$

where  $\bar{x}_2 = [x_1, x_2]^T$ , and the  $d_t$  can be regarded as the total disturbance of the manipulator system,  $y$  is the output variable of the manipulator system, and  $u$  is the input control variable of the manipulator system.

Then, in order to make the manipulator model better applied in ADRC control algorithm, and we can get the reequation as

$$\begin{cases} \dot{x}_1 = f(x_1, x_2) \\ \dot{x}_2 = f(\bar{x}_2, u) \\ y = x_1 \end{cases} \quad (6)$$

### 3 Controller Design Process

#### 3.1 Basic Structure of Controller

Due to the influence of measurement error and external disturbance, it is very difficult to get an accurate model of manipulator, and it is difficult to meet the requirement of high precision tracking signal with traditional controller. In order to solve the problems, the tracking accuracy of the system is improved by the ADRC nonlinear control law.

The designed controller consists of three parts. The first part is tracking differentiator, it arranges signal transition process to extract differential signal of original signal. The second part is the extended state observer, it estimates the system state according to the input and output signals of the controlled object. The third part is the backstepping-ADRC control law, it's used to reduce position tracking error of the system.

#### 3.2 Control Law Design of ADRC

**Tracking Differentiator.** Tracking differentiator (TD) is used to extract differential signal from original signal or signal disturbed by noise. Excellent differential signal directly affects the function of the whole controller. A fast tracking differentiator is used as follows

$$\begin{cases} \dot{v}_1 = v_2 \\ \dot{v}_2 = fhan(v_1 - v_{in}, v_2, r, h_0) \end{cases} \quad (7)$$

where the function  $fhan$  is

$$fhan(e, x_2, r, h) = \begin{cases} d = rh \\ d_0 = hd \\ y = e + hx_2 \\ a_0 = \sqrt{d^2 + 8r|y|} \\ a = \begin{cases} x_2 + \frac{a_0 - d}{2} sign(y), & |y| > d_0 \\ x_2 + \frac{y}{h}, & |y| \leq d_0 \end{cases} \\ fhan = \begin{cases} r sign(a), & |a| > d \\ r \frac{a}{d}, & |a| \leq d \end{cases} \end{cases} \quad (8)$$

where  $v_1, v_2$  is output signal,  $v_{in}$  is input signal, parameter  $r$  is the fast factor, and parameter  $h_0$  is the integral step length.

**Extended State Observer.** The extended state observer (ESO) is used to observe the total disturbance of the whole manipulator system. And promptly eliminate the impact of the disturbance. The form of the extended state observer is as follows

$$\begin{cases} e = z_1 - y \\ \dot{z}_1 = z_2 - \beta_1 e \\ \dot{z}_2 = z_3 - \beta_2 fal(e, 0.5, \sigma) + bu \\ \dot{z}_3 = -\beta_3 fal(e, 0.25, \sigma) \end{cases} \quad (9)$$

where the function of  $fal$  is

$$fal(e, \alpha, \delta) = \begin{cases} |e|^\alpha sign(e) & |e| > \delta \\ e/\delta^{1-\alpha} & |e| \leq \delta \end{cases} \quad (10)$$

where the  $\beta_1, \beta_2, \beta_3$  are constants that need to be adjusted,  $\alpha, \delta$  are related to the shape of the equation  $fal$ .  $z_1, z_2$  is the estimated value of the output signal and its differential signal of the controlled object,  $z_3$  is the estimate of the total disturbance of the system,  $e$  is the position signal error, and  $b$  is a constant determined by the characteristics of the system.

**Nonlinear Error Control Law.** The combination of position signal error and velocity signal error is used in nonlinear error control law. The nonlinear error control law is used to reduce the output signal error and ensure the control accuracy. In ADRC, the most commonly used control law is as follows

$$u = k_1 fal(e_1, \alpha_1, \sigma_1) + k_2 fal(e_2, \alpha_2, \sigma_2) - z_3/b \quad (11)$$

where the  $fal$  is as the Eq. (10),  $e_1 = v_1 - z_1$  is position signal error,  $e_2 = v_2 - z_2$  is differential signal error,  $k_1$  and  $k_2$  are adjustable parameters.

### 3.3 Control Law Design of Backstepping-ADRC

Based on the nonlinear error control law in ADRC, the backstepping method is used to improve the nonlinear error control law.

In this part, based on backstepping and ADRC methods, a new control law of backstepping-ADRC is proposed. The purpose is to make input signal tracked by output signal better.

Define position signal tracking error  $e_1$  as

$$e_1 = x_1 - z_d \quad (12)$$

where  $z_d$  is the position instruction signal. The derivative of Eq. (12) is

$$\dot{e}_1 = \dot{x}_1 - \dot{z}_d = x_1 + x_2 - \dot{z}_d \quad (13)$$

Define the first Lyapunov function as

$$V_1 = \frac{1}{2} e_1^2 \quad (14)$$

The Eq. (14) derivative is

$$\dot{V}_1 = e_1 \dot{e}_1 = e_1(x_2 - \dot{z}_d) \quad (15)$$

Let

$$x_2 = \dot{z}_d - c_1 e_1 + e_2 \quad (16)$$

where  $c_1 > 0$ ,  $z_2$ , and  $e_2$  is the virtual control quantity.

And from the Eq. (16), we can obtain

$$e_2 = x_2 + c_1 e_1 - \dot{z}_d \quad (17)$$

The derivative of Eq. (15) is

$$\dot{V}_1 = -c_1 e_1^2 + e_1 e_2 \quad (18)$$

If  $e_2 = 0$ ,  $\dot{V}_1 \leq 0$ , we need to design the next step.

Define the second Lyapunov function

$$V_2 = V_1 + \frac{1}{2} e_2^2 \quad (19)$$

The derivative of Eq. (17) is

$$\dot{e}_2 = -\frac{3C_0}{4ml^2} x_2 - \frac{3g}{4l} \cos x_1 + \frac{3}{ml^2} u + \frac{3}{ml^2} d_t + c_1 \dot{e}_1 - \ddot{z}_d \quad (20)$$

And we can obtain

$$\begin{aligned} \dot{V}_2 &= \dot{V}_1 + e_2 \dot{e}_2 \\ &= -c_1 e_1^2 + e_1 e_2 + e_2 \left( -\frac{3C_0}{4ml^2} x_2 - \frac{3g}{4l} \cos x_1 \right. \\ &\quad \left. + \frac{3}{ml^2} u + \frac{3}{ml^2} d_t + c_1 \dot{e}_1 - \ddot{z}_d \right) \end{aligned} \quad (21)$$

In order to make  $\dot{V}_2 \leq 0$ , the design control law is

$$u = \frac{ml^2}{3} \left( \frac{3C_0}{4ml^2} x_2 + \frac{3g}{4l} \cos x_1 - \frac{3}{ml^2} d_t - c_1 \dot{e}_1 + \ddot{z}_d - e_1 - c_2 e_2 \right) \quad (22)$$

Let

$$\begin{aligned} f(x) &= \left( \frac{3C_0}{4ml^2} x_2 + \frac{3g}{4l} \cos x_1 - \frac{3}{ml^2} d_t \right) \\ b &= \frac{ml^2}{3} \end{aligned}$$

And we can obtain

$$u = b(f(x, t) - c_1 \dot{e}_1 + \ddot{z}_d - e_1 - c_2 e_2) \quad (23)$$

where the  $c_1, c_2$  are bigger than 0.

Then we can obtain

$$\dot{V}_2 = -c_1 e_1^2 - c_2 e_2^2 \leq 0 \quad (24)$$

For Eq. (17), where  $x_2 = \dot{x}_1$ , it is the output item of the system and  $\ddot{z}_d$  is the expected trajectory of the system. So the Eq. (17) can be expressed as

$$e_2 = \dot{x}_1 - \dot{z}_d + c_1 e_1 = \dot{e}_1 + c_1 e_1 \quad (25)$$

And we can obtain

$$u = b(f(x) - (c_1 c_2 - 1)e_1 + (c_2 - 1)\dot{e}_1 + \ddot{z}_d - f(x) + \ddot{z}_d) \quad (26)$$

$$\begin{cases} k_p = c_1 c_2 - 1 \\ k_d = c_2 - 1 \end{cases} \quad (27)$$

$$u = b(k_p e_1 + k_d \dot{e}_1 - f(x) + \ddot{z}_d) \quad (28)$$

Due to the first two terms of Eq. (28) are  $k_p e_1 + k_d \dot{e}_1$ , the form is similar to that of Eq. (11), and  $b$  can be regarded as a constant. Then, we can combine Eq. (28) with Eq. (11) to form backstepping-ADRC control law.

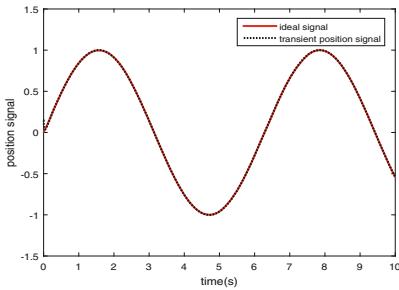
$$u = b(k_1 fal(e_1, \alpha_1, \sigma_1) + k_2 fal(e_2, \alpha_2, \sigma_2) - z_3/b - f(x) + \ddot{z}_d) \quad (29)$$

## 4 Simulation Results

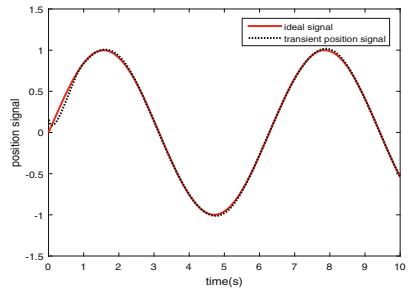
In order to verify the effectiveness of the proposed control law, the system simulation model is established under the Matlab/Simlink environment, and the designed controller is simulated. In the simulation, the object parameters are selected as:  $C_0 = 2.8 \text{ N.m.s/rand}$ ,  $m = 1.20 \text{ kg}$ ,  $l = 0.45 \text{ m}$ ,  $g = 9.8 \text{ m/s}^2$ ,  $d_t = x_2 \sin(x_1)$ , The initial value of the system  $x_1(0) = x_2(0) = [0.15, 0]^T$ ,

The parameters of the controller are  $b = 1/133, r = 1000, h = 0.01, \beta_1 = 100, \beta_2 = 300, \beta_3 = 1000, kp = 80, kd = 15$ .

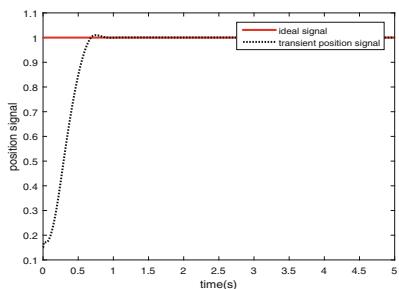
In order to compare the effect of the backstepping-ADRC control law designed in this paper, the traditional ADRC control law is taken as the contrast object. Sinusoidal and step signals are used to compare the advantages of the two control laws in the manipulator system. When the input is sinusoidal or step signal, it can be observed that the position tracking curve of backstepping-ADRC control law in Fig. 1 is better than that of ADRC control law in Fig. 2, and Fig. 3 is better than Fig. 4 in position tracking. Figure 5 and Fig. 6 are the position tracking error curves of the two control laws. As shown in Fig. 1, 2, 3, 4, 5 and 6, the results show that the control law designed in this paper has advantage in position signal tracking.



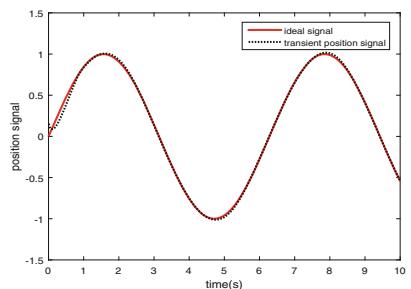
**Fig. 1.** Position tracking of backstepping-ADRC controller



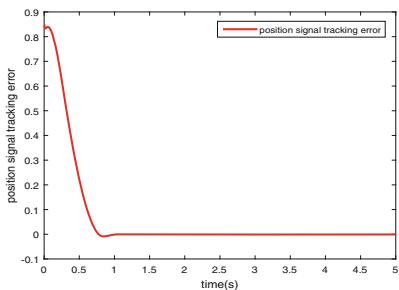
**Fig. 2.** Position tracking of ADRC controller



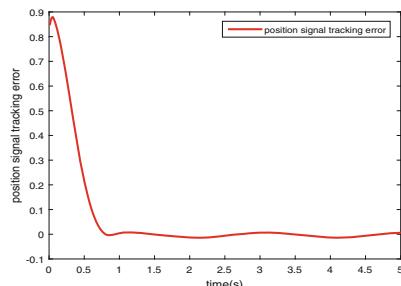
**Fig. 3.** Position tracking of backstepping-ADRC controller



**Fig. 4.** Position tracking of ADRC controller



**Fig. 5.** Position tracking error of backstepping-ADRC controller



**Fig. 6.** Position tracking error of ADRC controller

## 5 Conclusion

The innovation of this paper is that a method of ADRC and backstepping is designed to solve the influence of uncertain factors in the control model of manipulator. Aiming at improving the tracking accuracy of position signal in the manipulator system, an backstepping-ADRC control law is proposed. Based on this control law, the simulation experiment is done. By comparing the output curve of sinusoidal signal and step signal, the results show that the tracking error of position signal can be controlled in a very small range when considering the internal uncertainty and external interference. On the basis of this work, because the selected model is simpler, the usability in the complex multi DOF Manipulator needs to be further verified.

## References

1. Fan, X.U., Wang, J., Guodong, L.U.: Adaptive robust neural control of a two-manipulator system holding a rigid object with inaccurate base frame parameters. *J. Zhejiang Univ. Sci. C* **19**(11), 1316–1327 (2018)
2. Izadbakhsh, A., Khorashadizadeh, S.: Robust adaptive control of robot manipulators using Bernstein polynomials as universal approximator. *Int. J. Robust Non-linear Control* (2020). <https://doi.org/10.1002/rnc.4913>
3. Zhang, S., Dai, S., Zanchettin, A.M., Villa, R.: Trajectory planning based on non-convex global optimization for serial manipulators. *Appl. Math. Model.* (2020). <https://doi.org/10.1016/j.apm.2020.03.004>
4. Jun, J., Padois, V., Benamar, F.: Stability-based planning and trajectory tracking of a mobile manipulator over uneven terrains. In: *IEEE International Workshop on Advanced Robotics and its Social Impacts* (2015). <https://doi.org/10.1109/ARSO.2015.7428212>
5. Alvarezramirez, J., Cervantes, I., Kelly, R.: PID regulation of robot manipulators: stability and performance. *Syst. Control Lett.* **41**(2), 73–83 (2000)
6. Islam, S., Liu, X.P.: Robust sliding mode control for robot manipulators. *IEEE Trans. Ind. Electron.* **58**(6), 2444–2453 (2011)

7. Kali, Y., Saad, M., Benjelloun, K.: Control of robot manipulators using modified backstepping sliding mode. In: New Bioprocessing Strategies: Development and Manufacturing of Recombinant Antibodies and Proteins (2019). [https://doi.org/10.1007/978-981-13-2212-9\\_5](https://doi.org/10.1007/978-981-13-2212-9_5)
8. Guang, L., Min, W.: Singular perturbation approach for control of hydraulically driven flexible manipulator. *J. Central South Univ. Technol.* **12**(1), 238–242 (2005)
9. Qiu, Z., Wang, B., Zhang, X., Han, J.: Direct adaptive fuzzy control of a translating piezoelectric flexible manipulator driven by a pneumatic rodless cylinder. *Mech. Syst. Sig. Process.* **36**(2), 290–316 (2013)
10. Parlaktuna, O., Ozkan, M.: Adaptive control of free-floating space manipulators using dynamically equivalent manipulator model. *Robot. Auton. Syst.* **46**(3), 185–193 (2004)
11. Jiang, P., Woo, P., Unbehauen, R.: Iterative learning control for manipulator trajectory tracking without any control singularity. *Robotica* **20**(2), 149–158 (2002)
12. Jia, B., Liu, S., Liu, Y.: Visual trajectory tracking of industrial manipulator with iterative learning control. *Ind. Robot Int. J.* **42**(1), 54–63 (2015)
13. Han, J.Q.: From PID to active disturbance rejection control. *IEEE Trans. Ind. Electron.* **56**(3), 900–906 (2009)
14. Kanellakopoulos, I., Kokotovic, P.V., Morse, A.S.: Systematic design of adaptive controllers for feedback linearizable systems. In: American Control Conference (1991). <https://doi.org/10.1109/9.100933>
15. Melhem, K., Wang, W.: Global output tracking control of flexible joint robots via factorization of the manipulator mass matrix. *IEEE Trans. Robot.* **25**(2), 428–437 (2009)
16. Liu, B., Hong, J., Wang, L.: Linear inverted pendulum control based on improved ADRC. *Syst. Sci. Control Eng.* **7**(3), 1–12 (2019)



# Manipulator Calibration-Free Hand-Eye Coordination Based on ADRC Under Eye Fixation

Lijun Wang<sup>1,2(✉)</sup>, Tianyu Cao<sup>1,2</sup>, Jiaxuan Yan<sup>1,2</sup>, and Ningxi Liu<sup>1,2</sup>

<sup>1</sup> School of Automation and Electrical Engineering,  
University of Science and Technology Beijing, Beijing 100083, China  
[wanglj@ustb.edu.cn](mailto:wanglj@ustb.edu.cn)

<sup>2</sup> Key Laboratory of Knowledge Automation for Industrial Processes,  
Ministry of Education, Beijing 100083, China

**Abstract.** The forward kinematics of the manipulator and the Jacobian matrix of the manipulator are analysed and the D-H (Denavit-Hartenberg) method is used in manipulator modeling. The nonlinear relationship between the manipulator and the camera system is analyzed, and it is built into the form of ADRC (Active Disturbances Rejection Controller). The ESO (extended state observer) in ADRC is used to compensate the inaccuracy and unknown interference of system modeling, and realize the hand-eye coordinated control of the manipulator under the eye fixation. The feasibility and effectiveness of this method are verified by the simulation.

**Keywords:** ADRC · Hand-eye coordination · Calibration-free · Jacobian matrix

## 1 Introduction

Manipulator hand-eye coordination generally refers to the visual positioning and tracking of the manipulator. With the development of intelligent robots, research on hand-eye coordination has received increasing attention [1]. Traditional robot hand-eye coordination often uses a model-based control method, which is based on the accurate calibration of the hand-eye relationship model. This calibration method is essentially open-loop static control, and its performance is closely related to the accuracy of the calibration [2]. It is difficult to achieve in many practical application environments. Manipulator calibrated-free hand-eye coordination has great research potential [3,4].

The manipulator hand-eye coordination system is essentially in the case of the manipulator's hand-eye relationship is unknown, using the current visual feedback information to plan the robot path, and finally complete the target positioning and dynamic tracking tasks. The method based on the image Jacobian matrix is a more classic method [5–7]. It estimates the image Jacobian

matrix in real time online to plan the robot motion at the next moment, and completes the hand-eye coordination task finally.

At present, there has been a lot of research on the online estimation method of image Jacobian [8,9]. References [10,11] use artificial neural networks to approximate the inverse image Jacobian matrix, but this method requires offline training of a large number of sample points in the robot's motion space in advance, which seriously limits the practicability of the method. Literature [12] applied the idea of auto disturbance rejection controller to the field of hand-eye coordination of non-calibrated robots, but the designed coupled controller increases the complexity and increases the delay of the system.

In this paper, ADRC technology is used to achieve hand-eye coordination of ABBIRB2400 manipulator without calibration. First, the nonlinear mapping relationship between the opponent and camera is estimated and using the rough estimate as the system model. Then using ESO to estimate and compensate for the existing modeling error, dynamic uncertainty and unknown external interference in real time. Finally, the feedback gives the movement amount of the robot at the next moment, so as to realize the hand-eye coordinated control of the manipulator without calibration. In addition, this article removes the TD module and enhances the rapid performance of the system.

## 2 Problem Formulation

For six-degree-of-freedom manipulator that all joints are rotating joints, a world coordinate system can be established with the base as the origin. Assuming the position of the end of the manipulator is  $W = (x_w, y_w, z_w)^T$ , then the forward kinematics of the manipulator can be expressed as

$$W = K(q) \quad (1)$$

$q$ , is the rotation angle of joint. Derivatives on both sides

$$V = J_r \cdot \dot{q} \quad (2)$$

$V = \dot{W}$ , is a space translation velocity;  $J_r$  is the manipulator Jacobian matrix, which can be obtained by the vector product method or the differential transformation method. In this paper, the differential transformation method is used to obtain the robot arm Jacobian matrix  $J_r$ . Thus, the relationship between the moving speed of the end effector of the manipulator and the joint angle is obtained.

In the case of eye fixation, the coordinate of the end of the manipulator in the robot world coordinate system is  $W = (x_w, y_w, z_w)^T$ , and the coordinate in the image is  $P = (p_x, p_y)^T$ , then the mapping relationship between the target's image coordinates and three-dimensional space coordinates is

$$P = G(W) \cdot W \quad (3)$$

Differentiate both sides

$$\dot{P} = J_P \cdot U \quad (4)$$

$U = \dot{W}$ , hand grasping motion vector in robot coordinate system;  $J_P$ , image Jacobian matrix.

$$J_p = \begin{bmatrix} J_{p_x} \\ J_{p_y} \end{bmatrix} = \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \end{bmatrix} \quad (5)$$

From (2) and (3), we have

$$\begin{bmatrix} \dot{p}_x \\ \dot{p}_y \end{bmatrix} = \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \end{bmatrix} \cdot \begin{bmatrix} u_x \\ u_y \\ u_z \end{bmatrix} \quad (6)$$

that is

$$\begin{cases} \dot{p}_x = J_{11} \cdot u_x + J_{12} \cdot u_y + J_{13} \cdot u_z \\ \dot{p}_y = J_{21} \cdot u_x + J_{22} \cdot u_y + J_{23} \cdot u_z \end{cases} \quad (7)$$

### 3 Controller Design

For the two-dimensional XY plane, the vector of motion in the Z direction is 0, then

$$\begin{cases} \dot{p}_x = J_{11} \cdot u_x + J_{12} \cdot u_y \\ \dot{p}_y = J_{21} \cdot u_x + J_{22} \cdot u_y \end{cases} \quad (8)$$

Make the robot do the initial trial motion to get the initial estimate of the image comparable matrix  $\hat{J}_p$

$$\hat{J}_p = \begin{bmatrix} \hat{J}_{11} & \hat{J}_{12} \\ \hat{J}_{21} & \hat{J}_{22} \end{bmatrix}$$

Considering the errors caused by modeling, image detection, and external perturbations that introduce system errors  $\xi_1$  and  $\xi_2$ , Eq. (6) can be changed to ADRC's equation sub-model

$$\begin{cases} \dot{p}_x = (J_{11} - \hat{J}_{11}) \cdot u_x + J_{12} \cdot u_y + \xi_1 + \hat{J}_{11} \cdot u_x \\ \dot{p}_y = J_{21} \cdot u_x + (J_{22} - \hat{J}_{22}) \cdot u_y + \xi_2 + \hat{J}_{22} \cdot u_y \end{cases} \quad (9)$$

Suppose:

$$\begin{cases} \alpha_x = (J_{11} - \hat{J}_{11}) \cdot u_x + J_{12} \cdot u_y + \xi_1 \\ \alpha_y = J_{21} \cdot u_x + (J_{22} - \hat{J}_{22}) \cdot u_x + \xi_2 \end{cases} \quad (10)$$

Then (7) can be written as

$$\begin{cases} \dot{p}_x = \alpha_x + \hat{J}_{11} \cdot u_x \\ \dot{p}_y = \alpha_y + \hat{J}_{22} \cdot u_y \end{cases} \quad (11)$$

Thus, the original system is decoupled into two first-order systems. Where and are the total disturbances in the  $x$  and  $y$  directions of the system.

Particularly, for the first-order ADRC, the role of the TD module is to track the input signal, so removed the TD module can make the system faster and avoid the resulting time delay. Setting the step is  $h$ , the discrete ADRC of the system can be designed as:

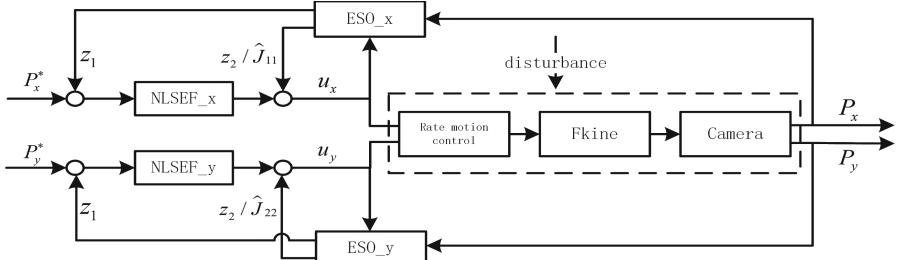
$$\begin{cases} e_1 = z_1(n) - p(n) \\ z_1(n+1) = z_1(n) + h \cdot (z_2(n) - \beta_{01} \cdot fal(e_1, \alpha_{01}, \delta_1) + b_0 \cdot u(n)) \\ z_2(n+1) = z_2(n) - h \cdot \beta_{02} \cdot fal(e_1, \alpha_{02}, \delta_2) \\ e_2 = p^*(n) - z_1(n) \\ u(n) = \beta_{03} \cdot fal(e_2, \alpha_{03}, \delta_3) - z_2(n)/b_0 \end{cases} \quad (12)$$

Where

$$fal(e, \alpha, \delta) = \begin{cases} |e|^\delta \cdot sign(e) & |e| \geq \delta \\ e/\delta^{1-\alpha} & |e| < \delta \end{cases} \quad (13)$$

$p(n)$  is the position of the paw in the image at the current moment;  $p^*(n)$  is the desired position of the paw in the image at the current moment;  $e_1$  is the observation error of ESO;  $e_2$  is the error value;  $z_1$  is the observation value of the system output;  $z_2$  is the observed estimate of the total disturbance;  $\beta_{01}$ ,  $\beta_{02}$ , are the gains of ESO,  $\beta_{03}$  is the gain of NLSEF;  $u$  is input signal;  $fal(e, \alpha, \delta)$  is a nonlinear function;  $b_0$  is corresponding to  $\hat{J}_{11}$  and  $\hat{J}_{22}$  in the  $x$  and  $y$  directions, respectively.

The block diagram of the entire system is shown in Fig. 1.



**Fig. 1.** Hand-eye coordination structure diagram

## 4 Simulation Results

In this section, the ABBIRB2400 manipulator is modeled using the D-H method, and the above control scheme is simulated through matlab to verify the effectiveness of the control.

The mechanical arm link model parameters are shown in Table 1.

**Table 1.** Mechanical arm link model parameters

$i$	$a_{i-1}$	$\alpha_{i-1}$	$d_i$	$\theta_i$	Joint variable range
1	100	-90°	0	90°	-180° ~ +180°
2	705	0	0	-90°	-100° ~ +110°
3	135	-90°	0	0	-400° ~ +400°
4	0	90°	755	0	-200° ~ +200°
5	0	-90°	0	0	-120° ~ +120°
6	0	0	0	0	-400° ~ +400°

Where  $a$  is the length of the connecting rod;  $\alpha$  is the torsion angle of the connecting rod;  $d$  is the joint distance;  $\theta$  is the initial joint rotation angle. The transformation from link coordinate system  $i - 1$  to coordinate system  $i$  is

$${}^{i-1}A_i = \begin{bmatrix} \cos \theta_i & -\sin \theta_i \cos \alpha_i & \sin \theta_i \cos \alpha_i & a_i \cos \theta_i \\ \sin \theta_i & \cos \theta_i \cos \alpha_i & -\cos \theta_i \sin \alpha_i & a_i \sin \theta_i \\ 0 & \sin \alpha_i & \cos \alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (14)$$

Suppose the internal and external parameters of the camera are: the camera focal length  $f = 6$  mm; the camera image plane quantization resolution are  $N_x = 1e - 04$ ,  $N_y = 1e - 04$ ; Camera image center position are  $X_c = 640$ ,  $Y_c = 512$ ; The rotational angles of the camera pose with respect to the robotic base coordinates are  $\varphi = 0$ ,  $\theta = \pi$ ,  $\phi = \pi$ ; the translational movement vector is  $T = [0, 855, 3500]^T$ ; the system step is  $h = 0.1$ .

The initial position of the hand is at the center point (640, 512) of the image plane. ADRCs with the same control parameters in the  $x$  and  $y$  directions, as shown in Table 2.

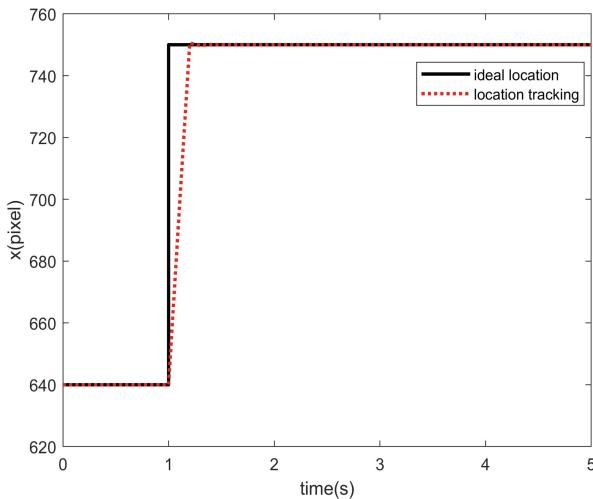
**Table 2.** The ADRC parameters used in simulation

ESO						NLSEF			$b_0$
$\alpha_{01}$	$\delta_1$	$\beta_{01}$	$\alpha_{02}$	$\delta_2$	$\beta_{02}$	$\alpha_{03}$	$\delta_3$	$\beta_{03}$	1.11
0.59	20	100	0.1	0.1	0.001	0.74	20	1.15	

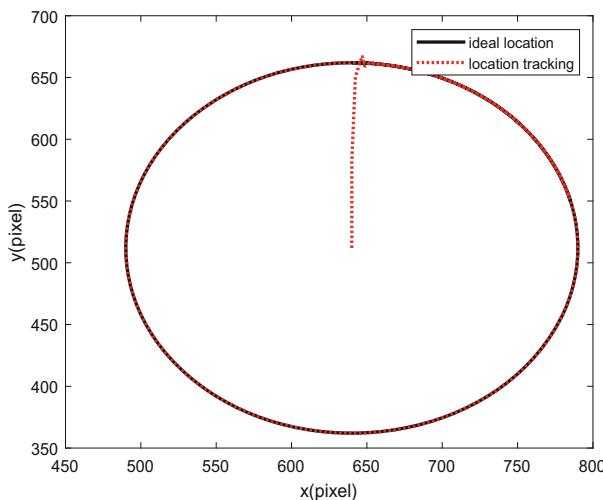
Suppose that the target is making a step movement in  $x$  direction, which is unknown to the robot controller. The response of the system is shown in the Fig. 2.

Suppose that the target is making elliptical movement, which is unknown to the robotic controller

$$\begin{cases} p_x^* = 640 + 150 \sin(t/5) \\ p_y^* = 512 + 150 \cos(t/5) \end{cases} \quad (15)$$



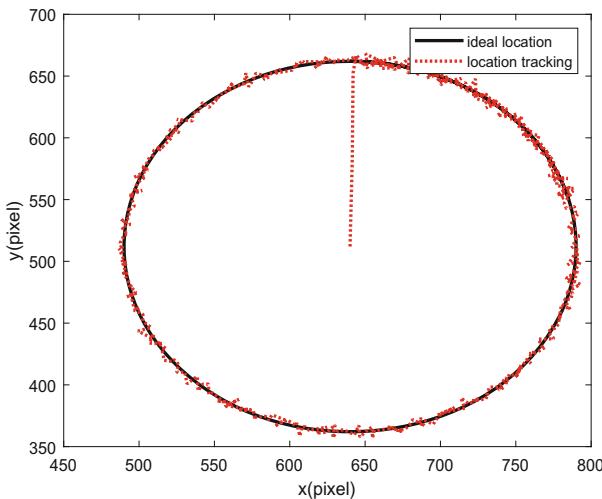
**Fig. 2.** The location tracking curve of the system when the ideal location is a step



**Fig. 3.** The location tracking curve of the system when the ideal location is an ellipse

The response of the system is shown in the Fig. 3.

Suppose that the external disturbance in both the x and y directions are normal distributed random noise with a maximum magnitude of  $\pm 3$  pixels. The response of the system is shown in the Fig. 4.



**Fig. 4.** The location tracking curve of the system when the ideal location add noise

## 5 Conclusion

In this paper, D-H method is used to model the ABBIRB2400 manipulator. The forward kinematics of the manipulator and the mapping relationship between the coordinates of the end of the manipulator and the camera are analyzed. First, the manipulator controlled to make a trial motion to get the initial value of the image Jacobian matrix. When the robot arm is moving, the ADRC is used to continuously compensate the error between the actual value and the initial estimate and the system caused by external disturbance error, until the entire exercise is completed. Finally, the simulation is used to verify the effectiveness of the control method.

## References

1. Hutchinson, S., Hager, G.D., Corke, P.I.: A tutorial on visual servo control. *IEEE Trans. Robot. Autom.* **RA-12**(5), 651–670 (1996)
2. Yoshimi, B.H., Allen, P.K.: Alignment using an uncalibrated camera system. *IEEE Trans. Robot. Autom.* **RA-11**(4), 516–521 (1995)
3. Gong, Z., Tao, B., Yang, H., Yin, Z., Ding, H.: An uncalibrated visual servo method based on projective homography. *IEEE Trans. Autom. Sci. Eng.* **15**(2), 806–817 (2018)
4. Tao, B., Gong, Z., Ding, H.: Research progress of uncalibrated visual servo control for robots. *J. Mech.* **48**(4), 767–783 (2016)
5. Hager, G.D., Chang, W.-C., Morse, A.S.: Robot feedback control based on stereo vision: towards calibration-free hand-eye coordination. In: Proceedings of the 1994 IEEE International Conference on Robotics and Automation, pp. 2850–2856 (1994)

6. Ghadi, M., Laouamer, L., Nana, L., et al.: A novel zero-watermarking approach of medical images based on Jacobian matrix model. *J. Secur. Commun. Netw.* **9**(18), 5203–5218 (2016)
7. Wang, X., Wei, W., Liu, F., et al.: Online estimation of image Jacobian matrix with time-delay compensation. *J. Adv. Comput. Intell. Intell. Inf.* **20**(2), 238–245 (2016)
8. Jiang, P., Bamforth, L.C.A., Feng, Z., et al.: Indirect iterative learning control for a discrete visual servo without a camera-robot model. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **37**(4), 863–876 (2007)
9. Cheah, C.C., Li, X., Yan, X., Sun, D.: Observer-based optical manipulation of biological cells with robotic tweezers. *IEEE Trans. Robot.* **30**(1), 68–80 (2014)
10. Su, J., Xi, Y., Hanebeck, U.D., Schmidt, G.: Nonlinear visual mapping model for 3-D visual tracking with uncalibrated eye-in-hand robotic system. *IEEE Trans. Syst. Man Cybern. (Cybern.) Part B* **34**(1), 652–659 (2004)
11. Su, J., Pan, Q., Xi, Y.: Dynamic coordination of uncalibrated hand/eye robotics system based on neural network. *J. Syst. Eng. Electron.* **12**(3), 45–50 (2001)
12. Wang, L., Su, J.: Disturbance rejection control for non-minimum phase systems with optimal disturbance observer. *J. ISA Trans.* **57**, 1–9 (2015)



# Adaptive Neural Consensus Tracking for Second-Order Nonlinear Multi-agent Systems with Full-State Constraints

Dan Liu and Lin Zhao<sup>(✉)</sup>

Qingdao University, Qingdao 266071, China  
zhaolin1585@163.com

**Abstract.** In this paper, a scheme to handle the consensus tracking problem of multi-agent systems that feature second-order and nonlinearity under the conditions of full-state constraints is further discussed. In order to make the output of each agent track the output of leader accurately without explosion of complexity problem in traditional backstepping, the backstepping method using the design of command filter is adopted. The filtering process will produce errors, so the compensation signal is adopted to further guarantee the tracking precision. Moreover, the innovative adaptive control law that need only one parameter is proposed and the output signal do not exceed the constrained region in the tracking process is proved. The neural network technology is introduced to approximate the dynamics that feature unknown nonlinearities. An example of mathematical simulation verifies the validity of involved method.

**Keywords:** Adaptive neural control · Full-state constraint · Command filter backstepping · Nonlinear multi-agent systems

## 1 Introduction

In the last few years, the researchers has been having their eyes on the topic that the applications of multi-agent systems (MASs) feature distributed cooperative control in astrovessel formation voyage, multi-unmanned aerial aerobats, multi-sensing element networks [1, 2]. The issues of MASs with unknown nonlinear have been discussed sufficient in [3, 4].

One of the most excellent solution to the tracking problems of systems that feature nonlinear is backstepping. The foundation of backstepping technology is based on [5] and the virtual control signals using the dynamic surface control scheme to avoid the problem that computational explosion are established in [6, 7]. In order to decrease the explosion degree of calculation in the project of backstepping, the command filter that features the input equation using virtual control signal has been introduced in [8, 9]. And the neural network method are adopted to approximate the project of dynamics that feature unknown nonlinearities in [10]. However, in the above-mentioned articles, the problem of state

constraint has not been solved well. The control scheme of full-state constraints has been designed in [11], but it only considers the single-agent system, while the multi-agent systems are used in [12] with only output constraint.

In order to deal with the state constrained problem, we will propose an innovative method based on distributed command filtered backstepping technique combined with neural network to analysis the solution of consensus tracking for nonlinear MASs feature uncertain dynamics, which can not only guarantee the whole convergence for states of agent to a predetermined common value, but also let all states be limited to the desired regions in the whole process of convergence.

## 2 System Description

The control scheme investigates nonlinear MAS feature one leader with  $N$  followers which the directed graph  $\bar{\mathcal{G}}$  indicates the reciprocal relations among them. To describe the  $i$ th follower, the dynamic equations are designed as

$$\begin{aligned}\dot{x}_{i,1} &= f_{i,1}(x_{i,1}) + g_{i,1}(x_{i,1})x_{i,2} \\ \dot{x}_{i,2} &= f_{i,2}(x_i) + g_{i,2}(x_i)u_i \\ y_i &= x_{i,1}, \quad i \in \mathcal{V}\end{aligned}\tag{1}$$

in which the state vector is described as  $x_i = [x_{i,1}, x_{i,2}]^T \in \mathbb{R}^2$ , the system output is described as  $y_i \in \mathbb{R}$  and the control input is described as  $u_i \in \mathbb{R}$ .  $f_{i,m}(\cdot), m = 1, 2$  is the uncertain nonlinear function that features unknown and smooth while the output equation of leader agent is described as  $r(t) \in \mathbb{R}$ .

All the states of the system are constrained in a compact set which  $|x_{i,m}| \leq c_{i,m}$  with  $c_{i,m} > 0, i \in \mathcal{V}, m = 1, 2$ .

**Assumption 1.**  $\bar{\mathcal{G}}$  features a spanning tree. Besides, the system's root node is represented by leader node.

**Assumption 2.** The function  $g_{i,m}(\cdot)$  is known, smooth and nonlinear with a bounded region, such as  $\rho_1 \leq |g_{i,m}(\cdot)| \leq \rho_2$ , where  $\rho_1 < \rho_2 < 0$  are known constants.

**Assumption 3.** Both of  $r(t)$  and  $\dot{r}(t)$  feature known, smooth and bounded with  $|r| \leq \gamma$ .

**Assumption 4.** The function  $g_{i,m}(\cdot)$  is known, smooth and nonlinear with a bounded region, such as  $\rho_1 \leq |g_{i,m}(\cdot)| \leq \rho_2$ , where  $\rho_1 < \rho_2 < 0$  are known constants.

## 3 Main Result

Consider the system local errors are defined as:

$$z_{i,1} = \sum_{j=1}^N a_{ij}(y_i - y_j) + b_i(y_i - r) \tag{2}$$

$$z_{i,2} = x_{i,2} - \pi_{i,2}$$

where  $\pi_{i,2}$  is the output of HOSM differentiator while the virtual controller  $\alpha_{i,1}$  represent the input. Then command filter with finite-time is designed as [13]

$$\begin{aligned}\dot{\zeta}_{i,1,1} &= \iota_{i,1,1} \\ \iota_{i,1,1} &= -r_{i,1,1}|\zeta_{i,1,1} - \alpha_{i,1}|^{\frac{1}{2}}\text{sign}(\zeta_{i,1,1} - \alpha_{i,1}) + \zeta_{i,1,2} \\ \dot{\zeta}_{i,1,2} &= -r_{i,1,2}\text{sign}(\zeta_{i,1,2} - \iota_{i,1,1})\end{aligned}\quad (3)$$

with  $\pi_{i,2}(t) = \zeta_{i,1,1}$  while its derivative is  $\dot{\pi}_{i,2}(t) = \iota_{i,2,1}$ , then we could get that

$$|\pi_{i,2} - \alpha_{i,1}| \leq \varphi_i \quad (4)$$

Considering the characteristics of the backstepping that feature distributed command filtered, we adopt  $\alpha_{i,m}$  as the virtual control functions:

$$\begin{aligned}\alpha_{i,1} &= \frac{1}{(d_i + b_i)g_{i,1}} \left( -k_{i,1}z_{i,1} - \frac{K_{i,1}\hat{\theta}_i S_{i,1}^T S_{i,1}}{2h_{i,1}^2} - \frac{1}{2}K_{i,1} + \sum_{j=1}^N a_{i,j}g_{j,1}x_{j,2} + b_i\dot{r} \right) \\ u_i &= \frac{1}{g_{i,2}} \left[ -k_{i,2}z_{i,2} - \frac{K_{i,2}\hat{\theta}_i S_{i,2}^T S_{i,2}}{2h_{i,2}^2} - \frac{1}{2}K_{i,2} + \dot{\pi}_{i,2} - \frac{K_{i,1}}{K_{i,2}}(d_i + b_i)g_{i,1}v_{i,2} \right]\end{aligned}\quad (5)$$

where  $k_{i,m}$  are designed positive constants,  $K_{i,m} = v_{i,m}/(l_{i,m}^2 - v_{i,m}^2)$  and positive parameters  $l_{i,m}$  will be designed later with  $m = 1, 2$ . The signals that used for tracking error compensation are adopted as

$$v_{i,m} = z_{i,m} - \xi_{i,m} \quad (6)$$

and the signals that used for error compensation are adopted as

$$\begin{aligned}\dot{\xi}_{i,1} &= -k_{i,1}\xi_{i,1} + (d_i + b_i)g_{i,1}(\pi_{i,2} - \alpha_{i,1} + \xi_{i,2}) \\ \dot{\xi}_{i,2} &= -k_{i,2}\xi_{i,2}\end{aligned}\quad (7)$$

The Lyapunov function is set up as

$$V_{i,1} = \frac{1}{2} \ln \frac{l_{i,1}^2}{l_{i,1}^2 - v_{i,1}^2} \quad (8)$$

further the derivative of  $V_{i,1}$  is calculated

$$\dot{V}_{i,1} = K_{i,1} \left[ \bar{f}_{i,1} + (d_i + b_i)g_{i,1}v_{i,2} - k_{i,1}v_{i,1} - \frac{K_{i,1}\hat{\theta}_i S_{i,1}^T S_{i,1}}{2h_{i,1}^2} - \frac{1}{2}K_{i,1} \right] \quad (9)$$

where  $\bar{f}_{i,1} = (d_i + b_i)f_{i,1} - \sum_{j=1}^N a_{i,j}f_{j,1}$  is approximated as

$$\bar{f}_{i,1} = W_{i,1}^T S_{i,1}(Z_{i,1}) + \delta_{i,1} \quad (10)$$

where  $Z_{i,1} = [x_{i,1}, x_{j,1}]^T$ ,  $|\delta_{i,1}| \leq \varepsilon_{i,1}$  and  $\forall \varepsilon_{i,1} > 0$ . So we could get

$$K_{i,1}\bar{f}_{i,1} \leq \frac{1}{h_{i,1}^2} K_{i,1}^2 \|W_{i,1}\|^2 S_{i,1}^T S_{i,1} + \frac{1}{2} (h_{i,1}^2 + K_{i,1}^2 + \varepsilon_{i,1}^2) \quad (11)$$

in which  $h_{i,1} > 0$ . Then, the result of (9) could be derived as

$$\begin{aligned} \dot{V}_{i,1} \leq & \frac{1}{2h_{i,1}^2} K_{i,1}^2 \left( \|W_{i,1}\|^2 - \hat{\theta}_i \right) S_{i,1}^T S_{i,1} + \frac{1}{2} (h_{i,1}^2 + \varepsilon_{i,1}^2) \\ & - k_{i,1} K_{i,1} v_{i,1} + (d_i + b_i) g_{i,1} K_{i,1} v_{i,2}. \end{aligned} \quad (12)$$

The second Lyapunov function is set up as

$$V_{i,2} = V_{i,1} + \frac{1}{2} \ln \frac{l_{i,2}^2}{l_{i,2}^2 - v_{i,2}^2} \quad (13)$$

similarly we could obtain

$$\begin{aligned} \dot{V}_{i,2} \leq & \sum_{m=1}^2 \left[ \frac{1}{2h_{i,m}^2} K_{i,m}^2 \left( \|W_{i,m}\|^2 - \hat{\theta}_i \right) S_{i,m}^T S_{i,m} \right. \\ & \left. + \frac{1}{2} h_{i,m}^2 + \frac{1}{2} \varepsilon_{i,m}^2 - k_{i,m} K_{i,m} v_{i,m} \right] \end{aligned} \quad (14)$$

where  $h_{i,2} > 0$ .

Denote  $\theta_i = \max\{\|W_{i,1}\|^2, \|W_{i,2}\|^2\}$  and  $\tilde{\theta}_i = \theta_i - \hat{\theta}_i$ , then

$$\dot{V}_{i,2} \leq \sum_{m=1}^2 \left( \frac{1}{2h_{i,m}^2} K_{i,m}^2 \tilde{\theta}_i S_{i,m}^T S_{i,m} + \frac{1}{2} h_{i,m}^2 + \frac{1}{2} \varepsilon_{i,m}^2 - k_{i,m} K_{i,m} v_{i,m} \right) \quad (15)$$

and the adaptive control law  $\hat{\theta}$  could be chosen as

$$\dot{\hat{\theta}}_i = -\lambda_i \mu_i \hat{\theta}_i + \sum_{m=1}^2 \frac{1}{2h_{i,m}^2} \lambda_i K_{i,m}^2 S_{i,m}^T S_{i,m} \quad (16)$$

where  $\lambda_i, \mu_i$  are all positive constants.

Then we set up the Lyapunov function to investigate the error compensation signals as

$$\bar{V} = \frac{1}{2} \sum_{i=1}^N \sum_{m=1}^2 \xi_{i,m}^2 \quad (17)$$

so  $\dot{\bar{V}}$  could be calculated as

$$\begin{aligned}
\dot{\bar{V}} &\leq - \sum_{i=1}^N \sum_{m=1}^2 k_{i,m} \xi_{i,m}^2 + \frac{1}{2} \rho_2 \sum_{i=1}^N (d_i + b_i) (\xi_{i,1}^2 + \xi_{i,2}^2 + \varphi_i^2 + \xi_{i,1}^2) \\
&= - \sum_{i=1}^N \left\{ [k_{i,1} - (d_i + b_i)\rho_2] \xi_{i,1}^2 + \left[ k_{i,2} - \frac{1}{2}(d_i + b_i)\rho_2 \right] \xi_{i,2}^2 \right\} \\
&\quad + \frac{1}{2} \rho_2 \sum_{i=1}^N (d_i + b_i) \varphi_i^2 \\
&\leq - p_1 \sum_{i=1}^N \sum_{m=1}^{n_i} \xi_{i,m}^2 + \frac{1}{2} \rho_2 \sum_{i=1}^N (d_i + b_i) \varphi_i^2 \\
&= - 2p_1 \bar{V} + q_1
\end{aligned} \tag{18}$$

where

$$\begin{aligned}
p_1 &= \min_i \left\{ [k_{i,1} - (d_i + b_i)\rho_2], \left[ k_{i,2} - \frac{1}{2}(d_i + b_i)\rho_2 \right] \right\} \\
q_1 &= \frac{1}{2} \rho_2 \sum_{i=1}^N (d_i + b_i) \varphi_i^2
\end{aligned} \tag{19}$$

and the parameters  $k_{i,1} - (d_i + b_i)\rho_2, k_{i,2} - \frac{1}{2}(d_i + b_i)\rho_2$  are all need to be positive.

Denote the Lyapunov function that covers the above parameters as

$$V = \sum_{i=1}^N V_{i,2} + \bar{V} + \sum_{i=1}^N \frac{1}{2\lambda_i} \tilde{\theta}_i^2 \tag{20}$$

then  $\dot{V}$  could be derived

$$\begin{aligned}
\dot{V} &= \sum_{i=1}^N \dot{V}_{i,2} + \dot{\bar{V}} - \sum_{i=1}^N \frac{1}{\lambda_i} \tilde{\theta}_i \dot{\tilde{\theta}}_i \\
&\leq - \sum_{i=1}^N \sum_{m=1}^2 k_{i,m} K_{i,m} v_{i,m} - 2p_1 \bar{V} + q_1 \\
&\quad + \sum_{i=1}^N \mu_i \tilde{\theta}_i \dot{\tilde{\theta}}_i + \frac{1}{2} \sum_{i=1}^N \sum_{m=1}^2 (h_{i,m}^2 + \varepsilon_{i,m}^2) \\
&\leq - \sum_{i=1}^N \sum_{m=1}^2 k_{i,m} \ln \frac{l_{i,m}^2}{l_{i,m}^2 - v_{i,m}^2} - 2p_1 \bar{V} + q_1 \\
&\quad + \sum_{i=1}^N \left( -\frac{\beta_i}{\lambda_i} \tilde{\theta}_i^2 + \frac{1}{2} \mu_i \eta_i \theta_i^2 \right) + \frac{1}{2} \sum_{i=1}^N \sum_{m=1}^2 (h_{i,m}^2 + \varepsilon_{i,m}^2) \\
&\leq - p_2 V + q_2
\end{aligned} \tag{21}$$

where

$$\begin{aligned} p_2 &= \min \{2k_{i,m}, 2p_1, 2\beta_i\} \\ q_2 &= q_1 + \frac{1}{2} \sum_{i=1}^N \sum_{m=1}^2 (h_{i,m}^2 + \varepsilon_{i,m}^2) + \frac{1}{2} \sum_{i=1}^N \mu_i \eta_i \theta_i^2 \\ \beta_i &= \frac{\lambda_i \mu_i (2\eta_i - 1)}{2\eta_i}, \eta_i > \frac{1}{2} \end{aligned} \quad (22)$$

and the parameters  $k_{i,m}$  are all need to be positive.

By solving the differential function

$$\dot{V} \leq -p_2 V + q_2$$

we could have

$$V(t) \leq \left[ V(0) - \frac{q_2}{p_2} \right] e^{-p_2 t} + \frac{q_2}{p_2} \leq V(0) + \frac{q_2}{p_2} \quad (23)$$

further we could have

$$\begin{aligned} |v_{i,m}| &\leq l_{i,m} \sqrt{1 - \exp \left[ -2 \left( V(0) + \frac{q_2}{p_2} \right) \right]} \leq l_{i,m} \\ |\xi_{i,m}| &\leq \sqrt{2 \left( V(0) + \frac{q_2}{p_2} \right)} \end{aligned} \quad (24)$$

Because the compensating signal  $\xi_{i,m}$  can be guaranteed to be bounded, then there exists a upper boundary for  $\xi_{i,m}$  such that  $|\xi_{i,m}| \leq C$  with  $C$  is a positive constant, so we can further obtain that  $|z_{i,m}| \leq l_{i,m} + C$ .

Based on the above formulas, for  $z_{i,1}$ , denote

$$\begin{aligned} Z_1 &= [z_{1,1}, z_{2,1}, \dots, z_{N,1}]^T \\ Y &= [y_1 - r, y_2 - r, \dots, y_N - r]^T \end{aligned} \quad (25)$$

considering Eq. (2) and  $H = D - A + B$ , we could know that  $Z_1 = HY$ , so we have  $Z_1^T Z_1 = Y^T (H^T H) Y$ , then we could get

$$\sigma^2 \sum_{i=1}^N (y_i - r)^2 \leq \sum_{i=1}^N z_{i,1}^2 \quad (26)$$

in which  $\sigma$  represents the minimum singular value for  $H$ . Then we get

$$|y_i - r| \leq \sqrt{\sum_{i=1}^N (y_i - r)^2} \leq \frac{1}{\sigma} \sqrt{\sum_{i=1}^N z_{i,1}^2} \leq \frac{\sqrt{N}}{\sigma} |z_{i,1}|_{\max} \quad (27)$$

in order to ensure  $|y_i| \leq c_{i,1}$ , it only needs  $|y_i| \leq |r| + (\sqrt{N}/\sigma)(l_{i,1} + C)_{\max} \leq (c_{i,1})_{\min}$ , then considering assumption 2,  $l_{i,1}$  should be chosen as  $(l_{i,1})_{\max} \leq -C + (\sigma/\sqrt{N})[(c_{i,1})_{\min} - \gamma]$ .

For  $z_{i,2}$ , denote  $|\pi_{i,2}| \leq \phi_i$ , then we have  $|x_{i,2}| \leq l_{i,2} + C + \phi_i$ , in order to ensure  $|x_{i,2}| \leq c_{i,2}$ , it only needs  $|x_{i,2}| \leq l_{i,2} + C + \phi_i \leq c_{i,2}$ , thus  $l_{i,2} \leq c_{i,2} - C - \phi_i$ .

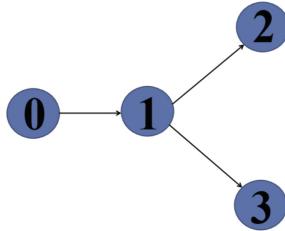
## 4 Numerical Results

For the sake of verifying the feasibility of above scheme, we adopt a nonlinear MAS that uses a directed graph  $\bar{\mathcal{G}}$  to describe the interactions in Fig. 1, in which

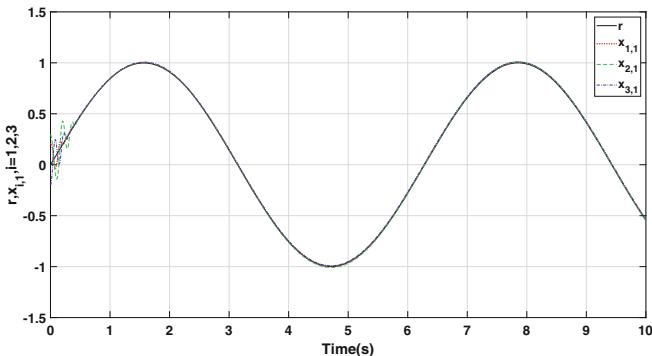
$$\begin{aligned} f_{1,1} &= \cos(x_{1,1}), f_{1,2} = x_{1,1}x_{1,2} \\ f_{2,1} &= \sin(x_{2,1}), f_{2,2} = x_{1,1}e^{-0.3x_{2,2}} \\ f_{3,1} &= \cos(-0.5x_{3,1}), f_{3,2} = 0.5x_{3,1}x_{3,2} \end{aligned} \quad (28)$$

$$g_{i,m} = 1, i = 1, 2, 3, m = 1, 2, r(t) = \sin t, c_{i,1} = 1.5, c_{i,2} = 12, x_1(0), x_2(0), x_3(0)$$

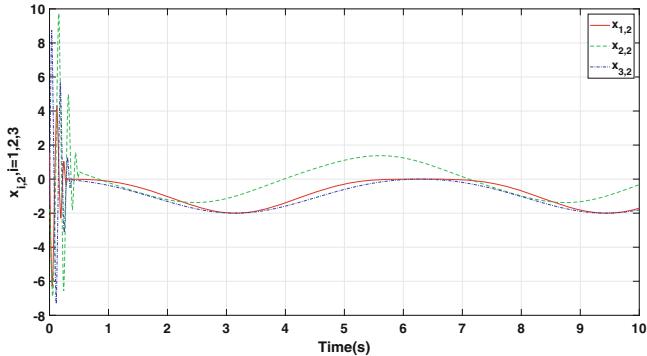
are  $[0.2, 1.6]^T, [0.3, -1.8]^T, [-0.2, -1.5]^T, L = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ . The value of control parameters are  $k_{i,m} = 100, r_{i,m,1} = 50, r_{i,m,2} = 100, h_{i,m} = 1, \lambda_i = 1, \mu_i = 1, l_{i,1} = 0.5, l_{i,2} = 2$ . The responses of  $r, x_{i,1}$  and  $x_{i,2}$  are showed in Fig. 2, Fig. 3, respectively. Obviously the consensus tracking of followers is achieved well while the state constraint are not violated.



**Fig. 1.** The interactions among agents



**Fig. 2.** The time-varying curves of  $r(t)$  and  $x_{i,1}, i = 1, 2, 3$



**Fig. 3.** The time-varying curves of  $x_{i,2}, i = 1, 2, 3$

## 5 Conclusion

From above discussion, a solution to consensus tracking issue for a species of nonlinear multi-agent systems that feature full-state constraints has been designed. In the engineered control scheme, the issue of complexity explosion is solved by the command filter while the error due to the filter is also compensated appropriately. And the neural network method is introduced to approximate the project of dynamics that feature unknown nonlinearities. To make the whole agent states can have accurate and speedy convergence to the leader signal, the virtual control signal is devised with the tracking error reduce to the allowable range. The control parameters  $l_{i,m}$  are introduced to ensure none of the state exceed the constrained region.

**Acknowledgment.** This work was supported by the Shandong Province Outstanding Youth Fund (ZR2018JL020), the Science and Technology Support Plan for Youth Innovation of Universities in Shandong Province (2019KJN033) and the Project funded by Qingdao Postdoctoral Science Foundation.

## References

1. Olfati-Saber, R., Murray, R.: Consensus problems in networks of agents with switching topology and time-delays. *IEEE Trans. Autom. Control* **49**(9), 1520–1533 (2004)
2. Jia, Y.: Robust control with decoupling performance for steering and traction of 4WS vehicles under velocity-varying motion. *IEEE Trans. Control Syst. Technol.* **8**(3), 554–569 (2009)
3. Zhao, L., Yu, J., Lin, C., Ma, Y.: Adaptive neural consensus tracking for nonlinear multiagent systems using finite-time command filtered backstepping. *IEEE Trans. Syst. Man. Cybern. Syst.* **48**(11), 2003–2012 (2018)
4. Jia, Y.: Alternative proofs for improved LMI representations for the analysis and the design of continuous-time systems with polytopic type uncertainty: a predictive approach. *IEEE Trans. Autom. Control* **48**(8), 1413–1416 (2003)

5. Zhao, L., Jia, Y.: Neural network-based adaptive consensus tracking control for multi-agent systems under actuator faults. *Int. J. Syst. Sci.* **47**(5–8), 1931–1942 (2016)
6. Swaroop, D., Hedrick, J., Yip, P., Gerdes, J.: Dynamic surface control for a class of nonlinear systems. *IEEE Trans. Autom. Control* **45**(10), 1893–1899 (2000)
7. Zhang, T., Ge, S.: Adaptive dynamic surface control of nonlinear systems with unknown dead zone in pure feedback. *Automatica* **44**(7), 1895–1903 (2008)
8. Zhao, L., Yu, J., Lin, C.: Command filter based adaptive fuzzy bipartite output consensus tracking of nonlinear coopetition multi-agent systems with input saturation. *ISA Trans.* **80**, 187–194 (2018)
9. Zhao, L., Yu, J., Lin, C.: Distributed adaptive output consensus tracking of nonlinear multi-agent systems via state observer and command filtered backstepping. *Inf. Sci.* **478**, 355–374 (2019)
10. Zhao, L., Yu, J., Yu, H., Lin, C.: Neuroadaptive containment control of nonlinear multi-agent systems with input saturations. *Int. J. Robust Nonlinear Control* **29**(9), 2742–2756 (2019)
11. Wang, C., Wu, Y., Yu, J.: Barrier Lyapunov Functions-based dynamic surface control for pure-feedback systems with full state constraints. *IET Control Theory Appl.* **11**(4), 524–530 (2016)
12. Sun, J., Liu, C.: Distributed zero-sum differential game for multi-agent systems in strict-feedback form with input saturation and output constraint. *Neural Netw.* **106**, 8–19 (2018)
13. Levant, A.: Higher-order sliding modes, differentiation and output-feedback control. *Int. J. Control.* **76**(9–10), 924–941 (2003)



# Adaptive Sliding Mode Control of Mismatched Quantization System

Qiaoyu Chen<sup>1(✉)</sup>, Wuneng Zhou<sup>1(✉)</sup>, Dongbing Tong<sup>2(✉)</sup>, and Yao Wang<sup>2</sup>

<sup>1</sup> College of Information Sciences and Technology, Donghua University,  
Shanghai 200051, China

goodluckqiaoyu@126.com, wnzhou@dhu.edu.cn

<sup>2</sup> College of Electronic and Electrical Engineering,  
Shanghai University of Engineering Science, Shanghai 201620, China  
tongdongbing@163.com

**Abstract.** In practical engineering applications, due to the lack of hardware, the parameters of the coding side and the decoding side of the quantizer are inconsistent. The sliding mode control (SMC) for the mismatched quantization system is studied. In the first place, an observer is given to predict the system state trajectory. In the next place, a sliding mode surface (SMS) and an observer are constructed to effectively get rid of the influence of mismatched quantization parameters. Furthermore, it can make the nonlinear system meet the requirement of reaching the SMS and ensure the asymptotically stable of mismatched quantization system. In the end, an example is used to prove the validity of the obtained results.

**Keywords:** Mismatched quantization system · Sliding mode control · Asymptotically stable

## 1 Introduction

In practice, the state of the system is not always known, and there may be unavoidable nondeterminacy, containing system errors and a variety of disturbances. At this time, the adaptive control method will have a better effect [4, 8, 9]. In addition, the sliding mode control (SMC) is a valid control strategy for dealing with uncertainties, parameters changes, and interference. The SMC strategy has been used in lots of actual system [6, 11], such as, automobile engine system and power system. Therefore, the design of SMC has attracted wide attention. In [7], the design of SMC for Markovian jump systems with digital communication constraints is studied.

It should be noted that most of the studies on quantization problems are based on the perfect channel for transmitting control signals (that is, the encoder and decoder are matched). In other words, the effect of hardware loss was not considered in [3, 10, 12]. [5] proposed the concept of quantitative parameter mismatch for the first time. However, it was required that the ratio of parameters on

both sides of encoder and decoder was fixed in [5]. That is to say, it is necessary to adjust the parameters of both sides of the encoder and decoder continuously to keep synchronization. Due to the problem of hardware defects, it is difficult to meet such conditions in the actual situation. After that, in [13], the ratio of the quantization parameters between the decoder and the encoder was promoted to be a positive number changing with time, and the asymptotically stable of mismatched quantization system was obtained through the feedback control strategy. Therefore, it is a new research direction to study the system with mismatched quantization parameters utilizing the SMC.

Based on the above facts, the SMC of unmatched quantization system is studied. Firstly, a suitable sliding mode surface (SMS) is designed. Secondly, the adaptive sliding mode controller based on an observer is given to counteract the effects of disturbance upper bound, quantization mismatch and nonlinearity, and to ensure the asymptotically stable of mismatched quantization system. In the end, an example is given to illustrate the validity of the obtained results.

## 2 Preliminaries and System Description

Taking the nonlinear system into account

$$\begin{cases} \dot{\theta}(t) = A\theta(t) + B(Q(\nu(t)) + g(\theta(t)) + w(t)), \\ \vartheta(t) = C\theta(t), \end{cases} \quad (1)$$

where  $\theta(t) \in \mathbb{R}^n$ ,  $Q(\nu(t)) \in \mathbb{R}^m$ ,  $g(\theta(t)) \in \mathbb{R}^m$ ,  $w(t) \in \mathbb{R}^m$  represent the state variable, the quantized input, the nonlinear function, external noise and all belong to  $\mathcal{L}_2[0, \infty)$ ,  $\vartheta(t) \in \mathbb{R}^q$  is the measurement output.  $A \in \mathbb{R}^{n \times n}$ ,  $B \in \mathbb{R}^{n \times m}$  and  $C \in \mathbb{R}^{q \times n}$ . And,

$$\|g(\theta(t))\| \leq m^{-0.5}(a + b\|\vartheta(t)\|), \quad (2)$$

where  $a$  and  $b$  are unknown positive constants.  $w(t)$  meets  $\|w(t)\| \leq m^{-0.5}w$ ,  $w$  is unclear.

Based on the control input  $\nu(t)$ , the control input  $Q(\nu(t))$  is quantized. It is usually represented by a rounding function.  $Q(\nu(t))$  is designed by

$$Q(\nu(t)) = \mu_d(t)\text{round}\left(\frac{\nu(t)}{\mu_c(t)}\right) = \mu_d(t)q\left(\frac{\nu(t)}{\mu_c(t)}\right), \quad (3)$$

where  $\mu_c(t)$  and  $\mu_d(t)$  are quantization parameters of encoder and decoder, respectively.

In the process of signal transmission, the quantization information  $q\left(\frac{\nu(t)}{\mu_c(t)}\right)$  is generated by the encoder and sent to the decoder through the communication channel. Then, the quantization information  $\mu_d(t)q\left(\frac{\nu(t)}{\mu_c(t)}\right)$  is decoded to generate the quantization signal (i.e.,  $Q(\nu(t))$ ).

In an ideal situation,  $\mu_c(t)$  and  $\mu_d(t)$  are considered equal. It is pointed out in [5] that due to hardware failure, they are not equal, that is,  $\mu_c(t)$  and  $\mu_d(t)$

are not matched. In [13], the problem is further studied, and the proportional parameter model with time variation is established. The scale parameters of  $\mu_c(t)$  and  $\mu_d(t)$  with time change are designed as

$$r(t) = \frac{\mu_d(t)}{\mu_c(t)}, \quad (4)$$

where  $r(t) \in (r_{\min}, r_{\max})$ , positive scalars  $r_{\min}$  and  $r_{\max}$  satisfy  $r_{\max} \geq r_{\min}$ .

*Remark 1.* In [1] and [12], the quantization parameters of encoder side and decoder side are the same, which is under ideal conditions. But in the actual system, this is not feasible. In this paper, the quantization parameters are inconsistent, so more general results are obtained.

*Remark 2.* It is a proportional parameter model that changes with time. This is a further development of [13]. Compared with the constant scale parameter, it is more general.

Combining Eqs. (3) and (4), the following quantized signals can be obtained

$$Q(\nu(t)) = r(t)\mu_c(t)q\left(\frac{\nu(t)}{\mu_c(t)}\right). \quad (5)$$

According to (1) and (5), it has

$$\begin{cases} \dot{\theta}(t) = A\theta(t) + Br(t)\mu_c(t)q\left(\frac{\nu(t)}{\mu_c(t)}\right) + Bg(\theta(t)) + Bw(t), \\ \vartheta(t) = C\theta(t). \end{cases} \quad (6)$$

By introducing the quantization error  $e_{\mu_c} = \mu_c(t)q\left(\frac{\nu(t)}{\mu_c(t)}\right)$  into the system (6), it has

$$\begin{cases} \dot{\theta}(t) = A\theta(t) + Br(t)(e_{\mu_c} + \nu(t)) + Bg(\theta(t)) + Bw(t), \\ \vartheta(t) = C\theta(t), \end{cases} \quad (7)$$

where the quantization error satisfies the inequality  $e_{\mu_c} \leq \Delta\mu_c$ ,  $\Delta = \frac{\sqrt{m}}{2}$ ,  $m$  is the dimension of  $\nu(t)$ .

### 3 The Quantitative SMC

Actually, the accurate information of interference are usually unknown. Therefore,  $a$ ,  $b$  and  $w$  are successively evaluated by  $\hat{a}(t)$ ,  $\hat{b}(t)$  and  $\hat{w}(t)$ . Let  $\hat{a}(t) - a = \tilde{a}(t)$ ,  $\hat{b}(t) - b = \tilde{b}(t)$  and  $\hat{w}(t) - w = \tilde{w}(t)$ .

### 3.1 Construct an Observer

In the first place, an observer is obtained to evaluate the unknown modules. In the second place, the SMS function is constructed by the estimator. According to system (1), the state observer is constructed by

$$\begin{cases} \dot{\hat{\theta}}(t) = A\hat{\theta}(t) + B(Q(\nu(t)) + B\nu_n(t) + N[\vartheta(t) - C\hat{\theta}(t)]), \\ \hat{\vartheta}(t) = C\hat{\theta}(t), \end{cases} \quad (8)$$

where  $\hat{\theta}(t) \in \mathbb{R}^n$  is the estimation of  $\theta(t)$ .  $N \in \mathbb{R}^{q \times n}$  is the observer gain matrix.  $\nu_n(t)$  is used to counteract the influence of  $g(\theta(t))$  and  $w(t)$ . For the following derivation, let  $S_e(t) = B^T P e(t)$ .

Set  $e(t) = \theta(t) - \hat{\theta}(t)$ . By (1) and (8), the error system is

$$\begin{aligned} \dot{e}(t) &= Ae(t) + Bg(\theta(t)) + Bw(t) - B\nu_n(t) - NCe(t) \\ &= (A - NC)e(t) + B[g(\theta(t)) + w(t) - \nu_n(t)]. \end{aligned} \quad (9)$$

The SMS is constructed by

$$S(t) = M\hat{\theta}(t) - \int_0^t M(A + BK)\hat{\theta}(\tau)d\tau, \quad (10)$$

where  $M \in \mathbb{R}^{m \times n}$ ,  $K \in \mathbb{R}^{m \times n}$ , and  $K$  is chosen so that  $A + BK$  is Hurwitz and  $MB$  is the positive definite matrix and nonsingular.

By deriving the two sides of (10), it has

$$\begin{aligned} \dot{S}(t) &= M\dot{\hat{\theta}}(t) - M[A + BK]\hat{\theta}(t) \\ &= MBQ(\nu(t)) + MB\nu_n(t) + MNCE(t) - MBK\hat{\theta}(t). \end{aligned} \quad (11)$$

Set  $\dot{S}(t) = 0$ . By (11), it yields

$$Q_{eq}(\nu(t)) = -\nu_n(t) - (MB)^{-1}MNCE(t) + K\hat{\theta}(t). \quad (12)$$

According to (12) and (8), it obtains

$$\dot{\hat{\theta}}(t) = (A + BK)\hat{\theta}(t) + [I - B(MB)^{-1}M]NCe(t). \quad (13)$$

The nonlinear input  $\nu_n(t)$  is designed as follows

$$\nu_n(t) = (\hat{a}(t) + \hat{b}(t)\|\vartheta(t)\| + \hat{w}(t) + \gamma_1)\frac{1}{\sqrt{m}}\text{sgn}(S_e(t)), \quad (14)$$

where  $\dot{\hat{a}}(t) = \frac{1}{\sqrt{m}}c_a\|S_e(t)\|$ ,  $\dot{\hat{b}}(t) = \frac{1}{\sqrt{m}}c_b\|S_e(t)\|\|\vartheta(t)\|$ ,  $\dot{\hat{w}}(t) = \frac{1}{\sqrt{m}}c_w\|S_e(t)\|$ ,  $c_a > 0$ ,  $c_b > 0$  and  $c_w > 0$ .

### 3.2 System Stability

In the following theorem, the asymptotically stable of system created by (9) and (13) is proved and let  $S = PN$ .

**Theorem 1.** *On the basis of the state observer (8), the SMS function is given in (10). If the matrices  $P > 0$ ,  $N$ ,  $L$  and  $S$  satisfy the following linear matrix inequality*

$$\begin{bmatrix} \Psi_3 & 0 & PE & 0 \\ * & \Psi_4 & 0 & C^T S^T \\ * & * & -P & 0 \\ * & * & * & -P \end{bmatrix} < 0, \quad (15)$$

$$B^T P = LC, \quad (16)$$

where  $\Psi_3 = PA + C^T L^T K + A^T P + K^T LC$ ,  $\Psi_4 = PA + A^T P - SC - C^T S^T$ . The system created by (9) and (13) is asymptotically stable. Furthermore,  $N = P^{-1}S$ .

Proof. Let the Lyapunov function be

$$V_1(\hat{\theta}, e, t) = \hat{\theta}^T(t)P\hat{\theta}(t) + e^T(t)Pe(t) + c_a^{-1}\tilde{a}^2(t) + c_b^{-1}\tilde{b}^2(t) + c_w^{-1}\tilde{w}^2(t). \quad (17)$$

Based on (9) and (13), it gets

$$\begin{aligned} \dot{V}_1 &= 2\hat{\theta}^T(t)P[(A + BK)\hat{\theta}(t) + (I - B(MB)^{-1}M)NCe(t)] \\ &\quad + 2e^T(t)P[(A - NC)e(t) + B(g(\theta(t)) + w(t) - \nu_n(t))] \\ &\quad + 2c_a^{-1}\tilde{a}(t)\dot{\tilde{a}}(t) + 2c_b^{-1}\tilde{b}(t)\dot{\tilde{b}}(t) + 2c_w^{-1}\tilde{w}(t)\dot{\tilde{w}}(t). \end{aligned} \quad (18)$$

Based on Lemma 4.1 in [2], it has

$$\begin{aligned} &2\hat{\theta}^T(t)P(I - B(MB)^{-1}M)NCe(t) \\ &\leq \hat{\theta}(t)PEP^{-1}E^TP\hat{\theta}(t) + e^T(t)C^TN^TPNCe(t), \end{aligned} \quad (19)$$

where  $E = I - B(MB)^{-1}M$ .

According to  $S_e(t) = B^T Pe(t)$  and inequality (18), we have

$$\begin{aligned} &2e^T(t)PB[g(\theta(t)) - \frac{1}{\sqrt{m}}(\hat{a}(t) + \hat{b}(t)\|\vartheta(t)\| + \gamma_1)\text{sgn}(S_e(t))] \\ &\quad + 2c_a^{-1}\tilde{a}(t)\dot{\tilde{a}}(t) + 2c_b^{-1}\tilde{b}(t)\dot{\tilde{b}}(t) \\ &\leq -\frac{2}{\sqrt{m}}\gamma_1\|S_e(t)\| \\ &< 0, \end{aligned} \quad (20)$$

and

$$\begin{aligned} &2e^T(t)PBw(t) + 2c_w^{-1}\tilde{w}(t)\dot{\tilde{w}}(t) - 2e^T(t)PB\frac{1}{\sqrt{m}}\hat{w}(t)\text{sgn}(S_e(t)) \\ &\leq \frac{2}{\sqrt{m}}\|S_e(t)\|(w + \tilde{w}(t) - \hat{w}(t)) \\ &= 0. \end{aligned} \quad (21)$$

By (18)–(21), it obtains

$$\begin{aligned}\dot{V}_1 &\leq 2\hat{\theta}^T(t)P(A+BK)\hat{\theta}(t) + \hat{\theta}^T P E P^{-1} E^T P \hat{\theta}(t) \\ &\quad + e^T C^T N^T P N C e(t) + 2e^T(t)P(A-NC)e(t) \\ &= \xi^T(t)\Omega\xi(t),\end{aligned}\tag{22}$$

where  $\xi(t) = [\hat{\theta}^T(t)e^T(t)]^T$ ,  $\Omega = \begin{bmatrix} \Psi_1 & 0 \\ * & \Psi_2 \end{bmatrix}$ ,  $\Psi_1 = P(A+BK) + (A+BK)^TP + PEP^{-1}E^TP$ ,  $\Psi_2 = PA + A^TP + C^TN^TPNC - PNC - C^TN^TP$ .

According to the Schur lemma and (15), we can get  $\Omega < 0$ . Therefore, the extended system composed of (9) and (13) is asymptotically stable.

Based on (16), we can get

$$\text{tr}[(B^TP - LC)^T(B^TP - LC)] = 0.\tag{23}$$

Based on Eq. (23), one variable  $\mu \geq 0$  can always be found to satisfy

$$(B^TP - LC)^T(B^TP - LC) < \mu I.\tag{24}$$

Then, according to the Schur lemma, it has

$$\begin{bmatrix} -\mu I & (B^TP - LC)^T \\ * & -I \end{bmatrix} < 0.\tag{25}$$

Therefore, the problem of the SMC on the basis of the observer will become the following minimization problem:  $\min \mu$ , s.t. LMIs (15) and (25). This is the end of proof.

### 3.3 Adaptive SMC

In the following subsection, the sliding mode controller will be provided to ensure the reaching condition. And the time-varying proportional parameters are assumed to satisfy  $\|r(t)\| \leq l$ , where  $l = r_{max}$ .

**Theorem 2.** *Assume that Theorem 1 holds. For system (1) and SMS function (10), the state trajectory of (8) will be driven to the SMS in a limited time. Moreover, the sliding mode controller is given by*

$$\nu(t) = \nu_a(t) + \nu_b(t),\tag{26}$$

where  $\nu_a(t)$  represents the linear control input part,  $\nu_b(t)$  represents the nonlinear control input part. The two control inputs  $\nu_a(t)$  and  $\nu_b(t)$  are

$$\begin{aligned}\nu_a(t) &= \frac{K}{r(t)}\hat{\theta}(t), \\ \nu_b(t) &= -(r(t))^{-1}(\delta(t) + \hat{a}(t)\|MB\| + \hat{b}(t)\|MB\|\|\vartheta(t)\| \\ &\quad + \hat{w}(t)\|MB\| + \gamma_1\|MB\| + \eta)(MB)^{-1}\text{sgn}(S(t)).\end{aligned}\tag{27}$$

Furthermore,  $\eta$  is a positive constant small enough,  $\delta(t) > 0$  and  $\delta(t) = \|MN\| \|\vartheta(t)\| + \|MNC\| \|\hat{\theta}(t)\| + l\Delta\mu_c \|MB\|$ . And  $\dot{a}(t) = \frac{1}{\sqrt{m}}c_a \|S_e(t)\|$ ,  $\dot{b}(t) = \frac{1}{\sqrt{m}}c_b \|S_e(t)\| \|\vartheta(t)\|$ ,  $\dot{w}(t) = \frac{1}{\sqrt{m}}c_w \|S_e(t)\|$ , where  $c_a$ ,  $c_b$  and  $c_w$  are positive numbers of design choices. The controller (26) can ensure the reachability of (8) in a limited time.

**Proof:** Based on (8), the Lyapunov function is selected by

$$V_2(t) = \frac{1}{2}S^T(t)S(t). \quad (28)$$

Based on (11), it gets

$$\begin{aligned} \dot{V}_2(t) &= S^T(t)\dot{S}(t) \\ &= S^T(t)[MBQ(\nu(t)) + MB\nu_n(t) + MNCE(t) - MBK\hat{\theta}(t)]. \end{aligned} \quad (29)$$

According to (26), it yields

$$\begin{aligned} \dot{V}_2(t) &= S^T(t)[MBr(t)e_{\mu_c} - MB\delta(t)(MB)^{-1}\text{sgn}(S(t)) + MB\nu_n(t) \\ &\quad + MNCE(t) - MB[(\hat{a}(t) + \hat{b}(t)\|\vartheta(t)\| + \hat{w}(t))\|MB\| \\ &\quad + \gamma_1\|MB\| + \eta](MB)^{-1}\text{sgn}(S(t))]. \end{aligned} \quad (30)$$

Based on inequality  $S^T(t)\frac{1}{\sqrt{m}}\text{sgn}(S_e(t)) < \|S(t)\|$ , it has

$$\begin{aligned} &-(\hat{a}(t)\|MB\| + \hat{b}(t)\|MB\|\|\vartheta(t)\| + \hat{w}(t)\|MB\| \\ &\quad + \gamma_1\|MB\|)S^T(t)\text{sgn}(S(t)) + S^T(t)MB\nu_n(t) \leq 0. \end{aligned} \quad (31)$$

By (27), we have

$$S^T[MBr(t)e_{\mu_c} - \delta(t)\text{sgn}(S(t)) + MNCE(t)] \leq 0. \quad (32)$$

By (30)–(32), it yields

$$\dot{V}_2(t) \leq -\eta \|S(t)\| < 0. \quad (33)$$

Then it can be concluded that  $S^T(t)\dot{S}(t) < 0$ , which means that the state trajectory of (8) can reach to the predetermined SMS in a limited time. This is the end of proof.

*Remark 3.* Noting that  $M$  is chosen to satisfy the nonsingularity of  $MB$ . In this paper,  $MB$  does not need to be positive. Therefore, the Lyapunov function  $V(t) = S^T(t)S(t)$  is constructed. Another  $V(t) = S^T(t)(MB)^{-1}S(t)$  is constructed in [7].

*Remark 4.* When the hardware failure is not considered, the communication channel is perfect, i.e. quantized parameter matching. In the above proof process, it can make  $r_{\max} = r_{\min} = 1$  and  $l = 1$ .

## 4 Numerical Example and Simulation

Consider system (1) with matrix parameters as below:

$$A = \begin{bmatrix} -1.86 & 1.5 & 0.2 \\ 0 & -3.279 & -2.3 \\ 0 & 2.928 & -2.082 \end{bmatrix}, B = \begin{bmatrix} 0.1 \\ 0.2 \\ -0.4 \end{bmatrix}, C = \begin{bmatrix} 0.5 & -0.2 & 0.1 \\ -0.2 & 0.1 & -0.5 \\ 0.2 & 0.2 & 0 \end{bmatrix}.$$

Select  $B^T I_{3 \times 3}$  as matrix  $M$ . And select the following matrix  $K = [-1.2 \ -1.6 \ 1.1]$ . The nonlinear function and the external disturbance are designed as  $g(\theta(t)) = 2 + 2 \sin(t)\theta_1(t)$ ,  $w(t) = e^{-t} \sin(2t)$ .  $\theta(0) = [2.5, -1.5, 0.9]^T$  and  $\hat{\theta}(0) = [0.5, -0.5, 0.9]^T$ . The designed parameters can be selected as  $\hat{a}(0) = 3$ ,  $\hat{b}(0) = 3$ ,  $\hat{w}(0) = 1$ ,  $c_a = c_b = c_w = 0.2$ ,  $\gamma_1 = 0.1$ . To prevent control signal jitter,  $\text{sgn}(S(t))$  and  $\text{sgn}(S_e(t))$  are replaced by  $\frac{S(t)}{\|S(t)\|+0.01}$  and  $\frac{S_e(t)}{\|S_e(t)\|+0.01}$ , respectively. By solving LMIs (15) and (25), the observer matrix is solved as

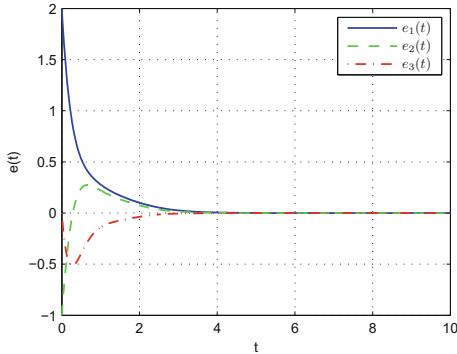
$$N = \begin{bmatrix} 1.5625 & 0.3125 & 1.4063 \\ -1.5625 & -0.3125 & 3.5937 \\ -0.9375 & -2.1875 & 0.1563 \end{bmatrix}.$$

Then,  $S(t) = [0.1 \ 0.2 \ -0.4] \hat{\theta}(t) - \int_0^t [-0.042 \ -2.012 \ 0.6288] \hat{\theta}(\tau) d\tau$ .

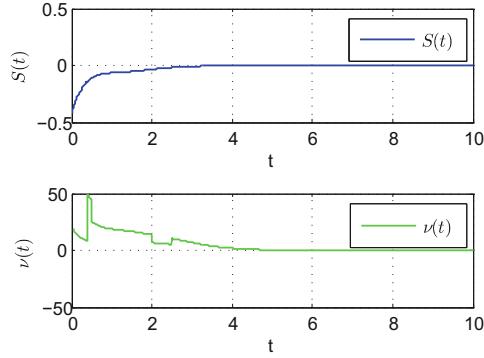
In addition, considering the problem of unmatched quantization parameters, the quantization parameters are selected as

$$\mu_c(t) = \begin{cases} 1, & t \leq 0.5 \\ 0.6, & 0.5 < t \leq 2 \\ 0.2, & t > 2 \end{cases}, \quad \mu_d(t) = \begin{cases} 1.2, & t \leq 0.4 \\ 0.2, & 0.4 < t \leq 2.5 \\ 0.1, & 2.5 < t < 6 \\ 0.04, & t > 6 \end{cases}$$

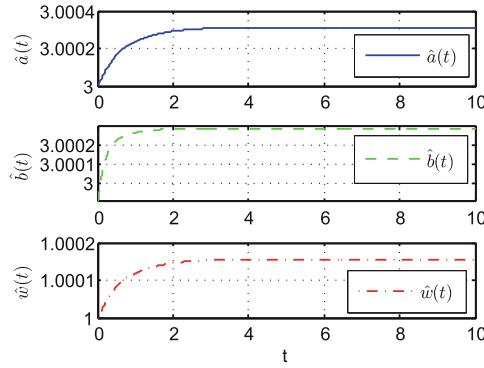
In order to show the validity of the design method, the state estimation error is shown in Fig. 1. Figure 2 depicts the SMS and the controller. Figure 3 shows the estimations for  $a$ ,  $b$  and  $w$ .



**Fig. 1.** Dynamic curves of estimation error.



**Fig. 2.** Dynamic curves of SMS and controller.



**Fig. 3.** Estimations of  $a$ ,  $b$  and  $w$ .

## 5 Conclusion

In this paper, SMC for mismatched quantization systems is studied when the nonlinear systems have external disturbances and unmeasurable states. The controller can adaptively match the unknown upper bound on the one hand, on the other hand, it can deal with the influence of mismatch quantization to ensure the reachability of system state trajectory. An example is used to prove the validity of the obtained results.

**Acknowledgements.** This work is partially supported by the National Natural Science Foundation of China (61673257; 61573095), and the China Postdoctoral Science Foundation (2019M661322).

## References

1. Hao, L., Park, J.H., Ye, D.: Integral sliding mode fault-tolerant control for uncertain linear systems over networks with signals quantization. *IEEE Trans. Neural Netw. Learn. Syst.* **28**(9), 2088–2100 (2017)
2. Hu, J., Wang, Z., Gao, H., Stergioulas, L.K.: Robust sliding mode control for discrete stochastic systems with mixed time delays, randomly occurring uncertainties, and randomly occurring nonlinearities. *IEEE Trans. Ind. Electron.* **59**(7), 3008–3015 (2012)
3. Jia, Y.: Robust control with decoupling performance for steering and traction of 4WS vehicles under velocity-varying motion. *IEEE Trans. Control Syst. Technol.* **8**(3), 554–569 (2000)
4. Jia, Y.: Alternative proofs for improved LMI representations for the analysis and the design of continuous-time systems with polytopic type uncertainty: A predictive approach. *IEEE Trans. Autom. Control* **48**(8), 1413–1416 (2003)
5. Kameneva, T., Nešić, D.: Robustness of quantized control systems with mismatch between coder/decoder initializations. *Automatica* **45**(3), 817–822 (2009)
6. Lin, F.J., Chang, C.K., Po-Kai, H.: FPGA-based adaptive backstepping sliding-mode control for linear induction motor drive. *IEEE Trans. Pow. Electron.* **22**(4), 1222–1231 (2007)
7. Liu, M., Zhang, L., Shi, P., Zhao, Y.: Sliding mode control of continuous-time Markovian jump systems with digital data transmission. *Automatica* **80**, 200–209 (2017)
8. Tong, D., Xu, C., Chen, Q., Zhou, W.: Sliding mode control of a class of nonlinear systems. *J. Franklin Inst.* **357**(3), 1560–1581 (2020)
9. Tong, D., Xu, C., Chen, Q., Zhou, W., Xu, Y.: Sliding mode control for nonlinear stochastic systems with Markovian jumping parameters and mode-dependent time-varying delays. *Nonlinear Dyn.* **100**(2), 1343–1358 (2020)
10. Wang, Y., Tong, D., Chen, Q., Zhou, W.: Exponential synchronization of chaotic systems with stochastic perturbations via quantized feedback control. *Circ. Syst. Sig. Process.* **39**(1), 474–491 (2020)
11. Xu, C., Tong, D., Chen, Q., Zhou, W., Shi, P.: Exponential stability of Markovian jumping systems via adaptive sliding mode control. *IEEE Trans. Syst. Man Cybern. Syst.* (2019). <https://doi.org/10.1109/TSMC.2018.2884565>
12. Xue, Y., Zheng, B., Yu, X.: Robust sliding mode control for T-S fuzzy systems via quantized state feedback. *IEEE Trans. Fuzzy Syst.* **26**(4), 2261–2272 (2018)
13. Zheng, B.C., Yang, G.H.:  $H_2$  control of linear uncertain systems considering input quantization with encoder/decoder mismatch. *ISA Trans.* **52**(5), 577–582 (2013)



# An Optimal Feedback Control Law for Spacecraft Reorientation with Attitude Pointing Constraints

Bin Li<sup>1</sup>, Yang Wang<sup>1</sup>, and Kai Zhang<sup>2(✉)</sup>

<sup>1</sup> Sichuan University, Chengdu, Sichuan 610065, China

[bin.li@scu.edu.cn](mailto:bin.li@scu.edu.cn), [wy@stu.scu.edu.cn](mailto:wy@stu.scu.edu.cn)

<sup>2</sup> Southwest Jiaotong University, Chengdu, Sichuan 610031, China

[kaizhang@swjtu.edu.cn](mailto:kaizhang@swjtu.edu.cn)

**Abstract.** A feedback control design for space reorientation mission with attitude pointing constraints is investigated in this paper. By introducing the convex parameterizations of sensor's field of view and antenna communication constraints, a novel potential function is proposed, which offers more design flexibility for the controller. Based on this potential function, a proportional differential-like attitude control law is then constructed to guarantee the stability of the system. Different from most of the existing works, which usually randomly choose the parameters of the control law within feasible region, the proposed design method further improves the control performance by optimizing these parameters. Simulation results are provided to show the effectiveness and superior of the proposed method.

**Keywords:** Spacecraft attitude control · Attitude constraints · Potential function · Optimal control

## 1 Introduction

Spacecraft reorientation is a challenging space mission. During the maneuvers, the scientific sensor's (space telescopes or star sensors) field of view (FOV) must be kept away from unwanted celestial objects (sun or moon) to avoid affecting the measurement precision. In addition, to guarantee the space-to-ground communication, the spacecraft must keep the antenna pointing to the earth station. Therefore, the spacecraft attitude is required not only to precisely reach the desired state, but also to rigorously satisfy certain attitude constraints.

Since the attitude control problem with constraints is a key issue, it has attracted much attentions. The existing works in the literatures can be divided into two categories: the path planning method [1–8] and potential function method [9–12]. The path planning method is to determine a feasible path before maneuvering. For example, in [1] and [2],  $A^*$  algorithm is used to search for feasible paths under the angular velocity and control constraints. [3] applies

the algorithms of enumeration points and graph search to find the shortest path with multiple triaxial attitude constraints. In [5], a feasible attitude maneuvering path with pointing constraint is obtained by semi-definite programming method. Although the path planning method can find a feasible path, its computational complexity and execution time are much larger than that of potential function method. A convex logarithmic potential function is constructed to design the attitude control laws in [10]. Based on the potential function method, [11] proposes a speed-free attitude control law. In the presence of attitude constraints, by integrating with a nonlinear disturbance observer and potential function method, [12] proposes a backstepping design method. However, the previous works can only guarantee the asymptotic stability, and parameters of the control law are randomly chosen within feasible region.

Motivated by this fact, this paper focuses on the attitude control for space-craft reorientation with two different types of attitude pointing constraints. By introducing the convex parameterizations of forbidden areas of sensor's FOV and mandatory areas of antenna communication, a novel potential function based attitude tracking control law is proposed, which can guarantee the satisfaction of attitude pointing constraints and asymptotic stability. Furthermore, a numerical optimization based tuning method is developed to optimize the control parameters. The main contributions are summarized as follows:

- 1) The proposed potential function introduces a design variable, which offers more flexibility on the controller design.
- 2) Most of previous works randomly choose the parameters of attitude control law within the feasible region. In this paper, an optimal parameters tuning method is proposed which highly improves the performance.

## 2 Preliminaries

### 2.1 Attitude Kinematics and Dynamics

In order to describe the relative orientation between of the body fixed frame with respect to a inertial frame is, the following unit quaternion is used

$$\mathbf{Q} = [q_1 \quad q_2 \quad q_3 \quad q_4]^T = [\mathbf{q}^T \quad q_4]^T \quad (1)$$

where the Euler parameters satisfy  $\mathbf{Q} \in \mathcal{U}_Q = \{\mathbf{Q} | \mathbf{q}^T \mathbf{q} + q_4^2 = 1\}$ . The quaternion conjugate is defined as  $\mathbf{Q}^* = [-\mathbf{q}^T \quad q_4]^T$ . For the quaternions  $\mathbf{Q} = [\mathbf{q}^T \quad q_4]^T$  and  $\mathbf{P} = [\mathbf{p}^T \quad p_4]^T$ , the quaternion multiplication is defined as

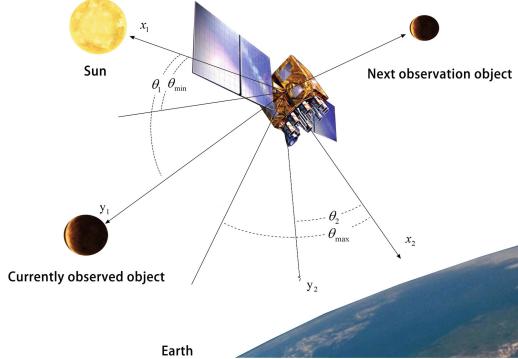
$$\mathbf{Q} \otimes \mathbf{P} = \begin{bmatrix} q_4 \mathbf{p} + p_4 \mathbf{q} + \mathbf{q} \times \mathbf{p} \\ q_4 p_4 - \mathbf{q}^T \mathbf{p} \end{bmatrix} \quad (2)$$

The spacecraft is assumed to be a rigid body with actuators that provide torques about three mutually perpendicular axes, the kinematics and dynamics model of a spacecraft can be described as follows:

$$\begin{aligned} \dot{\mathbf{Q}} &= \frac{1}{2} \mathbf{Q} \otimes \tilde{\boldsymbol{\omega}} \\ I\dot{\boldsymbol{\omega}} &= -\boldsymbol{\omega} \times I\boldsymbol{\omega} + \mathbf{u} \end{aligned} \quad (3)$$

where  $\mathbf{I} = \text{diag}(I_1, I_2, I_3)$  denotes the inertia matrix,  $\mathbf{Q} = [\mathbf{q}^T \ q_4]^T$  represents the attitude quaternion,  $\boldsymbol{\omega} = [\omega_1 \ \omega_2 \ \omega_3]^T$  is the angular velocity,  $\tilde{\boldsymbol{\omega}} = [\boldsymbol{\omega}^T \ 0]^T$ ,  $\mathbf{u} = [u_1 \ u_2 \ u_3]^T$  is the control torque.

## 2.2 Attitude Constraints



**Fig. 1.** The spacecraft attitude constraints

During the orientation, the attitude is required to satisfy many constraints. As shown in Fig. 1, the sensor fixed on spacecraft must avoid the direct exposure to certain celestial objects (e.g., the sun), otherwise they may affect the measurement precision. Thus, the angle  $\theta_1$  between the normalized boresight vector  $\mathbf{y}_1$  of fixed sensor and the normalized vector  $\mathbf{x}_1$  pointing toward a certain celestial object should be greater than the critical angle  $\theta_{min}$ . Further, in order to maintain the communication with the ground stations or the other spacecrafts, the angle  $\theta_2$  between the vector of antenna  $\mathbf{y}_2$  and the vector of receiver  $\mathbf{x}_2$  should be smaller than the critical angle  $\theta_{max}$ . Such attitude constraints can be expressed by

$$\mathbf{x}_1^T \mathbf{y}'_1 < \cos \theta_{min}, \quad \mathbf{x}_2^T \mathbf{y}'_2 > \cos \theta_{max} \quad (4)$$

where  $\mathbf{x}_1, \mathbf{y}'_1, \mathbf{x}_2$  and  $\mathbf{y}'_2$  are the unit vectors,  $\mathbf{y}'_1$  and  $\mathbf{y}'_2$  represent the vectors of sensors and antenna in the inertial coordinates, respectively, which can be obtained by

$$\mathbf{y}'_i = (q_4^2 - \mathbf{q}^T \mathbf{q}) \mathbf{y}_i + 2(\mathbf{q}^T \mathbf{y}_i) \mathbf{q} + 2q_4 (\mathbf{y}_i \times \mathbf{q}), \quad i = 1, 2 \quad (5)$$

Similar to constraint (4), we assume that there exist  $j$  constrained objectives specified by  $\mathbf{x}_1^j$  for the boresight vector  $\mathbf{y}_1$  and the angle  $\theta_{min}^j$  is the relevant constraint angle, then, refer to [10], the attitude quaternion  $\mathbf{Q}$  should be within the following set

$$\mathbf{U}_P = \left\{ \mathbf{Q} \in \mathbf{U}_Q \mid \mathbf{Q}^T \mathbf{M}_1^j (\mathbf{x}_1^j, \mathbf{y}_1, \theta_{min}^j) \mathbf{Q} < 0 \text{ and } \mathbf{Q}^T \mathbf{M}_2 (\mathbf{x}_2, \mathbf{y}_2, \theta_{max}) \mathbf{Q} > 0 \right\}$$

where  $j = 1, 2, \dots, n$ ,  $\mathbf{M}_2$  is defined analogous to  $\mathbf{M}_1^j$  in the following form

$$\mathbf{M}_1^j = \begin{bmatrix} \mathbf{A}_1^j & \mathbf{b}_1^j \\ \mathbf{b}_1^{j\top} & d_1^j \end{bmatrix} \quad (6)$$

with  $\mathbf{E}_3$  being a three dimensional identity matrix,  $\mathbf{A}_1^j = \mathbf{x}_1^j \mathbf{y}_1^\top + \mathbf{y}_1 \mathbf{x}_1^{j\top} - (\mathbf{x}_1^{j\top} \mathbf{y}_1 + \cos \theta_{min}^j) \mathbf{E}_3$ ,  $\mathbf{b}_1^j = \mathbf{x}_1^j \times \mathbf{y}_1$ ,  $d_1^j = \mathbf{x}_1^{j\top} \mathbf{y}_1 - \cos \theta_{min}^j$ .

According to [10], for  $\theta_{min}^j, \theta_{max} \in (0, \pi)$ , we have

$$-2 < \mathbf{Q}^\top \mathbf{M}_1^j \mathbf{Q} < 0, \quad 2 > \mathbf{Q}^\top \mathbf{M}_2 \mathbf{Q} > 0 \quad (7)$$

The objective of this paper is to design an attitude control law such that the quaternion  $\mathbf{Q}$  reaches the desired value  $\mathbf{Q}_d$  while satisfying attitude constraints (4). In addition, the stability has to be guaranteed and parameters of the control law are required to be chosen optimally under certain performance measure.

### 3 Controller Design

In this section, by constructing a novel potential function, an attitude control law is proposed. Both of the stability and the feasibility can be guaranteed under this control law.

#### 3.1 A Novel Potential Function

A novel potential function  $V(\mathbf{Q})$  is constructed which is defined as

$$V(\mathbf{Q}) = \|\mathbf{Q}_e\|^2 \left[ \sum_{j=1}^n k_1^j (e^{1/(-\delta\phi^j(\mathbf{Q}))} - 1) + k_2 (e^{1/(\delta\phi_2(\mathbf{Q}))} - 1) \right] \quad (8)$$

where  $\mathbf{Q}_e = \mathbf{Q} - \mathbf{Q}_d$ ,  $\|\mathbf{Q}_e\|$  denotes the Euclidian 2-norm, the matrices  $\mathbf{M}_1^j$  and  $\mathbf{M}_2$  represent the attitude constraints in (6),  $k_1^j, k_2 > 0$  are the weighting variables,  $\delta$  is a design variable to regulate the influence of potential function,  $\phi^j(\mathbf{Q}) = \frac{\mathbf{Q}^\top \mathbf{M}_1^j \mathbf{Q}}{2}$ ,  $\phi_2(\mathbf{Q}) = \frac{\mathbf{Q}^\top \mathbf{M}_2 \mathbf{Q}}{2}$ .

Then, (8) can be verified as a potential function by the following lemma.

**Lemma 1.** *The potential function (8) satisfies the following three conditions:*

- 1)  $V(\mathbf{Q}_d) = 0$ ,
- 2)  $V(\mathbf{Q}) > 0$ , for all  $\mathbf{Q} \in \mathbf{U}_P \setminus \{\mathbf{Q}_d\}$ ,
- 3)  $\Delta^2 V(\mathbf{Q})$  is positive definite for all  $\mathbf{Q} \in \mathbf{U}_P$ .

*Proof.* Firstly, for all  $\mathbf{Q}_d, \mathbf{Q} \in \mathbf{U}_P$ , it can be easily concluded from (7) that  $V(\mathbf{Q}_d) = 0$  and  $V(\mathbf{Q}) > 0$ , for all  $\mathbf{Q} \in \mathbf{U}_P \setminus \{\mathbf{Q}_d\}$ .

Then, because the summation of strictly convex functions is strictly convex, the proof can be simplified or omitted, that is

$$V(\mathbf{Q}) = \|\mathbf{Q}_e\|^2 \left[ \sum_{j=1}^n k_1^j (e^{1/(-\delta\phi^j(\mathbf{Q}))} - 1) \right] \quad (9)$$

From the relationship  $\frac{\partial}{\partial \mathbf{Q}} \|\mathbf{Q}_e\|^2 = \frac{\partial}{\partial \mathbf{Q}} (2 - 2\mathbf{Q}_d^T \mathbf{Q}) = -2\mathbf{Q}_d^T$ , the gradient of  $V(\mathbf{Q})$  is given by

$$\begin{aligned} \Delta V &= -2\mathbf{Q}_d^T \sum_{j=1}^n k_1^j (e^{1/(-\delta\phi^j(\mathbf{Q}))} - 1) \\ &\quad + \|\mathbf{Q}_e\|^2 \sum_{j=1}^n \delta k_1^j e^{1/(-\delta\phi^j(\mathbf{Q}))} \frac{\mathbf{Q}^T \mathbf{M}_1^j}{(-\delta\phi^j(\mathbf{Q}))^2} \end{aligned} \quad (10)$$

Further, its second order gradient is

$$\begin{aligned} \Delta^2 V &= \sum_{j=1}^n \delta k_1^j e^{1/(-\delta\phi^j(\mathbf{Q}))} \frac{-4\mathbf{Q}_d \mathbf{Q}^T \mathbf{M}_1^j}{(-\delta\phi^j(\mathbf{Q}))^2} + \|\mathbf{Q}_e\|^2 \sum_{j=1}^n e^{1/(-\delta\phi^j(\mathbf{Q}))} \\ &\quad \left[ \delta^2 k_1^j \frac{(\mathbf{Q}^T \mathbf{M}_1^j)^T \mathbf{Q}^T \mathbf{M}_1^j}{(-\delta\phi^j(\mathbf{Q}))^4} + 2\delta^2 k_1^j \frac{(\mathbf{Q}^T \mathbf{M}_1^j)^T \mathbf{Q}^T \mathbf{M}_1^j}{(-\delta\phi^j(\mathbf{Q}))^3} + \delta k_1^j \frac{\mathbf{M}_1^j}{(-\delta\phi^j(\mathbf{Q}))^2} \right] \end{aligned}$$

By pre- and post-multiplying  $\Delta^2 V$  by  $\mathbf{Q}^T$  and  $\mathbf{Q}$ , respectively, it leads to

$$\mathbf{Q}^T \Delta^2 V \mathbf{Q} = \sum_{j=1}^n \delta k_1^j e^{1/(-\delta\phi^j)} \left\{ \frac{(2 - 2\mathbf{Q}_d^T \mathbf{Q}) \delta (2\phi^j)^2 + \frac{\delta^2 (2\phi^j)^3}{2} (5 - 3\mathbf{Q}_d^T \mathbf{Q})}{(-\delta\phi^j)^4} \right\}$$

Obviously  $\mathbf{Q}_d^T \mathbf{Q} \in [-1, 1]$ , therefore  $\mathbf{Q}^T \Delta^2 V \mathbf{Q} > 0$  for all  $\mathbf{Q}_d, \mathbf{Q} \in \mathbf{U}_P$ , then the Hessian of  $V(\mathbf{Q})$  is positive definite. In summary, the function  $V(\mathbf{Q})$  is smooth and strictly convex, it has a global minimum when  $\mathbf{Q} = \mathbf{Q}_d$ . This completes the proof.

Different from the existing potential functions such as [10], (8) has an extra design parameter  $\delta$ , which can offer more design flexibilities. This will be shown later in Sect. 5.

### 3.2 The Control Law

By associating with the novel potential function  $V(\mathbf{Q})$  in (8), the following Lyapunov function is obtained

$$V_2 = k_q V(\mathbf{Q}) + \frac{1}{2} \boldsymbol{\omega}^T \mathbf{I} \boldsymbol{\omega} \quad (11)$$

where  $k_q$  is the positive constant.

Since  $\boldsymbol{\omega} \times \mathbf{I}$  is skew symmetric, then, by taking the derivative of  $V_2$  along the dynamics (3), it follows that given by

$$\dot{V}_2 = \boldsymbol{\omega}^T \left( k_q \text{Vec} \left[ -\nabla V^*(\mathbf{Q}) \otimes \frac{1}{2} \mathbf{Q} \right] + \mathbf{u} \right) \quad (12)$$

where the operator  $\text{Vec}[\cdot]$  represents the vector part of  $[\cdot]$ . Then, we chose the control input in the following form:

$$\mathbf{u} = \frac{1}{2} k_q \text{Vec} [\nabla V^*(\mathbf{Q}) \otimes \mathbf{Q}] - k_\omega \boldsymbol{\omega} \quad (13)$$

where  $k_\omega$  is the positive constant.

In what follows, we need to verify the system stability under the control law (13). By substituting (13) into (12), we have  $\dot{V}_2 = -k_\omega \boldsymbol{\omega}^T \boldsymbol{\omega}$ . According to LaSalle's invariance principle,  $\boldsymbol{\omega} = 0$  is within the invariant set  $\mathbf{N}_p = \{(\mathbf{Q}, \boldsymbol{\omega}) | \dot{V}_2 = 0\}$ . From (3), it follows that  $\mathbf{u} = 0$  if  $\boldsymbol{\omega} = 0$ .

By observing (13), if  $\mathbf{u} = 0$  and  $\boldsymbol{\omega} = 0$ ,  $\mathbf{Q} \neq 0$ , then  $\nabla V^*(\mathbf{Q}) = 0$ . Therefore, from Lemma 1, we can reach the following conclusions straight forward:

$$\{\mathbf{Q} | \nabla V^*(\mathbf{Q}) = 0\} \Leftrightarrow \{\mathbf{Q} | V(\mathbf{Q}) = 0\} \Leftrightarrow \{\mathbf{Q} | \mathbf{Q} = \mathbf{Q}_d\}$$

In general, the largest invariant set in  $\mathbf{N}_p$  is  $\mathbf{M}_p = \{(\mathbf{Q}, \boldsymbol{\omega}) | \mathbf{Q} = \mathbf{Q}_d, \boldsymbol{\omega} = 0\}$ . By applying the LaSalle's invariance principle, it follows that

$$\lim_{t \rightarrow \infty} \boldsymbol{\omega}(t) = 0, \quad \lim_{t \rightarrow \infty} \mathbf{Q}(t) = \mathbf{Q}_d \quad (14)$$

Thus, the stability of the system can be guaranteed under the proposed control law (13).

The feasibility of the attitude constraints (4) under control law (13) can be verified by the following arguments. If  $\mathbf{Q}$  goes to the boundary of the set  $\mathbf{U}_P$ , we have  $V_2 \rightarrow \infty$ , which contradicts with the fact that the (negative) gradient flow property. Therefore, the feasibility of the constraints (4) can be ensured under the proposed control law (13).

To proceed, we derive the exact form of the feedback control law by substituting (10) into (13), which yields

$$\mathbf{u} = -k_q \text{Vec} \left[ \left( l_{g1} \mathbf{Q}_d^* - \frac{1}{\delta} \|\mathbf{Q}_e\|^2 \mathbf{L}_{g2}^* \right) \otimes \mathbf{Q} \right] - k_\omega \boldsymbol{\omega} \quad (15)$$

where

$$\begin{aligned} l_{g1} &= \left( \sum_{j=1}^n k_1^j (e^{1/(-\delta\phi^j(\mathbf{Q}))} - 1) + k_2 (e^{1/(\delta\phi_2(\mathbf{Q}))} - 1) \right) \\ \mathbf{L}_{g2} &= \frac{2}{\delta} \left( \sum_{j=1}^n k_1^j e^{1/(-\delta\phi^j(\mathbf{Q}))} \frac{\mathbf{M}_1^j \mathbf{Q}}{(2\phi^j(\mathbf{Q}))^2} - k_2 e^{1/(\delta\phi_2(\mathbf{Q}))} \frac{\mathbf{M}_2 \mathbf{Q}}{(2\phi_2(\mathbf{Q}))^2} \right) \end{aligned}$$

## 4 An Optimal Parameters Tuning Method

In this section, an optimal parameters tuning algorithm is proposed to optimize attitude controller parameters. The tuning algorithm is based on the control parametrization method [13].

By choosing the performance index in quadric form, the optimal parameter tuning problem for the attitude controller (15) can be formulated as below:

$$\min_k J = \int_0^{t_f} (\mathbf{Q}_e^T \mathbf{R}_Q \mathbf{Q}_e + \boldsymbol{\omega}^T \mathbf{R}_\omega \boldsymbol{\omega} + \mathbf{u}^T \mathbf{R}_u \mathbf{u}) dt \quad (16)$$

$$\text{s.t. (3), (15)} \quad (17)$$

where  $\mathbf{R}_Q$ ,  $\mathbf{R}_\omega$  and  $\mathbf{R}_u$  are the positive definite weighting matrices,  $\mathbf{k} = [k_q \ k_w]^T$  are the decision variables.

According to [13], problem (16)-(17) is an optimal parameter selection problem, which is, in fact, a non-linear program and solved by the standard non-linear program methods, such as the sequential quadratic program (SQP) method.

In this paper, SQP method is adopted. Since it is a gradient based method, then the gradient formula for the objective function (16) is required, which is given by the following theorem. The proof of the theorem can be obtained based on Theorem 5.2.1 in [13] and it is omitted here for brevity.

**Theorem 1.** *The gradient formula for the objective function (16) is*

$$\frac{\partial J}{\partial \mathbf{k}} = \int_0^{t_f} \frac{\partial H_0(\mathbf{Q}(t), \boldsymbol{\omega}(t), \mathbf{k}, \boldsymbol{\lambda}_0(t))}{\partial \mathbf{k}} dt \quad (18)$$

where  $H_0(\mathbf{Q}(t), \boldsymbol{\omega}(t), \mathbf{k}, \boldsymbol{\lambda}_0(t))$  is the Hamilton function given by

$$H_0 = \mathbf{Q}_e^T \mathbf{R}_Q \mathbf{Q}_e + \boldsymbol{\omega}^T \mathbf{R}_\omega \boldsymbol{\omega} + \mathbf{u}^T \mathbf{R}_u \mathbf{u} + \frac{1}{2} \boldsymbol{\lambda}_{01}^T \mathbf{Q} \otimes \tilde{\boldsymbol{\omega}} + \boldsymbol{\lambda}_{02}^T \mathbf{I}^{-1} (-\boldsymbol{\omega} \times \mathbf{I} \boldsymbol{\omega} + \mathbf{u})$$

$$\frac{\partial H_0}{\partial \mathbf{k}} = -2 \left( \mathbf{u}^T \mathbf{R}_u + \boldsymbol{\lambda}_{02}^T \mathbf{I}^{-1} \right) \left[ \text{Vec} \left[ \left( l_{g1} \mathbf{Q}_d^* - \frac{1}{\delta} \|\mathbf{Q}_e\|^2 \mathbf{L}_{g2}^* \right) \otimes \mathbf{Q} \right] - \boldsymbol{\omega} \right]$$

and the costate vector  $\boldsymbol{\lambda}_0(t) = [\boldsymbol{\lambda}_{01}^T \ \boldsymbol{\lambda}_{02}^T]^T$  is obtained by the following differential equations:

$$\dot{\boldsymbol{\lambda}}_0(t) = -\frac{\partial H_0(\mathbf{Q}(t), \boldsymbol{\omega}(t), \mathbf{k}, \boldsymbol{\lambda}_0(t))}{\partial \mathbf{x}(t)} \quad (19)$$

with the boundary condition

$$\boldsymbol{\lambda}_0(t_f) = 0$$

$$\text{Here } \mathbf{x}(t) = [\mathbf{Q}^T(t) \ \boldsymbol{\omega}^T(t)]^T, \ \frac{\partial H_0}{\partial \mathbf{x}} = \left[ \left( \frac{\partial H_0}{\partial \mathbf{Q}} \right)^T \ \left( \frac{\partial H_0}{\partial \boldsymbol{\omega}} \right)^T \right]^T,$$

$$\frac{\partial H_0}{\partial \mathbf{Q}} = 2 \mathbf{R}_Q \mathbf{Q}_e + \frac{1}{2} \left( \begin{bmatrix} \mathbf{c} \\ 1 \end{bmatrix} \otimes \tilde{\boldsymbol{\omega}} \right)^T \boldsymbol{\lambda}_{01}$$

$$\frac{\partial H_0}{\partial \boldsymbol{\omega}} = 2 \mathbf{R}_\omega \boldsymbol{\omega} + \frac{1}{2} \left( \mathbf{Q} \otimes \begin{bmatrix} \mathbf{c} \\ 0 \end{bmatrix} \right)^T \boldsymbol{\lambda}_{01} + (-\mathbf{c} \times \mathbf{I} \boldsymbol{\omega} - \boldsymbol{\omega} \times \mathbf{I} \mathbf{c})^T (\mathbf{I}^{-1})^T \boldsymbol{\lambda}_{02}$$

$$\text{and } \mathbf{c} = [1 \ 1 \ 1]^T.$$

## 5 Simulation Results

We assume that the spacecraft carries a single light-sensitive element along the  $z$  axis of the spacecraft's body frame. In addition, a high gain antenna is mounted on the spacecraft and its scope is along the  $y$  axis of body frame. The inertia matrix is set as  $\mathbf{I} = \text{diag}(10, 12, 14)$ . The initial and desired attitude quaternion are set as  $\mathbf{Q}(0) = [-0.5386, 0, 0.2436, 0.8066]$ ,  $\mathbf{Q}_d = [0.1178, 0, 0.5065, 0.8541]$ . The spacecraft's sensor needs to avoid the influence of three celestial bodies and maintain communication with a ground station, the related vectors  $\mathbf{x}_1^j, \mathbf{x}_2$  and angles  $\theta_{min}^j, \theta_{max}^j (j = 1, 2, 3)$  of attitude constraints in (4) and (6) are listed in Table 1. The weight parameters of the objective function are selected as  $\mathbf{R}_Q = \mathbf{E}_3$ ,  $\mathbf{R}_\omega = \mathbf{R}_u = 10\mathbf{E}_3$ .

**Table 1.** Attitude constraints parameters

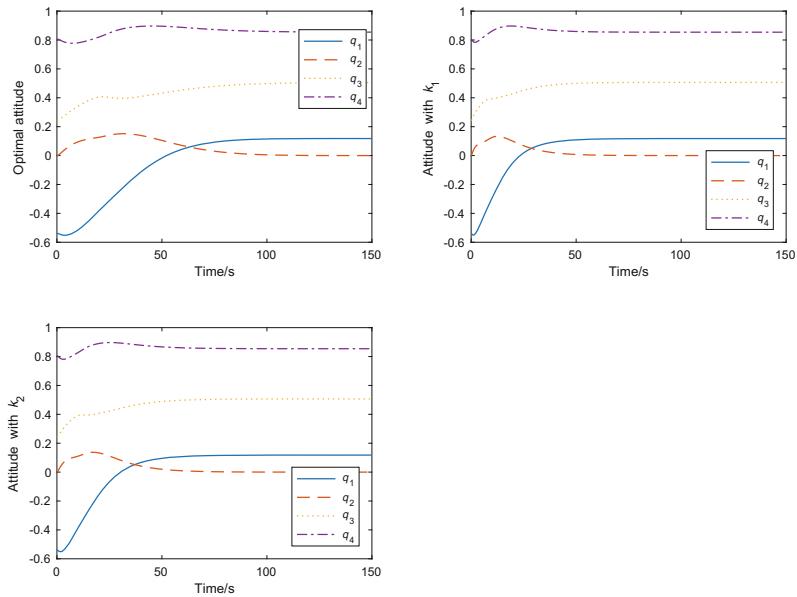
	Related unit vector	Related angle
Communication constraint	$\mathbf{x}_2 = [0.2087, -0.2726, 0.3488]$	$\theta_{max} = 80 \text{ deg}$
	$\mathbf{x}_1^1 = [-0.5713, -0.6509, 0.5000]$	$\theta_{min}^1 = 40 \text{ deg}$
Sensor constraint	$\mathbf{x}_1^2 = [0.94, 0, 0.342]$	$\theta_{min}^2 = 30 \text{ deg}$
	$\mathbf{x}_1^3 = [0, 0.866, 0.5]$	$\theta_{min}^3 = 30 \text{ deg}$

For comparison, we fix  $\mathbf{k} = [k_q, k_\omega]$  with two different values, which are denoted as  $\mathbf{k}_1$  and  $\mathbf{k}_2$ , and then compare the corresponding function values and trajectories with the optimal solution. The details are listed in Table 2, and the trajectories of attitude quaternion  $\mathbf{Q}$ , angular velocity  $\boldsymbol{\omega}$  and control torque  $\mathbf{u}$  are plotted in Figs. 2 and 3, respectively.

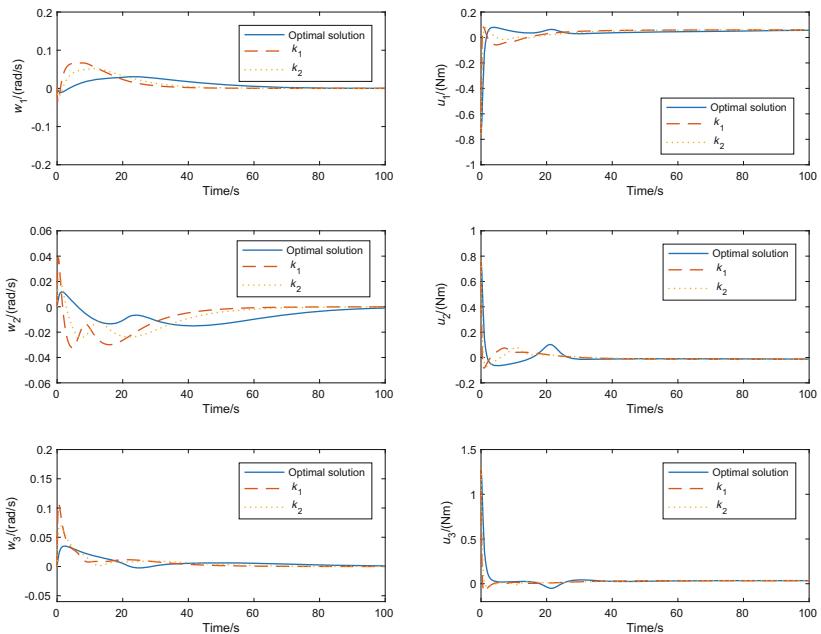
**Table 2.** Control parameters

	$k_1$	$k_2$	Optimal solution
$k_q, k_\omega$	10,18	5,7	1.75,1.48
Function value $J$	6.44	2.76	1.57

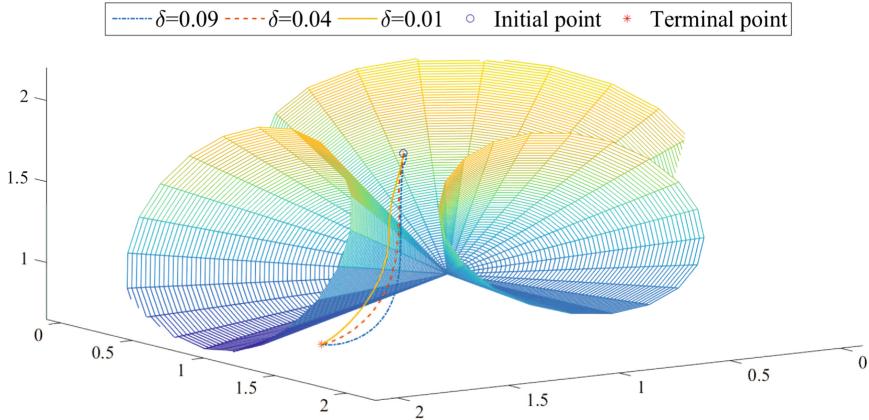
As expected, the function value of the optimal solution is much smaller than those of the fixed  $\mathbf{k}$  in Table 2. In order to verify the design flexibility of the parameter  $\delta$  in (8), we plot the sensor pointing trajectories with different  $\delta$  in Figs. 4 and 5. It can be easily seen that by increasing  $\delta$ , we can make the trajectory far from the forbidden area, which provides more flexibility for our design.



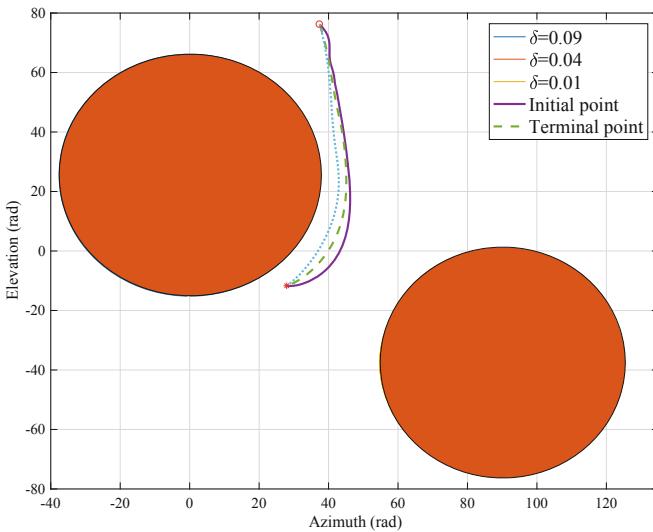
**Fig. 2.** The attitude with optimal solution,  $k_1$  and  $k_2$



**Fig. 3.** The angular velocity and control torque



**Fig. 4.** Sensor pointing trajectory in 3-D



**Fig. 5.** Sensor pointing trajectory in 2-D

## 6 Conclusions

In this paper, an attitude control problem for spacecraft reorientation has been investigated. A potential function based attitude control law was proposed to ensure the stability of the system and feasibility of the attitude constraints. An optimal parameter tuning method was developed for determining the control parameters. As shown in the numerical simulations, the performance of the tuned controller outperform the ones without carrying out optimization. In addition, by adjusting the design parameter in the potential function, the sensor point

trajectory behaves differently in the simulations, which provides more flexibility during the controller design.

## References

1. Kjellberg, H.C., Lightsey, E.G.: Discretized constrained attitude pathfinding and control for satellites. *J. Guid. Control Dynam.* **36**(5), 1301–1309 (2013)
2. Kjellberg, H.C., Lightsey, E.G.: Discretized quaternion constrained attitude pathfinding. *J. Guid. Control Dyn.* **39**(3), 713–718 (2016)
3. Tanygin, S.: Fast three-axis constrained attitude pathfinding and visualization using minimum distortion parameterizations. *J. Guid. Control Dyn.* **38**(12), 2324–2336 (2015)
4. Melton, R.G.: Hybrid methods for determining time-optimal, constrained spacecraft reorientation maneuvers. *Acta Astronaut.* **94**(1), 294–301 (2014)
5. Kim, Y., Mesbah, M., Singh, G., Hadaegh, F.Y.: On the convex parameterization of constrained spacecraft reorientation. *IEEE Trans. Aerosp. Electron. Syst.* **46**(3), 1097–1109 (2010)
6. Tam, M., Glenn Lightsey, E.: Constrained spacecraft reorientation using mixed integer convex programming. *Acta Astronaut.* **127**, 31–40 (2016)
7. De Angelis, E.L., Giulietti, F., Avanzini, G.: Single-axis pointing of underactuated spacecraft in the presence of path constraints. *J. Guid. Control Dyn.* **38**(1), 143–147 (2014)
8. Xu, R., Wang, H., Xu, W., Cui, P., Zhu, S.: Rotational-path decomposition based recursive planning for spacecraft attitude reorientation. *Acta Astronaut.* **143**, 212–220 (2018)
9. McInnes, C.R.: Large angle slew maneuvers with autonomous sun vector avoidance. *J. Guid. Control Dyn.* **17**(4), 875–877 (1994)
10. Lee, U., Mesbah, M.: Feedback control for spacecraft reorientation under attitude constraints via convex potentials. *IEEE Trans. Aerosp. Electron. Syst.* **50**(4), 2578–2592 (2014)
11. Shen, Q., Yue, C., Goh, C.H.: Velocity-free attitude reorientation of a flexible spacecraft with attitude constraints. *J. Guid. Control Dyn.* **40**(5), 1293–1299 (2017)
12. Cheng, Y., Ye, D., Sun, Z.S.: Spacecraft reorientation control in presence of attitude constraint considering input saturation and stochastic disturbance. *Acta Astronaut.* **144**, 61–68 (2018)
13. Teo, K.L., Goh, C.J., Wong, K.H.: A Unified Computational Approach for Optimal Control Problems. Longman, New York (1991)



# Active Detection Based Mobile Robot Radioactive Source Localization Method and Data Processing

Han Gao, Zhengguang Ma, and Yongguo Zhao<sup>(✉)</sup>

Institute of Automation, Qilu University of Technology (Shandong Academy of Sciences), Shandong Provincial Key Laboratory of Robot and Manufacturing Automation Technology, Jinan Shandong 250014, China

sduzyg@sdas.org

**Abstract.** In view of the frequent occurrence of radioactive source loss or out-of-control events, mobile robots that replace manual search for radioactive sources are increasingly subject to market demand. In this article, a mobile robot, which can be used for locating the radioactive source autonomously is developed. Against whether the nuclide information of the radioactive source is known, the corresponding localization method is designed separately. Because the collected data have interference errors during the robot detection process, this article proposes a filtering algorithm that sliding average filtering algorithm combines a variable forgetting factor. Simulation results show that, the improved filtering algorithm can better filter out the interference data in the radioactive source data. The accuracy of the localization of the out-of-control radioactive source can be further improved. The algorithm can meet the design requirements for mobile robots, which can replace humans to search and locate the out-of-control radioactive source.

**Keywords:** Radioactive source · Robot · Sliding average filtering · Variable forgetting factor

## 1 Introduction

In recent years, with the vigorous development of the nuclear industry worldwide, radioactive sources have been widely used in industrial, agricultural, military, medical and other fields. The safety problems in the process of use also are becoming increasingly prominent. Incidents of radioactive source lost or out of control have frequently occurred. Due to the unique radioactive hazards of radioactive sources, once they were out of control, they could directly or indirectly endanger human health and cause social panic [1–3]. Therefore, it is very meaningful to develop a mobile robot for automatic detection and disposal of radioactive sources [4–7], the mobile robot can replace humans to enter the area where the uncontrolled radioactive sources may be located.

Data processing is the key to locate the uncontrolled radioactive sources precisely [8]. During the detection process of the robot, its motion state may

appear rapid acceleration, deceleration, sharp turns and rough road conditions, etc. There are many dynamic interference factors, which result in large errors in the collected radioactive source detection data. It is difficult to solve such problems through the hardware system, but software filtering can effectively reduce the interference of error data and have better stability. In order to filter out the noise mixed in the radioactive source data, digital filtering has been extensively used. Common digital filtering methods include median filtering [9], arithmetic average filtering [10], and sliding average filtering [11, 12].

Based on this, this paper designs a robot that can autonomously locate out-of-control radioactive source, and proposes a new filtering algorithm that combines a sliding average filter algorithm with a variable forgetting factor [13, 14] to improve the positioning accuracy of radioactive sources.

## 2 Robot System Structure

The robot system is composed of a robot mobile chassis, a robot control module, a wireless network and data transmission module, a radioactive source detection module and a remote control terminal. Inertial sensors and other instruments are essential for intelligent mobile platforms, it can realize the robot's self-positioning. The robot receives the control command from the remote control terminal, further, the corresponding equipment executes the specified operations through the robot control module. In addition, the mobile robot feeds back all kinds of information collected by sensors to the remote control terminal, the mobile robot part is shown in Fig. 1(a).

The remote control terminal, which is a convenient platform for human-computer interaction, can communicate with the mobile robot by wireless network and data transmission module. The control command can be sent to the robot control module to complete the corresponding task including movement, information collection and camera rotation et al. The remote control terminal screen is designed with a real-time display for the robot's running status and sensor data. The remote control terminal is shown in Fig. 1(b).



(a) The structure of mobile robot

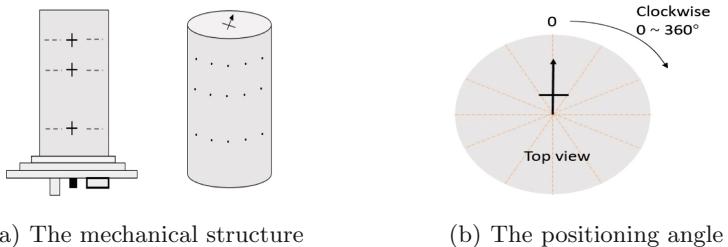


(b) The remote control terminal

**Fig. 1.** The system of the mobile robot

The radioactive source detection module is equipped with radioactive source searching detector, aiming detector, interface module and data processing

software. Among them, radioactive source searching detector is a sensor for detecting radioisotopes. The mechanical structure of the detector is a cylinder, the mechanical structure of the detector is shown in Fig. 2(a). The center of the detector is equipped with a shielded collimator measurement system. The detector uses these photon counters to measure the absorbed dose rate of air in the environment. According to the number of accumulated absorbed photons in different azimuth photon counters judges the absorbed dose rate of gamma in the air, and measures the angle corresponding to the maximum  $\gamma$  air absorbed dose rate. It provides accurate data for the subsequent positioning of the radioactive source. The detection angle of the radiation source of the detector is  $0 \sim 360^\circ$  clockwise from the direction of the arrow on the top of the sensor, as shown in Fig. 2(b).



**Fig. 2.** The radioactive source searching detector

### 3 Principles of Radioactive Source Localization

In the actual detection process, the method of locating the radioactive source depends on whether the nuclide information of the radioactive source is known. Aiming at these two situations, the corresponding method is designed to achieve the accurate localization of the radioactive sources.

#### 3.1 Radioactive Source with Known Nuclide Information

When the nuclide information of the out-of-control radioactive source is known, the activity of radioactive source can be estimated by the activity of the pre-lost radioactive source  $A_0$ , the half-life of radioactive source  $T_{\frac{1}{2}}$  and the lost time  $t$ :

$$A = A_0 e^{-\frac{0.693t}{T_{1/2}}} \quad (1)$$

where  $A$  is the activity of the radioactive source after loss, which represents the number of core decay per unit time. The unit of activity of the radioactive source  $A$  is  $Bq$ .

For the radioactive source searching detector, during its work, at time  $t_1$ , the output of the detector is the air absorbed dose rate  $D_1(t_1)$  of the robot and the angle  $\theta_1(t_1)$  of the radioactive source relative to the robot. The distance between the radioactive source and the robot has the following relationship:

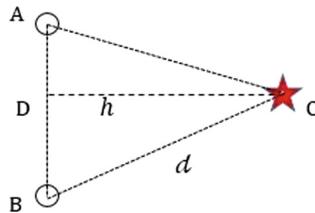
$$R = \sqrt{\frac{K_y \cdot A}{D}} \quad (2)$$

where  $K_y$  is the air kerma constant of the radioactive source, its size depends on the photon energy, the activity of the radioactive source, the shape of the radioactive source and the distance from the radioactive source. For a point source, its exposure rate in air depends solely on the nature of the source. The unit of the air kerma constant  $K_y$  is  $Gy \cdot m^2 \cdot Bq^{-1} \cdot h^{-1}$ .

According to the outputs of the detector and Eq. (2), the relative angle  $\theta_1$  and distance  $R_1$  between the radioactive source and the robot are known. Then, the position of the point  $P_1$  of the radioactive source can be determined by polar coordinates. With the continuous movement process of the mobile robot, at time  $t_1, \dots, t_n$ , the detector obtains the position information of a series of radioactive sources  $P_1, \dots, P_n$ .

### 3.2 Radioactive Source with Unknown Nuclide Information

When the type or time of loss of the out-of-control source is unknown, neither the activity information of the radioactive source nor the corresponding air kerma constant can be obtained. It is no longer possible to use Eq. (2) to calculate the distance between the radioactive source and the robot. At this time, only the angle information between the radioactive source and the robot is available, so the position of the radioactive source cannot be determined. In this case, the position of the radioactive source is determined by the triangulation method. The principle is shown in Fig. 3. Point  $A$  is the initial position of the robot, and point  $O$  is the position of the radioactive source.



**Fig. 3.** Triangle locating method for radioactive source

At time  $t_1$ , the robot starts to execute the detection work at point  $A$  and the angle of the radioactive source returned by the detector relative to the robot is  $\theta_1$ . The robot continues to move in a straight line at a certain speed. At time  $t_2$ , the robot reaches point  $B$ , and the angle of the radioactive source returned by the detector relative to the robot is  $\theta_2$ . The distance  $l$  of the line segment  $AB$  can be obtained by calculating the speed and running time of the robot.

According to the triangle correlation theorem, the length  $h$  of the high  $OD$  on the side of the triangle  $AB$  can be obtained:

$$h = \frac{l \cdot \tan(\theta_1) \cdot \tan(\theta_2)}{\tan(\theta_1) + \tan(\theta_2)} \quad (3)$$

According to theorem of right triangle, the length  $d$  of the triangle  $OB$  side can be calculated. The distance  $d$  between the robot and the radioactive source:

$$d = \frac{h}{\sin(\theta_2)} = \frac{l \cdot \tan(\theta_1) \cdot \tan(\theta_2)}{\sin(\theta_2)(\tan(\theta_1) + \tan(\theta_2))} \quad (4)$$

Control the robot to move to point  $O$  in the direction of  $BO$ , point  $O$  is the position of the radioactive source.

## 4 Data Filtering

In the process of detecting radioactive sources, there are many interference factors. In order to improve the measurement accuracy, the data needs to be digitally filtered. The filtered data can be used for related operations. Common filtering methods include median filtering, arithmetic average filtering and sliding average filtering.

The median filtering: firstly, the measured data is continuously sampled  $N$  times, then, the sampled data is sorted according to size, finally the intermediate value is taken as the effective value. This algorithm is simple in operation, it effectively overcomes the fluctuation interference. But it is not suitable for filtering data in a rapidly changing process.

The arithmetic average filtering: firstly, the measured data is continuously sampled  $N$  times, then, the sampled values are arithmetically averaged, finally the average value obtained is taken as the effective value. This algorithm is very effective for filtering random errors mixed in the measured data, but it is not easy to eliminate the errors caused by pulse interference.

The sliding average filtering: firstly, the  $N$  consecutively acquired measurement values are temporarily stored in a queue, where the length of the queue is a fixed value  $N$ . Each time a new data is updated, the data at the head of the queue is first removed, the data queue is moved forward by one whole, then the updated data is placed at the end of the queue, finally the effective data are obtained from the arithmetic average of the updated data in the queue. The formula of the algorithm is shown in Eq. (5):

$$\bar{X}_i = \frac{1}{N} \sum_{i=0}^{N-1} X_i \quad (5)$$

Where  $\bar{X}_i$  is the filtered output of the  $i$ -th sample,  $X_i$  is the unfiltered  $i$ -th sample value, and  $N$  is the number of sliding average terms. The algorithm has high smoothness and low sensitivity, but it has a poor suppression effect on accidental impulsive interference.

#### 4.1 Improved Sliding Average Filtering Algorithm

In the sliding average filtering algorithm, the value of adjacent sampling points have the same contribution for the calculation. However, during the radioactive source detection task, when the distance between the robot and the radioactive source continues to decrease, the accuracy of the detector is gradually improved, so the accuracy of the collected data is higher, it is not reasonable to continue to use this equal weight smoothing formula. In order to better overcome the adverse effects of errors, a new algorithm is designed, which combines sliding average filtering with a variable forgetting factor. The key of this algorithm is to appropriately select the proportion of the data. That is, the data at different times are multiplied by a different forgetting factor. The formula of this algorithm is shown in Eq. (6):

$$\bar{X}_i = \sum_{i=0}^{N-1} \lambda_i X_i \quad (6)$$

Where  $\lambda_i$  is a variable forgetting factor, and  $\lambda_0 + \lambda_1 + \dots + \lambda_{N-1} = 1$ .  $\lambda_i$  can be calculated with the function  $W(i)$ :

$$\lambda_i = \frac{W(i)}{\sum_{i=0}^{N-1} W(i)} \quad (7)$$

According to the technical performance of the detector, when the distance between the robot and the radioactive source gradually decreases, the accuracy of the detector is higher, and the air absorption dose rate  $D$  and the angle information  $\theta$  are more accurate. The size of forgetting factor is closely related to the change of dose rate. The correlation between the two can be expressed by the following function:

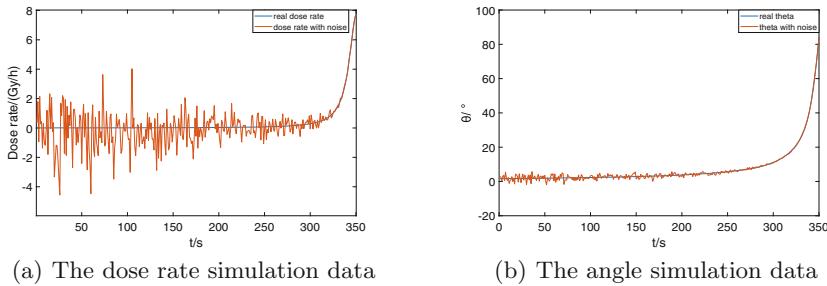
$$W(i) = \frac{1}{1 + e^{-\frac{1}{30} D_i}} \quad (8)$$

It can be seen from the above formula, the variable forgetting factor is positively related to the change of the air absorbed dose rate. When  $D_i$  increases,  $\lambda$  increases, otherwise,  $\lambda$  decreases.  $\lambda$  changes between  $[0, 1]$ , and the change is slow, which can buffer the error problem caused by pulse sharp interference. If  $\lambda$  is closer to 1, then the proportion of new data will be larger, the faster the old data is forgotten, the data has a better tracking effect.

### 5 Simulation and Analysis

In this simulation, the radioactive source  $^{60}Co$  is taken as the simulation object. The air kerma constant is  $13.2 \text{ Gy} \cdot \text{m}^2 \cdot \text{Bq}^{-1} \cdot \text{h}^{-1}$ , and the half-life of the nuclide is 5.27 years. Assuming that the activity of the radioactive source before losing is  $100 \text{ Bq}$ , the loss time is 4.0 years, then the current activity is  $59.1 \text{ Bq}$ . Considering the situation that the radioactive source is on a plane, the initial position of the robot is assumed to be set at the origin  $(0, 0)$ . For convenience, the radioactive source can be placed at a given position, namely at coordinates  $(10, 350)$ .

Combining with the principle of radioactive source detection, the raw data of the air absorption dose rate  $D$  of the detector and the relative angle  $\theta$  with the robot are generated. The robot is disturbed by uncertain random factors. According to the characteristics of the smaller the distance between the detector and the detected object, the higher the detection accuracy and the smaller the error, the Gaussian white noise is added to the raw data for simulation experiments to verify the superiority of the improved algorithm filtering effect, in which the added Gaussian white noise gradually weakens as the distance between the robot and the radioactive source decreases. The simulation data of dose rate and angle are shown in Fig. 4(a) and Fig. 4(b) respectively.

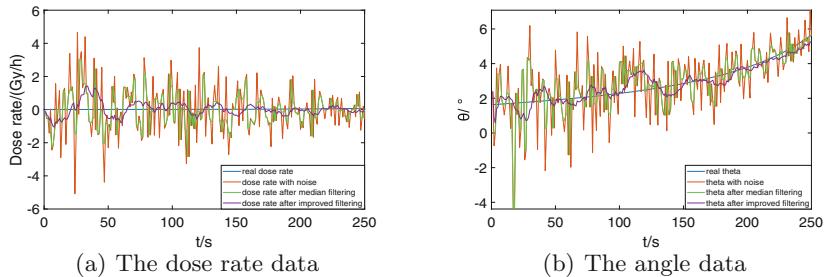


**Fig. 4.** The simulation data

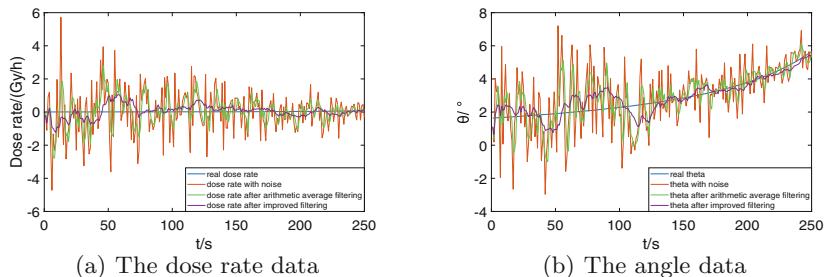
Through MATLAB software, apply median filtering, arithmetic average filtering and improved filtering algorithm to simulate respectively, and draw the change trend of the filtered data. Among them, Fig. 5, Fig. 6 are the simulation result graphs of median filtering, arithmetic average filtering compared with improved filtering algorithm respectively.

It can be seen from Fig. 5 that the median filtering algorithm can overcome the fluctuation interference caused by accidental factors to a certain extent, and weaken the impulse noise with a large peak value. But this algorithm has a relatively poor suppression effect on the overall noise data. Combined with Fig. 6, it can be seen that the arithmetic average filtering algorithm is better than the median filtering algorithm in suppressing noisy data, but the algorithm has a relatively poor suppression effect on peak errors, and there is still peak noise after filtering. After filtering, the smoothness of the data curves of the two algorithms is relatively low. However, the improved filtering algorithm can effectively filter out the random noise mixed in the radioactive source detection data, and the accidental peak interference noise is also effectively suppressed. After the filtering process, the smoothness of the data curve is significantly higher than the median filtering and arithmetic average filtering. The processed data by the improved filtering algorithm and the real data maintain a uniform error, which can reflect the detailed information of the real data and provide accurate data support for further positioning of the radioactive source. Therefore, the improved filtering

algorithm adopted in this paper has a better filtering effect and the radioactive source localization is more accurate.



**Fig. 5.** Median filtering compared with improved filtering algorithm



**Fig. 6.** Arithmetic average filtering compared with improved filtering algorithm

## 6 Conclusion

This paper designs a mobile robot system, which can autonomously locate a out-of-control radioactive source. The robot can be remotely controlled through the wireless communication network, and the remote control terminal can display the robot working scene environment in real time. Aiming at the two cases of whether the nuclide information of the radioactive source is known, this paper designs two localization methods. In order to reduce the influence of noisy data and improve the localization accuracy, this paper proposes a filtering algorithm that combines the sliding average filtering algorithm with the variable forgetting factor. This algorithm can effectively filter out the noise interference in the detected data, accurately track the change of the real data. In the future, it can be applied to the actual localization of radioactive source.

**Acknowledgments.** The research is supported by key Research and Development Program of Shandong Province(2017CXGC0916) and Shandong Academy of Sciences Collaborative Innovation Fund(2018CXY-3) and Youth Science Funds of Shandong Academy of Sciences(2019QN0015) and Shandong key research and development

plan (Public welfare science and technology research) (2019GGX104005) and Shandong key research and development plan(Major scientific and technological innovation projects)(2019JZZY010430). Yongguo Zhao is the corresponding author (e-mail: sduzyg@sdas.org).

## References

1. Liu, C.Z., Zhi, Y., Deng, J.S., et al.: Study on accident response robot for nuclear power plant and analysis of key technologies. *Nucl. Sci. Eng.* **33**(1), 97–105 (2013)
2. Kim, I.S., Choi, Y., Jeong, K.M.: A new approach to quantify safety benefits of disaster robots. *Nucl. Eng. Technol.* **49**(7), 1412–1422 (2017)
3. Bogue, R.: Robots in the nuclear industry: a review of technologies and applications. *Ind. Robot.* **38**(2), 113–118 (2011)
4. Nagatani, K., Kiribayashi, S., Okada, Y., et al.: Emergency response to the nuclear accident at the fukushima daiichi nuclear power plants using mobile rescue robots. *J. Field Robot.* **30**(1), 44–63 (2013)
5. Li, Z.I., Li, C., Zeng, G.Q.: The mobile radiation monitoring platform and terminal design. *Nucl. Electron. Detect. Technol.* **33**(7), 813–816 (2013)
6. Jiang, M., Zhao, Y., Liu, G., et al.: The design of remote terminal for searching radiation source robot. In: 2017 Chinese Automation Congress (CAC), pp. 793–796. IEEE, Jinan (2017)
7. Zhu, L., Zhao, Y., Ma, Z., et al.: Design and implementation of the robot for radiation source detection and disposal. In: 2019 Chinese Automation Congress (CAC), pp. 3431–3434. IEEE, Hangzhou (2019)
8. Matía, F., Jiménez, A.: Multisensor fusion: an autonomous mobile robot. *J. Intell. Robot. Syst.* **22**(2), 129–141 (1998)
9. Storath, M., Weinmann, A.: Fast median filtering for phase or orientation data. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(3), 639–652 (2018)
10. Deivalakshmi, S., Palanisamy, P.: Improved tolerance based selective arithmetic mean filter for detection and removal of impulse noise. In: 2010 5th International Conference on Industrial and Information Systems, pp. 309–313. IEEE, Mangalore (2010)
11. Ren, L., Xu, T.H.: Research on smoothing filtering algorithm of BDS/GPS slow deformation monitoring sequence. In: 9th China Satellite Navigation Conference, pp. 33–44. Springer, Harbin (2018)
12. Belge, M., Miller, E.L.: A sliding window RLS-like adaptive algorithm for filtering alpha-stable noise. *IEEE Signal Process. Lett.* **7**(4), 86–89 (2000)
13. White, K.G.: Forgetting functions. *Animal Learn. Behav.* **29**(3), 193–207 (2001)
14. Paleologu, C., Benesty, J., Ciochina, S.: A robust variable forgetting factor recursive least-squares algorithm for system identification. *IEEE Signal Process. Lett.* **15**, 597–600 (2008)



# Distributed Adaptive Consensus Tracking Control for Multiple AUVs with State Constraints

Jingzi Fan and Lin Zhao<sup>(✉)</sup>

Qingdao University, Qingdao 266071, China  
zhaolin1585@163.com

**Abstract.** The problem of the adaptive consensus tracking for multiple AUV (autonomous underwater vehicle) systems with state constraints is studied in this paper. A command filtered backstepping control scheme with the positions state constraints based on the neural network is proposed. The unknown nonlinear dynamics are approximated by a neural network in this scheme. It can make the state of following AUV tracks the leader's state with acceptable accuracy. Not only do we eliminate the errors caused by filtering, but we can also solve complex calculation problems well. Through simulation specific examples, proved the feasibility of the control scheme.

**Keywords:** Multiple AUV systems · Command filtered backstepping · State constraints · Neural network

## 1 Introduction

Due to the distributed collaborative control of multiple AUVs in the field of deep sea exploration and oceanographic research, it has received extensive attention [1]. Recently, many control schemes for instance the backstepping control schemes, the homogeneous control schemes, and the adaptive control schemes have been proposed to achieve the anticipant control performance for multiple AUVs [2–6].

In the backstepping control scheme designed in [2,3], The virtual control signal is the state of AUV, we need to calculate its derivative at same time, which will lead to the problem of computationally complex. Although the dynamic surface control proposed later solves the above problems well [7], new errors are generated due to the use of filters. The command filtered backstepping control scheme proposed in [4], solved the problems in traditional backstepping control and dynamic surface control while ensuring the control effect. But in the actual situation, the multiple AUV systems has a high probability to be in a constrained state. This is not considered by the current control scheme.

Based on existing research results, We will solve the tracking control problem of multiple AUV systems with position state constraints. A command filtering backstepping control scheme with position state constraints is proposed to ensure that the tracking error of position state converges to the expected neighborhood.

© The Editor(s) (if applicable) and The Author(s), under exclusive license

to Springer Nature Singapore Pte Ltd. 2021

Y. Jia et al. (Eds.): CISC 2020, LNEE 705, pp. 317–325, 2021.

[https://doi.org/10.1007/978-981-15-8450-3\\_34](https://doi.org/10.1007/978-981-15-8450-3_34)

## 2 System Descriptions and Preliminaries

We consider a system with a weighted directed graph  $\bar{\mathcal{G}}$  to describe the interaction between them. This system has one leader and  $N$  AUVs, and all the AUVs have fixed attitudes [4]. Denote the translational dynamics about  $i$ -th following AUV ( $i \in \mathcal{V}, \mathcal{V} = \{1, 2, \dots, N\}$ ) as

$$\begin{aligned}\dot{p}_i &= R_i(\Theta_i)v_i \\ M_i\dot{v}_i &= -D_i v_i - g_i(\Theta_i) + \tau_i \\ y_i &= p_i\end{aligned}\tag{1}$$

where  $p_i = [p_{x_i}, p_{y_i}, p_{z_i}]^T$  is the position vector,  $\Theta_i = [\phi_i, \theta_i, \psi_i]^T$  is the attitude vector.  $y_i$  is the output vector,  $v_i = [u_i, \nu_i, \omega_i]^T$  is translational velocity vector in the body-fixed reference frame,  $M_i$  is the inertia matrix,  $D_i(v_i)$  is the damping matrix,  $g_i(\Theta_i)$  is the restoring force vector, and  $\tau_i \in \mathbf{R}^3$  is the control force vector.  $R_i(\Theta_i)$  is the kinematic transformation matrix, and defined as

$$R_i(\Theta_i) = \begin{bmatrix} c_{\psi_i}c_{\theta_i} & -s_{\psi_i}c_{\phi_i} + s_{\phi_i}s_{\theta_i}c_{\psi_i} & s_{\psi_i}s_{\phi_i} + s_{\theta_i}c_{\psi_i}c_{\phi_i} \\ s_{\psi_i}c_{\theta_i} & c_{\psi_i}c_{\phi_i} + s_{\phi_i}s_{\theta_i}s_{\psi_i} & -c_{\psi_i}s_{\phi_i} + s_{\theta_i}s_{\psi_i}c_{\phi_i} \\ -s_{\theta_i} & s_{\phi_i}c_{\theta_i} & c_{\phi_i}c_{\theta_i} \end{bmatrix}$$

here  $s_\alpha = \sin \alpha, c_\alpha = \cos \alpha$ .  $M_i = \text{diag}\{m_{i1}, m_{i2}, m_{i3}\}, D_i(v_i) = \text{diag}\{d_{L_{i1}} + d_{Q_{i1}}|u_i|, d_{L_{i2}} + d_{Q_{i2}}|\nu_i|, d_{L_{i3}} + d_{Q_{i3}}|\omega_i|\}, m_{ij}, d_{L_{ij}}, d_{Q_{ij}} > 0, j = 1, 2, 3, g_i(\Theta_i) = [(W_i - B_i)s_{\theta_i}, -(W_i - B_i)c_{\theta_i}s_{\phi_i}, -(W_i - B_i)c_{\theta_i}c_{\phi_i}]^T$  where  $W_i$  stands for gravity,  $B_i$  stands for buoyancy forces. We also need to pay attention to  $R_i R_i^T = I$ .

**Assumption 1.**  $d_{L_{ij}}, d_{Q_{ij}}, W_i$  and  $B_i$  are unknown bounded constants.

Find the derivative of  $\dot{p}_i$  with respect to time, assume  $x_{i,1} = p_i$  and  $x_{i,2} = \dot{p}_i$ , we can get the following equation:

$$\begin{aligned}\dot{x}_{i,1} &= x_{i,2} \\ \dot{x}_{i,2} &= f_i + \bar{\tau}_i \\ y_i &= x_{i,1}\end{aligned}\tag{2}$$

where  $f_i = \dot{R}_i v_i - R_i M_i^{-1} D_i v_i - R_i M_i^{-1} g_i$  and  $\bar{\tau}_i = R_i M_i^{-1} \tau_i$ .

**Assumption 2.**  $\bar{\mathcal{G}}$  contains a spanning tree and the leader node is the root node.

**Assumption 3.** All the states of the system are constrained in a compact set which  $|x_{i,mq}| \leq c_{i,mq}$  with  $c_{i,mq}$  is a positive constant,  $i \in \mathcal{V}, m = 1, 2, q = 1, 2, 3$ .

**Assumption 4.** The output signal of leader agent  $r(t)$  and its derivative  $\dot{r}(t)$  are smooth, bounded and known functions with  $|r_q| \leq \eta_q, q = 1, 2, 3$ .

### 3 Main Result

The local errors are given as:

$$\begin{aligned} z_{i,1} &= \sum_{j=1}^N a_{ij}(y_i - y_j) + b_i(y_i - r) \\ z_{i,2} &= x_{i,2} - \pi_i, i \in \mathcal{V} \end{aligned} \quad (3)$$

Command filter is designed as:

$$\begin{aligned} \dot{\varphi}_{i,1,\gamma} &= \iota_{i,1,\gamma} \\ \iota_{i,1,\gamma} &= -r_{i,1} |\varphi_{i,1\gamma} - \alpha_{i,1,\gamma}|^{\frac{1}{2}} \text{sign}(\varphi_{i,1,\gamma} - \alpha_{i,1,\gamma}) + \varphi_{i,2,\gamma} \\ \dot{\varphi}_{i,2,\gamma} &= -r_{i,2} \text{sign}(\varphi_{i,2,\gamma} - \iota_{i,1,\gamma}), \gamma = 1, 2, 3 \end{aligned} \quad (4)$$

with  $\pi_i(t) = \varphi_{i,1}(t)$ ,  $\dot{\pi}_i(t) = \iota_{i,1}(t)$  as the output, with the  $\alpha_{i,1}$  as the input [8].

Then we could get that:

$$\|(\pi_i - \alpha_{i,1})\| \leq \varpi_i \quad (5)$$

Define the compensated tracking error signals as:

$$\nu_{i,m} = z_{i,m} - \xi_{i,m}, i \in \mathcal{V}, m = 1, 2 \quad (6)$$

Pseudocontrol signals  $\alpha_{i,1}$  can be written as:

$$\begin{aligned} \alpha_{i,1} &= \frac{1}{(d_i + b_i)} (-k_{i,1}z_{i,1} + \sum_{j=1}^N a_{ij}x_{j,2} + b_i\dot{r}) \\ \alpha_{i,2} &= -k_{i,2}z_{i,2} + \dot{\pi}_i - \frac{1}{2}(K_{bi,211}, K_{bi,222}, K_{bi,233})^T \\ &\quad - (d_i + b_i)\left(\frac{K_{bi,111}v_{i,21}}{K_{bi,211}}, \frac{K_{bi,122}v_{i,22}}{K_{bi,222}}, \frac{K_{bi,133}v_{i,23}}{K_{bi,233}}\right)^T \\ &\quad - \left(\frac{K_{bi,211}^2\hat{\theta}_{i,1}S_{i,1}^TS_{i,1}}{2h_{i,1}^2}, \frac{K_{bi,222}^2\hat{\theta}_{i,2}S_{i,2}^TS_{i,2}}{2h_{i,2}^2}, \frac{K_{bi,233}^2\hat{\theta}_{i,3}S_{i,3}^TS_{i,3}}{2h_{i,3}^2}\right)^T \end{aligned} \quad (7)$$

where  $i \in \mathcal{V}$ . The parameters  $k_{i,m}, m = 1, 2$  are designed positive constants,

$$K_{bi,mq} = \frac{v_{i,m}^T I_q}{k_{bi,m}^T I_q k_{bi,m} - v_{i,m}^T I_q v_{i,m}}, k_{bi,m}$$

will be designed later.

And the error compensating vector is defined as  $\xi_i = [\xi_{i,1}^T, \xi_{i,2}^T]^T$ .

$$\begin{aligned} \dot{\xi}_{i,1} &= -k_{i,1}\xi_{i,1} + (d_i + b_i)(\pi_i - \alpha_{i,1}) + (d_i + b_i)\xi_{i,2} \\ \dot{\xi}_{i,2} &= -k_{i,2}\xi_{i,2} \end{aligned} \quad (8)$$

with  $\xi_{i,m}(0) = 0 (i \in \mathcal{V}, m = 1, 2)$ .

Define a compact set  $\Omega_{v_{i,1q}} = \{|v_{i,1q}| < k_{bi,1q}, q = 1, 2, 3\}$ , then we construct the Lyapunov function as:

$$V_{i,1} = \sum_{q=1}^3 \frac{1}{2} \log \frac{k_{bi,1q}^2}{k_{bi,1q}^2 - v_{i,1q}^2} = \sum_{q=1}^3 \frac{1}{2} \log \frac{k_{bi,1}^T I_q k_{bi,1}}{k_{bi,1}^T I_q k_{bi,1} - v_{i,1}^T I_q v_{i,1}} \quad (9)$$

where  $I_1 = \text{diag}(1, 0, 0)$ ,  $I_2 = \text{diag}(0, 1, 0)$  and  $I_3 = \text{diag}(0, 0, 1)$ .

Then the derivative of  $V_{i,1}$  could be given as

$$\begin{aligned}\dot{V}_{i,1} &= \sum_{q=1}^3 K_{bi,1q}[(d_i + b_i)z_{i,2} + (d_i + b_i)\alpha_{i,1} + (d_i + b_i)(\pi_i - \alpha_{i,1}) \\ &\quad - \sum_{j=1}^N a_{ij}x_{i,2} - b_i\dot{r} - \dot{\xi}_{i,1}]\end{aligned}\tag{10}$$

Substituting  $\alpha_{i,1}$  and  $\dot{\xi}_{i,1}$  into (10), we obtain

$$\dot{V}_{i,1} = -\sum_{q=1}^3 k_{i,1}K_{bi,1q}v_{i,1} + \sum_{q=1}^3 (d_i + b_i)K_{bi,1q}v_{i,2}\tag{11}$$

Construct another Lyapunov function

$$V_{i,2} = V_{i,1} + \sum_{q=1}^3 \frac{1}{2} \log \frac{k_{bi,2q}^2}{k_{bi,2q}^2 - v_{i,2q}^2}\tag{12}$$

Its derivative could be calculate as:

$$\dot{V}_{i,2} = \dot{V}_{i,1} + \sum_{q=1}^3 K_{bi,2q}(f_i + \bar{\tau}_i - \dot{\pi}_i - \dot{\xi}_{i,2})\tag{13}$$

Use the RBF neutral network to approximate  $f_{i,q}$  ( $i \in \mathcal{V}, q = 1, 2, 3$ ) [9]. It can be rewritten as:

$$f_{i,q} = W_{i,q}^T S_{i,q} + \sigma_{i,q}, q = 1, 2, 3\tag{14}$$

where the approximation errors satisfy bounded conditions  $\|\sigma_{i,q}\| \leq \sigma_{i,q}^*$ ,  $\sigma_{i,q}^* > 0$ . We can get the following inequality:

$$K_{bi,2qq}(W_{i,q}^T S_{i,q} + \sigma_{i,q}) \leq \frac{h_{i,q}^2}{2} + \frac{1}{2}K_{bi,2qq}^2 + \frac{(\sigma_{i,q}^*)^2}{2} + \frac{K_{bi,2qq}^2 \|W_{i,q}\|^2 S_{i,q}^T S_{i,q}}{2h_{i,q}^2}\tag{15}$$

where  $h_{i,q} > 0$  is a constant.

Denote  $\theta_{i,q} = \|W_{i,q}\|^2$  ( $i \in \mathcal{V}, q = 1, 2, 3$ ),  $\tilde{\theta}_{i,q} = \theta_{i,q} - \hat{\theta}_{i,q}$ , substituting  $\bar{\tau}_i = \alpha_{i,2}$  and  $\dot{\xi}_{i,2}$  into (13), we have

$$\dot{V}_{i,2} \leq -\sum_{q=1}^3 \sum_{m=1}^2 k_{i,m} K_{bi,mq} v_{i,m} + \sum_{q=1}^3 \frac{h_{i,q}^2 + (\sigma_{i,q}^*)^2}{2} + \sum_{q=1}^3 \frac{K_{bi,2qq}^2 \tilde{\theta}_{i,q} S_{i,q}^T S_{i,q}}{2h_{i,q}^2}\tag{16}$$

Choose the Lyapunov function as:

$$\bar{V} = \frac{1}{2} \sum_{i=1}^N \sum_{m=1}^2 \xi_{i,m}^T \xi_{i,m}\tag{17}$$

Then, we have

$$\dot{\bar{V}} = \sum_{i=1}^N [-k_{i,1}\xi_{i,1}^T\xi_{i,1} + (d_i + b_i)\xi_{i,1}^T(\pi_i - \alpha_{i,1}) + (d_i + b_i)\xi_{i,1}^T\xi_{i,2} - k_{i,2}\xi_{i,2}^T\xi_{i,2}] \quad (18)$$

According to inequality (5), and denote  $\rho_1 = \max\{d_i + b_i, (i \in \mathcal{V})\}$  the derivative of  $\bar{V}$  could be calculate as:

$$\begin{aligned} \dot{\bar{V}} &\leq \sum_{i=1}^N [-k_{i,1}\xi_{i,1}^T\xi_{i,1} - k_{i,2}\xi_{i,2}^T\xi_{i,2} + \frac{\rho_1}{2}(\xi_{i,1}^T\xi_{i,1} + \varpi_i^2) + \frac{\rho_1}{2}(\xi_{i,1}^T\xi_{i,1} + \xi_{i,2}^T\xi_{i,2})] \\ &= -\sum_{i=1}^N (k_{i,1} - \rho_1)\xi_{i,1}^T\xi_{i,1} - \sum_{i=1}^N (k_{i,2} - \frac{\rho_1}{2})\xi_{i,2}^T\xi_{i,2} + \sum_{i=1}^N \frac{\rho_1}{2}\varpi_i^2 \end{aligned} \quad (19)$$

We have:  $\dot{\bar{V}} \leq -\rho_2\bar{V} + \rho_3$ , with  $\rho_2 = \min\{(k_{i,1} - \rho_1), (k_{i,2} - \frac{\rho_1}{2}), i \in \mathcal{V}\}$ ,  $\rho_3 = \sum_{i=1}^N \frac{\rho_1}{2}\varpi_i^2$  and the parameters  $(k_{i,1} - \rho_1), (k_{i,2} - \frac{\rho_1}{2})$  are all need to be positive.

A new Lyapunov function is given as:

$$V = \sum_{i=1}^N V_{i,2} + \bar{V} + \sum_{i=1}^N \sum_{q=1}^3 \frac{\tilde{\theta}_{i,q}^2}{2\lambda_{i,q}} \quad (20)$$

Let's construct the adaptive function of  $\theta_i$  as:

$$\dot{\theta}_{i,q} = -\lambda_{i,q}\mu_{i,q}\hat{\theta}_{i,q} + \frac{\lambda_{i,q}K_{bi,2qq}^2S_{i,q}^T S_{i,q}}{2h_{i,q}^2}, (i \in \mathcal{V}, q = 1, 2, 3) \quad (21)$$

where  $\lambda_{i,q}, \mu_{i,q}$  are all positive constants.

The derivative of  $V$  could be calculate as:

$$\begin{aligned} \dot{V} &= \sum_{i=1}^N \dot{V}_{i,2} + \dot{\bar{V}} - \sum_{i=1}^N \sum_{q=1}^3 \frac{\tilde{\theta}_{i,q}\dot{\hat{\theta}}_{i,q}}{\lambda_{i,q}} \\ &\leq -\sum_{i=1}^N \sum_{m=1}^2 \sum_{q=1}^3 k_{i,m} K_{bi,mq} v_{i,m} + \sum_{i=1}^N \sum_{q=1}^3 \mu_{i,q} \tilde{\theta}_{i,q} \hat{\theta}_{i,q} \\ &\quad + \sum_{i=1}^N \sum_{q=1}^3 \frac{h_{i,q}^2 + (\sigma_{i,q}^*)^2}{2} - \rho_2\bar{V} + \rho_3 \\ &\leq -\sum_{i=1}^N \sum_{m=1}^2 \sum_{q=1}^3 k_{i,m} K_{bi,mq} v_{i,m} - \rho_2\bar{V} \\ &\quad - \sum_{i=1}^N \sum_{q=1}^3 \frac{\mu_{i,q}\lambda_{i,q}\tilde{\theta}_{i,q}^2}{2\lambda_{i,q}} + \rho_3 + \sum_{i=1}^N \sum_{q=1}^3 \frac{h_{i,q}^2 + (\sigma_{i,q}^*)^2 + \mu_{i,q}\theta_{i,q}^2}{2} \end{aligned} \quad (22)$$

because the following inequality holds

$$\log \frac{k_{bi,m}^T I_q k_{bi,m}}{k_{bi,m}^T I_q k_{bi,m} - v_{i,m}^T I_q v_{i,m}} \leq \frac{v_{i,m}^T I_q v_{i,m}}{k_{bi,m}^T I_q - v_{i,m}^T I_q v_{i,m}} \quad (23)$$

We get:

$$\begin{aligned} \dot{V} &\leq - \sum_{i=1}^N \sum_{m=1}^2 \sum_{q=1}^3 k_{i,m} \log \frac{k_{bi,m}^T I_q k_{bi,m}}{k_{bi,m}^T I_q k_{bi,m} - v_{i,m}^T I_q v_{i,m}} \\ &\quad - \rho_2 \bar{V} - \sum_{i=1}^N \sum_{q=1}^3 \frac{\mu_{i,q} \lambda_{i,q} \tilde{\theta}_{i,q}^2}{2 \lambda_{i,q}} + \rho_3 + \sum_{i=1}^N \sum_{q=1}^3 \frac{h_{i,q}^2 + (\sigma_{i,q}^*)^2 + \mu_{i,q} \theta_{i,q}^2}{2} \end{aligned} \quad (24)$$

this obviously gets the following inequality

$$\dot{V} \leq -e_1 V + e_2 \quad (25)$$

where  $e_1 = \min\{2k_{i,m}, \rho_2, \mu_{i,q} \lambda_{i,q}\}$ ,  $e_2 = \sum_{i=1}^N \sum_{q=1}^3 \frac{h_{i,q}^2 + (\sigma_{i,q}^*)^2 + \mu_{i,q} \theta_{i,q}^2}{2} + \rho_3$  and the parameters  $k_{i,m}, \rho_2, R_{i,q}$  are all need to be positive.

By solving the differential function  $\dot{V} \leq -e_1 V + e_2$  we get:

$$V(t) \leq [V(0) - \frac{e_2}{e_1}] e^{-e_1 t} + \frac{e_2}{e_1} \leq V(0) + \frac{e_2}{e_1} \quad (26)$$

then from Eq. (20) we could have

$$\begin{aligned} \frac{1}{2} \log \frac{k_{bi,m}^T I_q k_{bi,m}}{k_{bi,m}^T I_q k_{bi,m} - v_{i,m}^T I_q v_{i,m}} &\leq V(0) + \frac{e_2}{e_1} \\ \frac{1}{2} \xi_{i,m}^T \xi_{i,m} &\leq V(0) + \frac{e_2}{e_1} \end{aligned} \quad (27)$$

so get

$$\begin{aligned} \|v_{i,m}^T I_q\| &\leq k_{bi,m} \sqrt{1 - e^{-2[V(0) + \frac{e_2}{e_1}]}} < k_{bi,m} \\ \|\xi_{i,m}\| &\leq \sqrt{2[V(0) + \frac{e_2}{e_1}]} \end{aligned} \quad (28)$$

Because the compensating signal  $\xi_{i,m}$  can be guaranteed to be bounded, then there exists a positive constant  $C$  such that  $|\xi_{i,m,q}| \leq C$ . Then we will get  $|z_{i,m,q}| \leq k_{bi,m q} + C, q = 1, 2, 3$ .

Based on the above formulas, we can know that  $Z_1 = [z_{1,1}^T, z_{2,1}^T, \dots, z_{N,1}^T]^T$ ,  $Y = [y_1^T - r^T, y_2^T - r^T, \dots, y_N^T - r^T]^T$  considering Eq. (3) and  $H = D - A + B$  [10], we could know that  $Z_1 = HY$ , so we could get  $Z_1^T Z_1 = Y^T (H^T H) Y$  then:

$$\beta^2 \sum_{i=1}^N \sum_{q=1}^3 (y_{i,q} - r_q)^2 \leq \sum_{i=1}^N \sum_{q=1}^3 z_{i,1q}^2 \quad (29)$$

where  $\beta$  is the minimum singular value of  $H$ . Thus

$$|y_{i,q} - r_q| \leq \sqrt{\sum_{i=1}^N \sum_{q=1}^3 (y_{i,q} - r_q)^2} \leq \frac{1}{\beta} \sqrt{\sum_{i=1}^N \sum_{q=1}^3 z_{i,1q}^2} \leq \frac{\sqrt{3N}}{\beta} |z_{i,1q}|_{max} \quad (30)$$

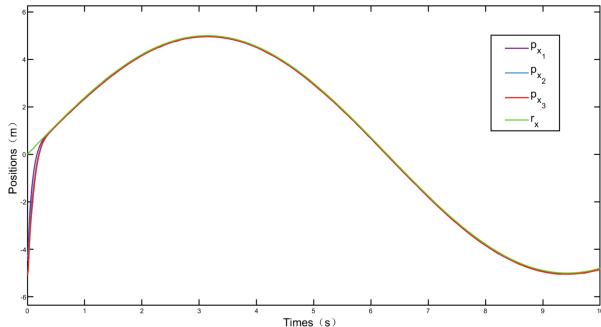
in order to guarantee  $|y_{i,q}| \leq c_{i,1q}$ , we only need  $|y_{i,q}| \leq |r_q| + \frac{\sqrt{3N}}{\beta} (k_{bi,1q} + C)_{max} \leq (c_{i,1q})_{min}$ , according assumption 4, we could have  $(k_{bi,1q})_{max} \leq -C + \frac{\beta}{\sqrt{3N}} [(c_{i,1q})_{min} - \eta_q]$ .

For  $z_{i,2q} (i \in \mathcal{V}, q = 1, 2, 3)$ , denote  $\|\pi_i\| \leq \phi_i$ , we can get  $|x_{i,2q}| \leq k_{bi,2q} + C + \phi_{i,q}$ , in order to ensure  $|x_{i,2q}| \leq c_{i,2q}$ , it only needs  $|x_{i,2q}| \leq k_{bi,2q} + C + \phi_i \leq c_{i,2q}$ , therefore  $k_{bi,2q} \leq c_{i,2q} - C - \phi_i$ .

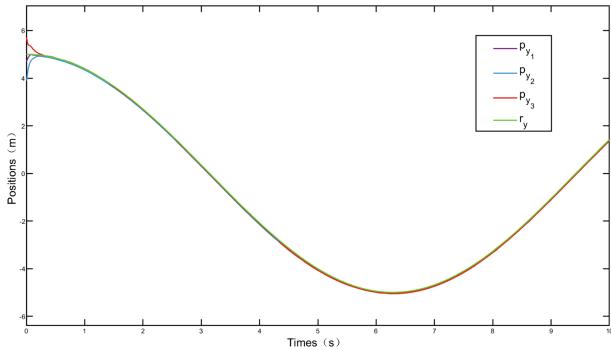
## 4 Simulations

Consider the Laplacian matrix of  $\bar{\mathcal{G}}$  is  $\begin{bmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$ , the leader adjacency matrix

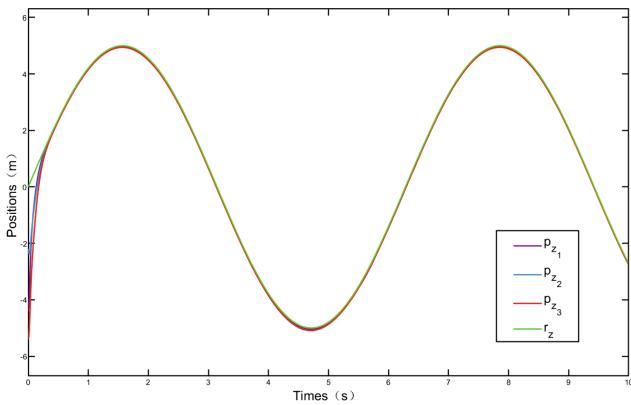
is  $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ , and their parameters is  $M_i = diag\{184.5, 135.5, 135.5\}$ ,  $D_i = diag\{140 + 80|u_i|, 100 + 80|\nu_i|, 160 + 80|\omega_i|\}$ ,  $W_i = 1259N$ ,  $B_i = 1219N$ ,  $\phi_i = \frac{\pi}{5}$ ,  $\theta_i = -\frac{\pi}{10}$ ,  $\psi_i = \frac{\pi}{15}$ ,  $i \in \mathcal{V}$ , and their positions at  $t = 0$  are  $p_1(0) = [-4.5, 4.7, -4.4]^T$ ,  $p_2(0) = [-4.3, 3.9, -2.2]^T$ ,  $p_3(0) = [-5.1, 5.7, -5.4]^T$  and all are static. The leader position is  $r(t) = [5 \sin(0.5t), 5 \cos(0.5t), 5 \sin(t)]$ . The state constraint parameter is  $c_{i,mq} = 6$ . The control parameters are  $k_{i,m} = 30$ ,  $r_{i,m,1} = 450$ ,  $r_{i,m,2} = 2 \times 10^3$ ,  $k_{bi,m} = 0.5$ ,  $h_{i,q} = 1$ ,  $\lambda_{i,q} = 1$ ,  $\mu_{i,q} = 1$ . The RBF NN has 10 neurons.



**Fig. 1.** Response curves of  $p_{x_i}, i = 1, 2, 3$  and  $r_x$ .



**Fig. 2.** Response curves of  $p_{y_i}, i = 1, 2, 3$  and  $r_y$ .



**Fig. 3.** Response curves of  $p_{z_i}, i = 1, 2, 3$  and  $r_z$ .

Figures 1-3 shows the response curves of  $p_i (i = 1, 2, 3)$  and  $r$  under the backstepping control scheme we designed above. Obvious, the anticipant control performance is achieved.

## 5 Conclusion

In this paper, we studied the problem of adaptive consensus tracking for multiple AUV systems with the positions state constraints. A command filtering backstepping control scheme with the positions state constraints is proposed, and use the NN approximation technology to approximate uncertain nonlinear dynamics. The above simulation examples prove that the designed control system is feasible.

**Acknowledgment.** This work was supported by the Shandong Province Outstanding Youth Fund (ZR2018JL020), the Science and Technology Support Plan for Youth Innovation of Universities in Shandong Province (2019KJN033) and the Project funded by Qingdao Postdoctoral Science Foundation.

## References

1. Fossen, T.I.: Marine Control Systems: Guidance, Navigation and Control of Ships, Rigs and Underwater Vehicles. Marine Cybernetics, Trondheim (2002)
2. Cui, R., Ge, S.S., How, B.V.E., Choo, Y.S.: Leader-follower formation control of underactuated autonomous underwater vehicles. *Ocean Eng.* **37**(17–18), 1491–1502 (2010)
3. Yang, E., Gu, D.: Nonlinear formation-keeping and mooring control of multiple autonomous underwater vehicles. *IEEE/ASME Trans. Mechatron.* **12**(2), 164–178 (2007)
4. Zhao, L., Yu, J., Yu, H.: Distributed adaptive consensus tracking control for multiple AUVs. In: Seventh International Conference on Information Science and Technology (ICIST), Da Nang, pp. 480–484 (2017)
5. Jia, Y.: Robust control with decoupling performance for steering and traction of 4WS vehicles under velocity-varying motion. *IEEE Trans. Control Syst. Technol.* **8**(3), 554–569 (2009)
6. Jia, Y.: Alternative proofs for improved LMI representations for the analysis and the design of continuous-time systems with polytopic type uncertainty: a predictive approach. *IEEE Trans. Autom. Control* **48**(8), 1413–1416 (2003)
7. Tong, S., Li, Y., Feng, G., Li, T.: Observer-based adaptive fuzzy backstepping dynamic surface control for a class of MIMO nonlinear systems. *IEEE Trans. Syst. Man Cybern. B Cybern.* **41**(4), 1124–1135 (2011)
8. Levant, A.: Higher-order sliding modes, differentiation and output feedback control. *Int. J. Control.* **76**(9–10), 924–941 (2003)
9. Yu, J., Shi, P., Dong, W., Chen, B., Lin, C.: Neural network-based adaptive dynamic surface control for permanent magnet synchronous motors. *IEEE Trans. Neural Netw. Learn. Syst.* **26**(3), 640–645 (2015)
10. Hu, J., Feng, G.: Distributed tracking control of leader-follower multi-agent systems under noisy measurement. *Automatica* **46**(8), 1382–1387 (2010)



# Intelligent Wireless Propagation Model with Environmental Adaptability

Xiaoyu Qu<sup>(✉)</sup> and Jiangyun Wang

School of Automation Science and Electrical Engineering,  
BeiHang University, Beijing, China  
835312287@qq.com

**Abstract.** Reasonable deployment of base stations can provide users with better services due to the development of 5G. During the deployment, wireless propagation models are used to predict signal coverage. Based on Deep Learning, this paper proposes an intelligent model, which can be adapted to a variety of environments and predict the user Reference Signal Receiving Power (RSRP). After introducing the data preparation, this paper presents the construction and training of the neural network, and then compares the intelligent model with the existing model. The results suggest that the intelligent model proposed can predict RSRP more accurately with stronger environmental adaptability. Therefore, in some cases, the model proposed can replace the existing model for deployment of base stations.

**Keywords:** Propagation model · Deep learning · Neural network · RSRP

## 1 Introduction

Wireless propagation models can contribute significantly to the deployment of 5G networks. In order to meet user requirements, operators need wireless propagation models to forecast the radio coverage and to thoughtfully choose the site of the 5G base station. However, classic wireless propagation models, such as Cost 231-Hata and Okumura-Hata model, are always limited by environment, which will cause huge errors of prediction. For example, Singh has proved that path loss predicted by Okumura-Hata model is much lower than the actual measured value in the street environment [1].

Currently, these models should be corrected in their application according to the environment on the propagation path [2]. Thus, these models are complex and changeable, which brings great difficulties for the use of the models [5]. To obtain a model that matches the environment, a large amount of data needs to be collected to correct the model. The establishment of the propagation model is thus essentially a function fitting process. We propose an intelligent wireless propagation model, which uses Deep Learning to excavate the connections between various data. By adding environment data to the training dataset, we can adapt the model to a variety of environments for more extensive usage.

© The Editor(s) (if applicable) and The Author(s), under exclusive license

to Springer Nature Singapore Pte Ltd. 2021

Y. Jia et al. (Eds.): CISC 2020, LNEE 705, pp. 326–332, 2021.

[https://doi.org/10.1007/978-981-15-8450-3\\_35](https://doi.org/10.1007/978-981-15-8450-3_35)

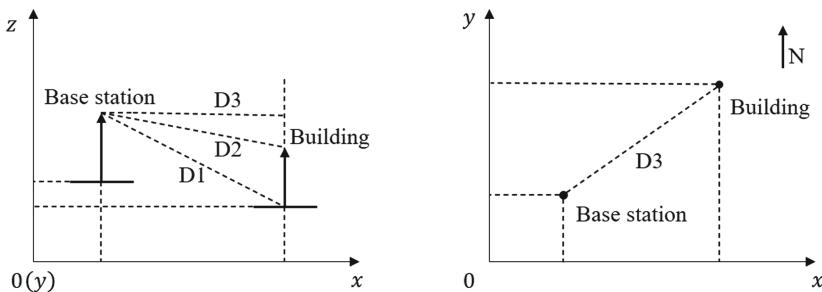
## 2 Data Preparation

The original data come from the 16th China post-graduate mathematical contest in modeling. The data have 18 columns, with the first 9 columns being engineering parameter data, the middle 8 columns map data, and the last column the RSRP label data used for learning. We clean the original data to form a dataset for neural network training. During data cleaning, we removed the error data, and finally formed a data set containing 10 million pieces of data.

Based on the correction factors in the Okumura-Hata model, we reanalyze the data and extract some new data features for neural network training, including distance parameters, angle parameters and area parameters.

### 2.1 Distance Parameters

Distance parameter mainly contains various distances between the measurement point and the base station (Fig. 1).



**Fig. 1.** Distance parameters

To calculate the distance from the base station to the bottom and top of the building, we obtain the distance parameters  $D_1$  and  $D_2$ . The projection of  $D_1$  and  $D_2$  on the horizontal plane is  $D_3$ , which measures the horizontal distance between the base station and the measurement point.

$$D_1 = \sqrt{(\Delta X)^2 + (\Delta Y)^2 + (\Delta Z - \text{BuildingHeight})^2} \quad (1)$$

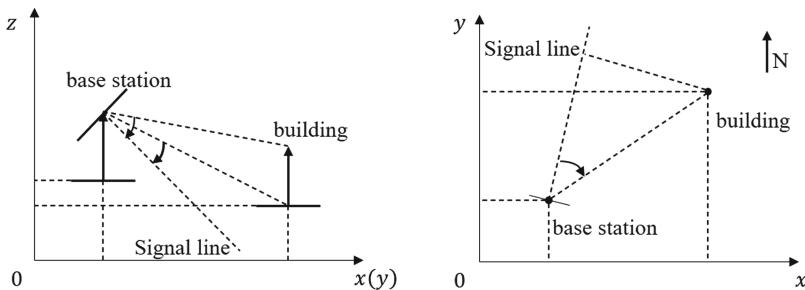
$$D_2 = \sqrt{(\Delta X)^2 + (\Delta Y)^2 + (\Delta Z)^2} \quad (2)$$

$$D_3 = \sqrt{(\Delta X)^2 + (\Delta Y)^2} \quad (3)$$

### 2.2 Angle Parameters

The angle parameter mainly contains two angles between the measurement point and signal line, as shown in Fig. 2.

Angle parameters are similar with the distance parameter, with two angle parameters in the vertical direction and one angle in the horizontal direction.



**Fig. 2.** Angle parameters

The vertical angle is:

$$\varphi_1 = \text{ElectricalDowntilt} + \text{MechanicalDowntilt} \pm \arccos \frac{D_3}{D_2} \times \frac{180^\circ}{\pi} \quad (4)$$

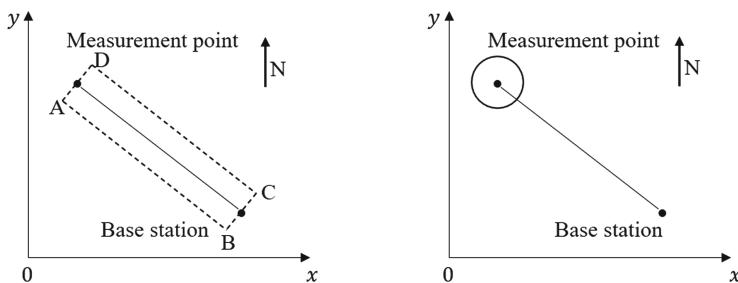
$$\varphi_2 = \text{ElectricalDowntilt} + \text{MechanicalDowntilt} \pm \arccos \frac{D_3}{D_1} \times \frac{180^\circ}{\pi} \quad (5)$$

The horizontal angles are:

$$\theta = \pm(90^\circ + \arctan \frac{\Delta Y}{\Delta X} \times \frac{180^\circ}{\pi}) - \text{Azimuth} \quad (6)$$

### 2.3 Area Parameters

Area parameters are different from the previous two types of parameters, involving the influence of buildings, forests, and waters in the area around the propagation path and the measurement points (Fig. 3).



**Fig. 3.** Area parameters

To calculate regional parameters, it is necessary to count the cell index in the target area; thus, we calculate the building density, floor area ratio, water proportion and forest proportion in the area.

Building density:

$$\rho = \frac{\sum_{i=1} \rho(i) \times \Delta s_i}{\sum_{i=1} \Delta s_i} \quad (7)$$

Floor area ratio:

$$P = \frac{\sum_{i=1} P(i) \times \Delta s_i}{\sum_{i=1} \Delta s_i} \quad (8)$$

Water and Forest proportion:

$$K_{ocean} = \frac{\sum i_{ocean}}{\sum i} \quad (9)$$

$$K_{forest} = \frac{\sum i_{forest}}{\sum i} \quad (10)$$

### 3 Intelligent Wireless Propagation Model

#### 3.1 Building a Neural Network

Based on characteristics of data and TensorFlows framework, we build a deep neural network (DNN). DNN is a neural network with a multi-layer hidden layer structure, and uses back propagation to update network parameters, the DNN we used is consist of 13 layers of neurons, including 1 input layer, 11 hidden layers, and 1 output layer. The length of the input data is 33, so the input layer contains 33 neuron nodes. RSRP is the label data, so there is only one neuron in the output layer. The number of neurons included in hidden layers are 64, 128, 256, 512, 1024, 2048, 1024, 512, 256, 128 and 32. This neural network is spindle-shaped, the upper part mainly mines the data connections, and the lower part can extract the most valuable data features. Based on experience and experiments, we constantly adjust the number of neural network layers and the number of neurons, finally we determine the structure of the above neural network, which can get best results.

The analysis shows that the model is a regression instead of a classification model; thus, we use mean-square error (MSE) as the loss function. After considering the size of the data set, we set the batch size to 2000 and learning rate to 0.0001.

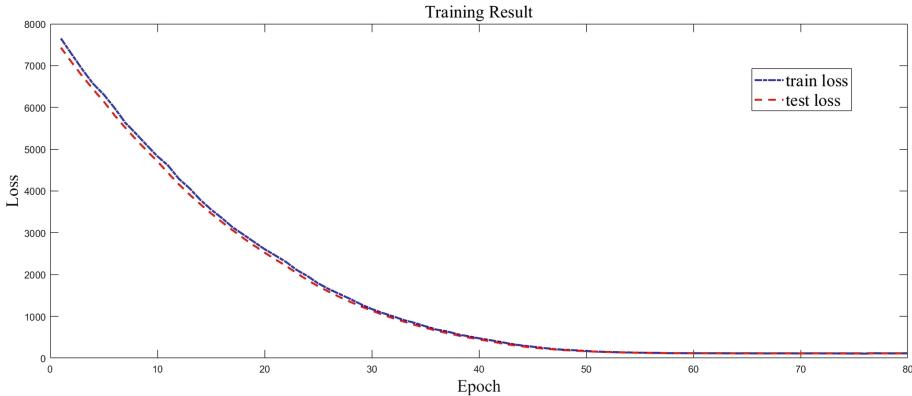
#### 3.2 Training the Neural Network

Using TensorFlow, we read a certain amount of data in each epoch to train and update the neural network parameters with Mini-Batch Gradient Descent (MBGD) algorithm, and then repeat this process until the loss does not decline. After the loss converges, we can save the neural network model which is the intelligent wireless propagation model.

## 4 Results

### 4.1 Neural Network Training Results

During the training, we recorded the training error and test error of each epoch (Fig. 4). The blue line records the change in the training error and the red line the test error.

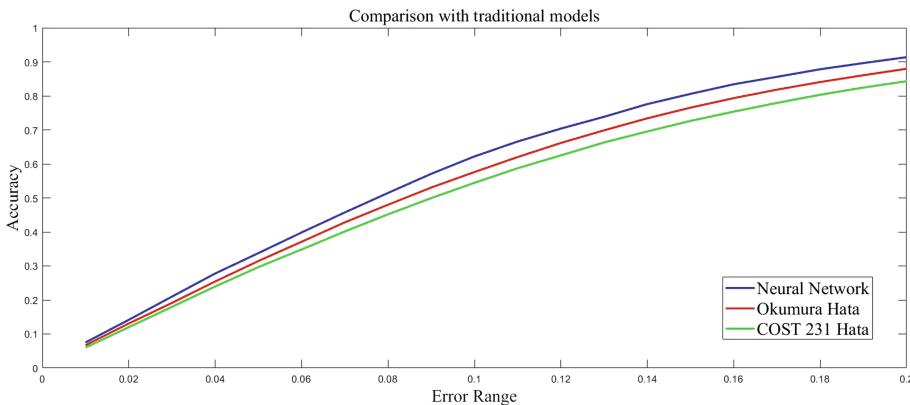


**Fig. 4.** Training result

It can be seen from the Fig. 4 that both lines are gradually decreasing and converging. This shows that the neural network gradually learns the data features during the training, which gradually reduces the learning error. Eventually, the models error (MSE) is stable at 108.5264.

### 4.2 Comparison with Traditional Models

We used another dataset containing various environment data to test Cost 231-Hata, Okumura-Hata and neural network model. Because the output of various models is the predicted value of RSRP, which is not convenient for comparison, we have adopted a special approach to compare several models. We set a parameter called the error range. When the predicted RSRP is within the error range of the true value, the prediction is accurate. We count the number of accurate predictions, divide it by the total amount of data, and then obtain the prediction accuracy. The comparison of the three models accuracy is shown in the Fig. 5. Obviously, the prediction accuracy of the neural network is higher than the Okumura-Hata and Cost 231-Hata model, which suggests that the neural network can achieve better prediction results and have stronger environmental adaptability.



**Fig. 5.** Comparison of model accuracy

## 5 Discussion and Conclusion

The existing models such as Okumura-Hata and Cost 231-Hata are limited by the environment, so we have developed and trained a neural network for an intelligent model which can adapt to different environments. By adding distance, angle and area parameters, environmental data is added to the model, making the model adaptable to the environment. Thus, our RSRP model can obtain more accurate prediction results. The comparison between existing models and neural network model shows higher accuracy of the neural network. Therefore, in some cases the neural network could replace the existing model to predict RSRP.

Many researchers have corrected and improved the propagation model on the basis of the existing model, such as the small base station [3] and street model [4]. In spite of the Improvement from these studies, limitations still exist in environment adaptability. Our model, however, has shown better environmental adaptability.

For the data preparation of this study, the data we use has a small spatial coverage, and in a certain space, there is only one base station. Thus, our research has limitations on signal superposition and interference between multiple base stations, which will lead to more research in the future.

## References

1. Singh, Y.: Comparison of okumura, hata and COST-231 models on the basis of path loss and signal strength. *Int. J. Comput. Appl.* **59**(11) (2012)
2. Jain, R., Shrivastava, L.: Performance analysis of path loss propagation models in wireless ad-hoc network. *Int. J. Res. Rev. Comput. Sci.* **2**(3), 822–826 (2011)
3. Haroon, M.S., Abbas, Z.H., Muhammad, F., Abbas, G.: Analysis of coverage-oriented small base station deployment in heterogeneous cellular networks. *Phys. Commun.* **38**, 100908 (2020)

4. Medeisis, A., Kajackas, A.: On the use of the universal Okumura-Hata propagation prediction model in rural areas (2000)
5. Xia, J., Li, C., Lai, X., Lai, S., Zhu, F., Deng, D., Fan, L.: Cache-aided mobile edge computing for B5G wireless communication networks. *EURASIP J. Wirel. Commun. Netw.* **2020**(99), 15 (2020)
6. Sinanović, D., Šišul, G., Kurdija, A.S., Ilić, Ž.: Multiple transmit antennas for low PAPR spatial modulation in SC-FDMA: single vs. multiple streams. *EURASIP J. Wirel. Commun. Netw.* **2020**(2016), 1–15 (2020)
7. Kapoor, R., Gupta, R., Kumar, R., Jha, S.: New scheme for underwater acoustically wireless transmission using direct sequence code division multiple access in MIMO systems. *Wirel. Netw.* **25**(8), 4541–4553 (2019)
8. Naveed, M., Qazi, S., Atif, S.M., Khawaja, B.A., Mustaqim, M.: SCRAS server-based crosslayer rate-adaptive video streaming over 4G-LTE for UAV-based surveillance applications. *Electronics* **8**(8), 910 (2019)
9. Ma, L., Jin, N., Zhang, Y., Xu, Y.: RSRP difference elimination and motion state classification for fingerprint-based cellular network positioning system. *Telecommun. Syst.* **71**(2), 191–203 (2019)
10. Simpson, O., Sun, Y.: LTE RSRP, RSRQ, RSSNR and local topography profile data for RF propagation planning and network optimization in an urban propagation environment. *Data Brief* **21**, 1724–1737 (2018)



# The Local Navigation and Positioning System of Unmanned Ground Vehicles

Shaowei Li<sup>1(✉)</sup>, Qingquan Feng<sup>2</sup>, Jiangang Wang<sup>1</sup>, Zhiyong Li<sup>1</sup>,  
Dongxiao Wang<sup>1</sup>, and Shizhao Liu<sup>1</sup>

<sup>1</sup> Air Force Early Warning Academy, Wuhan 430014, Hubei, China  
[shaowei\\_nudt@outlook.com](mailto:shaowei_nudt@outlook.com)

<sup>2</sup> College of Information and Communication,  
National University of Defense Technology, Wuhan 430014, Hubei, China  
[fqq0053@163.com](mailto:fqq0053@163.com)

**Abstract.** Navigation and positioning system plays an important role in autonomous driving vehicles. In view of the existing problems, this paper mainly introduces the design and the implementation of local navigation and positioning system of UGVs. At first, this paper introduces all kinds of dead-reckoning algorithms for local navigation and positioning system of UGVs, including DR algorithm based on odometer, IMU, and the combination of IMU and odometer. Then, it analyzes the advantages, disadvantages and existing problems of these algorithms, and demonstrates the rationality of DR algorithm based on IMU and odometer. Finally, the navigation and positioning results of different algorithms are compared by experiments. Experiments show that the algorithm based on IMU and odometer is better than other algorithms. It is feasible to use to complete local navigation and positioning system of UGVs.

**Keywords:** Unmanned ground vehicles · Local navigation and positioning system · Dead-reckoning

## 1 Introduction

In recent years, unmanned ground vehicles (UGVs) has made great progress. Driverless system is a complex system composed of environment perception module, navigation and positioning module, path-planning module, motion-control module and so on. Navigation and positioning module plays a very important role for UGVs. As one of the basic modules, navigation and positioning module serves for other modules at the same time.

According to different coordinate systems and positioning principles, methods can be generally divided into local navigation and positioning, global navigation and positioning. This paper mainly studies the technical problems existing in local navigation and positioning of UGVs, and puts forward some corresponding solutions.

## 1.1 Demand Analysis

This section will introduce the demand analysis of “localpose” module. The module has three features: the high output frequency, the smooth positioning track and small accumulative error in a short time. These features of “localpose” module determine its important role in the UGVs, which is mainly reflected in the following aspects.

First, it can complete sensor data alignment.

Second, it can provide motion prior information for environment perception module. “Localpose” module can be used as the prior information of perception module, such as lidar SLAM algorithm, to make the result of positioning and mapping more accurate.

Third, it can provide information for path-planning module. In path-planning module, the path sent by the planning layer is always based on vehicle’s coordinate system, which represents the movement trend of UGVs. Then the relationship between vehicle’s body and surface road is constantly updated by using the information of “localpose” module. In addition, the output of control module is also constantly adjusted.

## 1.2 Research Status of Dead-Reckoning

Local navigation and positioning aims to determine the position of robots in the local coordinate system. The main principle is dead-reckoning (DR) algorithms [1]. These algorithms integrate step lengths and orientation estimations at each detected second, so as to compute the absolute position and orientation of a robot. The accuracy of positioning results is affected by many factors, such as different sensors and the accuracy of sensors. In addition, Inertial Measurement Units (IMU) and odometer are commonly used sensors.

Dead-reckoning algorithms mainly use observation information of sensors for integral operation. These algorithms are based on vehicle’s dynamic model, odometer, IMU or lidar. In terms of theory, using the vehicle’s dynamic model and the input in a past period of time, the motion state of the system can be calculated. The dynamic model is usually based on the locations of instantaneous centers of rotation (ICRs) of treads [2]. However, in practice, due to the influence of the inertial force, the ICRs of treads will change. In addition, the vehicle’s dynamic model involves many parameters. It is not easy to estimate these parameters accurately [3,4].

Odometer is a kind of sensor to measure the forward distance of vehicle. According to the geometric calculation, the displacement of vehicle in a past period of time can be estimated. Reference [5] proposed a scheme of dead-reckoning algorithm based on vehicle’s odometer when the satellite signal is lost.

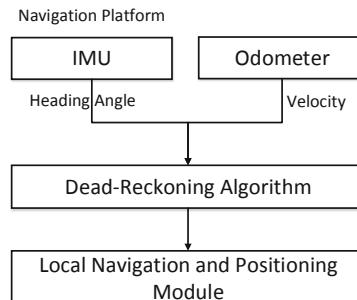
Inertial Measurement Units (IMU) is composed of accelerometer and gyroscope, which can measure the three-axis acceleration and the three-axis angular velocity. The dead-reckoning algorithm based on IMU is a typical independent navigational method. The advantages are obvious. Its work will not be affected

by external environment. And IMU is small and easy to use. The disadvantages are also obvious. Due to the existence of noise, the error of dead-reckoning algorithm based on IMU is very large. Reference [6] studied the influence of IMU's noise on navigation and positioning results.

In recent years, the emergence of lidar provides a new choice for the navigation and positioning module of UGVs [7–10]. This paper focuses on a low-cost navigation and positioning method. As the price of lidar is too expensive at present, there is no in-depth research on lidar related algorithms in this paper.

### 1.3 Research Status of Local Navigation and Positioning Module

After years of engineering practice, it can be found that the navigation and positioning module has two main functions for UGVs. First, it can output motion state information of vehicles to serve for path-planning module and environment perception module. Second, it can locate the vehicle's position and attitude in the global map or global coordinate systems, and understand the task file of UGVs. In order to realize the first function, this paper designs a local navigation and positioning module of UGVs in Fig. 1.



**Fig. 1.** Local navigation and positioning module

The main function of local navigation and positioning module is to calculate vehicle's position and attitude in the local coordinate system, and output vehicle's motion state information, mainly including the vehicle's position in the local coordinate system  $x, y$ , attitude angle  $roll, pitch, yaw$ , angular velocity  $\omega_x, \omega_y, \omega_z$ , acceleration  $acc_x, acc_y, acc_z$ , forward-speed of car body  $v$ , throttle  $oil$ , brake  $brake$ , gear  $gear$ , etc. The input of local navigation and positioning module is the data of IMU, odometer, chassis servomechanism. With different dead-reckoning algorithms, the input of module is also different.

In addition, local navigation and positioning module contains four data-computing nodes, three of which are used to receive, analyze and transmit data of sensors. The “localpose-computing” node is used to fuse sensor's data and compute the position and attitude in the local coordinate system of UGVs.

The output of local navigation and positioning module is a structure data, called “localpose-data”. This structure data is broadcast to other systems of vehicles, such as environment perception module, path-planning module and motion-control module.

This paper analyzes and compares the advantages and disadvantages of typical dead-reckoning algorithms, and then designs and implements a local dead-reckoning algorithm based on IMU and odometer. For convenience, local navigation and positioning module is written as “localpose” module in this paper.

## 2 Problem Formulation

For structuring “localpose” module, the following assumptions shall be made.

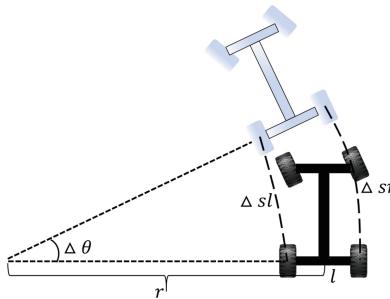
First, it is assumed that the movement of vehicle’s wheels on the ground is pure rolling, that is, the slipping displacement of wheels is ignored.

Second, in the process of each iteration, it is assumed that vehicle makes a uniform circular motion or a uniform straight motion in a very short period of time, such as 10 ms.

### 2.1 DR Algorithm Based on Odometer

Dead-reckoning algorithm based on odometer is shown in Fig. 2, Odometer is installed on each rear wheel of UGVs.

Suppose that the number of pulses of left and right odometers in unit time is  $n_l, n_r$  and each pulse represents a distance of  $k$ . Thus the distance of left and right wheels moving in unit time is  $\Delta s_l = k * n_l, \Delta s_r = k * n_r$ . Suppose that the position of the center of both rear wheels at the last moment is  $x_{k-1}, y_{k-1}, \theta_{k-1}$ , and at the current moment is  $x_k, y_k, \theta_k$ .  $l$  is the track width of both rear wheels,  $r$  is the radius of uniform circular motion, and  $\Delta s$  is the distance moved in the past period of time.



**Fig. 2.** Dead-reckoning algorithm based on Odometer

Therefore, the calculation of vehicle's position at the current time is shown in Eq. (1).

$$\begin{aligned}\Delta\theta &= \frac{\Delta sr - \Delta sl}{l}, \\ \theta_k &= \theta_{k-1} + \Delta\theta, \\ \Delta s &= \frac{\Delta sr + \Delta sl}{2}, \\ x_k &= x_{k-1} + \Delta s \cos \theta_k, \\ y_k &= y_{k-1} + \Delta s \sin \theta_k,\end{aligned}\tag{1}$$

The solution of Eq. (1) is vehicle's position  $x_k, y_k, \theta_k$ . In addition, the calculation of  $\Delta\theta$  is very important. In practice, due to the limitation of the sensor error, the error of  $\Delta\theta$  is very large.

## 2.2 DR Algorithm Based on IMU

An IMU is installed in the center of rear wheels. Assume that the acceleration measured by IMU at the current time is  $a_x, a_y, a_z$ , the angular velocity is  $\omega_x, \omega_y, \omega_z$ . The position of the center of both rear wheels at the last moment is  $x_{k-1}, y_{k-1}, \theta_{k-1}$ , and at the current moment is  $x_k, y_k, \theta_k$ .

According to Newton's law of motion, the calculation of vehicle's position at the current time is shown in Eq. (2).

$$\begin{aligned}v_k &= v_{k-1} + a_y \Delta t, \\ \Delta s &= v_k \Delta t, \\ \theta_k &= \theta_{k-1} + \omega_z \Delta t, \\ x_k &= x_{k-1} + \Delta s \cos \theta_k, \\ y_k &= y_{k-1} + \Delta s \sin \theta_k,\end{aligned}\tag{2}$$

Due to the error of IMU and the value of  $a_y$  is always not 0, the cumulative error of  $v_k$  is increasing with time, and the cumulative error of  $\Delta s$  is also increasing.

## 2.3 DR Algorithm Based on Odometer and IMU

One of the characteristics of dead-reckoning algorithm is that the position of current time depends on the position of the previous time. If there is an error in the previous position, the error will also accumulate to the position in the future. Because of twice integration of acceleration, the accumulated error of DR algorithm based on IMU will be multiplied. In order to reduce the accumulative error as much as possible, it is a feasible method to use the data of odometer and IMU at the same time. We can use odometer to measure the distance moved in the past period of time, and use IMU to measure the angular velocity of UGVs.

Odometers are installed on each rear wheel of UGVs and an IMU is installed in the center of rear wheels. Suppose that the number of pulses of left and right odometers in unit time is  $n_l, n_r$  and each pulse represents a distance of  $k$ . Thus the distance of left and right wheels moving in unit time is  $\Delta sl = k * n_l, \Delta sr = k * n_r$ .

The distance  $\Delta s$  moved in the past period of time is the average of the distance measured by both odometers. That is

$$\Delta s = \frac{\Delta sl + \Delta sr}{2}, \quad (3)$$

Assume that the angular velocity measured by IMU at the current time is  $\omega_x, \omega_y, \omega_z$ . The calculation of vehicle's position at the current time is shown in Eq. (4).

$$\begin{aligned} \theta_k &= \theta_{k-1} + \omega_z \Delta t, \\ x_k &= x_{k-1} + \Delta s \cos \theta_k, \\ y_k &= y_{k-1} + \Delta s \sin \theta_k, \end{aligned} \quad (4)$$

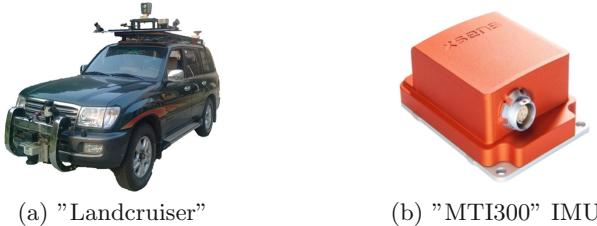
The Eq. (4) and the Eq. (2) are the same in form. The difference is that  $\Delta s$  is calculated in different ways.

### 3 Experiment Results

In view of the algorithm proposed above, the corresponding experimental verification is carried out in this section.

#### 3.1 Experiment Platform

The experiment platform is shown in Fig. 3. The UGV used in experiment was adapted from “landcruiser” produced by Toyota. The IMU used in experiment is the mti300 series produced by Xsens. The MTI 300-series features vibration-rejecting gyroscopes and accelerometers, and offers high-quality position, velocity, acceleration, and orientation, even in challenging environments.



**Fig. 3.** Experiment platform

At the same time, the accurate position of UGVs is output by SPAN-CPT, a high-precision integrated inertial navigation system produced by NOVATEL.

### 3.2 Experiment Results and Analysis

All the following data are collected on a same experiment platform at the same time. The data of IMU and odometer used in different algorithms are the same.

**Experiment Result of DR Algorithm Based on Odometers.** The experiment result of DR algorithm based on odometer is shown in Fig. 4. In Fig. 4(a), the red trajectory is calculated only by odometer, and the pink track is the accurate trajectory. Comparing the two trajectories, it can be found that the deformation of red trajectory is very large. In Fig. 4(b), the red curve is the heading angle calculated by odometer, and the pink curve is the accurate heading angle of vehicle. The error of heading angle is shown in Fig. 4(c). It can be found that the error of heading angle calculated by odometer is very large.

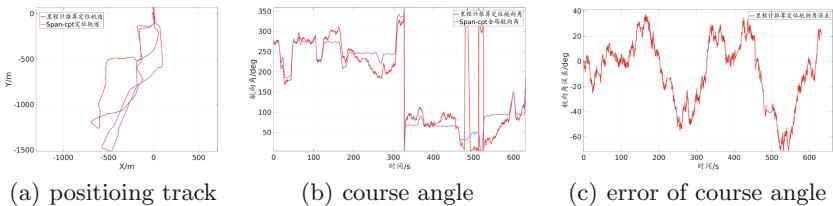
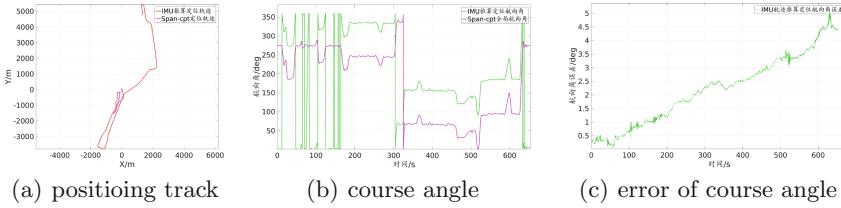


Fig. 4. Experiment result of DR algorithm based on odometer

Through the analysis above, it can be concluded that the DR algorithm based on odometers is not suitable for local navigation and positioning module due to the large error of heading angle.

**Experiment Result of DR Algorithm Based on IMU.** The experiment result of DR algorithm based on odometer is shown in Fig. 5. In Fig. 5(a), the red trajectory is calculated only by IMU, and the pink track is the accurate trajectory. Comparing the two trajectories, the red trajectory based on IMU is similar to the real trajectory in shape, but its length is several times larger than the real vehicle's trajectory. In addition, it can be found that the trajectory calculated by IMU is not closed.

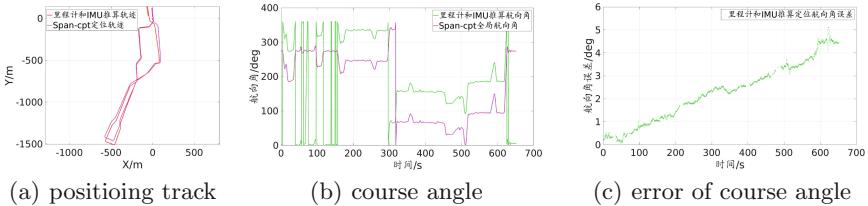
In Fig. 5(b), the green curve is the heading angle calculated by IMU, and the pink curve is the accurate heading angle of vehicle. It can be found that the difference between the heading angle based on IMU and the real heading angle of vehicle is very small. It can be seen from Fig. 5(c) that the heading angle error calculated by IMU is increasing slowly. Through the analysis above, it can be concluded that the error of acceleration measured by IMU is very large. However, the angular velocity measured by IMU is relatively accurate, and in a short period of time, the accumulated error of heading angle is relatively small. So the angular velocity measured by IMU is useful for DR algorithm.



**Fig. 5.** Experiment result of DR algorithm based on IMU

In conclusion, the distance measured by odometer is relatively accurate, and the angular velocity measured by IMU is accurate. Therefore, it is feasible to use the combination of IMU and odometer to realize the DR algorithm.

**Experiment Result of DR Algorithm Based on Odometers and IMU.** The experiment result of DR algorithm based on odometers and IMU is shown in Fig. 6. In Fig. 6(a), the red trajectory is calculated by odometers and IMU, and the pink track is the accurate trajectory. Comparing the two trajectories, the red trajectory is similar to the real trajectory. In Fig. 6(b) and Fig. 6(c), it is easy to find that the error of heading angle is the same as the last experiment.



**Fig. 6.** Experiment result of DR algorithm based on Odometers and IMU

Through the analysis above, it can be concluded that the deformation and heading angle error calculated by IMU and odometer are smaller than other DR algorithms. In conclusion, it is feasible to use the combination of IMU and odometer to realize the DR algorithm which is useful to complete the local navigation and positioning system of unmanned ground vehicles.

## 4 Conclusions

This paper mainly introduces the design and the implementation of local navigation and positioning system of UGVs, including the development of UGVs, the research status of deadreckoning, the demand analysis, the problem formulation, the experiment result and conclusions.

In this paper, a local dead-reckoning algorithm based on IMU and odometer is designed and implemented.

At first, the paper introduces all kinds of dead-reckoning algorithms for local navigation and positioning system of UGVs, including DR algorithm based on odometer, IMU, and the combination of IMU and odometer.

Then, it analyzes the advantages, disadvantages and existing problems of these algorithms, and demonstrates the rationality of DR algorithm based on IMU and odometer.

Finally, the navigation and positioning results of different algorithms are compared by experiments.

Experiments show that the algorithm based on IMU and odometer is better than other algorithms. It is feasible to use to complete local navigation and positioning system of UGVs.

## References

1. Jimenez, A.R., Seco, F., Prieto, C., Guevara, J.: A comparison of pedestrian dead-reckoning algorithms using a low-cost mems IMU. In: 2009 IEEE International Symposium on Intelligent Signal Processing, pp. 37–42. IEEE (2009)
2. Martínez, J.L., Morales, J., Mandow, A., Pedraza, S., García-Cerezo, A.: Inertia-based ICR kinematic model for tracked skid-steer robots. In: 2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR), pp. 166–171. IEEE (2017)
3. Li, H., Zhao, Y., Lin, F., Zhu, M.: Nonlinear dynamics modeling and rollover control of an off-road vehicle with mechanical elastic wheel. *J. Braz. Soc. Mech. Sci. Eng.* **40**(2), 51 (2018)
4. Zhao, Y., Yang, Z., Song, C., Xiong, D.: Vehicle dynamic model-based integrated navigation system for land vehicles. In: 2018 25th Saint Petersburg International Conference on Integrated Navigation Systems (ICINS), pp. 1–4. IEEE (2018)
5. Dicu, N., Andreeșcu, G.-D., HoratiuGurban, E.: Automotive dead-reckoning navigation system based on vehicle speed and yaw rate. In: 2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI), pp. 000 225–000 228. IEEE (2018)
6. Kepper IV, J.H., Claus, B.C., Kinsey, J.C.: A navigation solution using a MEMS IMU, model-based dead-reckoning, and one-way-travel-time acoustic range measurements for autonomous underwater vehicles. *IEEE J. Oceanic Eng.* **99**, 1–19 (2018)
7. Segal, A., Haehnel, D., Thrun, S.: Generalized-ICP. In: *Robotics: Science Systems*, vol. 2, no. 4, p. 435 (2009)
8. Wang, Y.-H., Qi, J.-Y., Xiong, L., Wei, S.: A implementation method of indoor slam using velodyne laser radar for mobile robot. *DEStech Transactions on Computer Science and Engineering*, no. cnai (2018)
9. Gomez-Ojeda, R., Moreno, F.-A., Zuñiga-Noël, D., Scaramuzza, D., Gonzalez-Jimenez, J.: PL-SLAM: a stereo SLAM system through the combination of points and line segments. *IEEE Trans. Robot.* **35**(3), 734–746 (2019)
10. Demim, F., Nemra, A., Boucheloukh, A., Kobzili, E., Hamerlain, M., Bazoula, A.: SLAM based on adaptive SVSF for cooperative unmanned vehicles in dynamic environment. *IFAC-PapersOnLine* **52**(8), 73–80 (2019)



# A Novel 3D Lidar-IMU Calibration Method Based on Hand-Eye Calibration System

Lei Ji, Long Zhao<sup>(✉)</sup>, Jingyun Duo, and Chao Wang

Digital Navigation Center, School of Automation Science and Electrical Engineering,  
Beihang University, Beijing 100191, China  
[flylong@buaa.edu.cn](mailto:flylong@buaa.edu.cn)

**Abstract.** In this paper, we proposed a method of targetless and automatic Lidar-IMU (Inertial Measurement Unit) calibration. Our approach is an extension of hand-eye calibration framework. Unlike global-shutter cameras, lidar collects a succession of 3D-points generally grouped in scans. By using the calibrated extrinsic matrix, IMU message can eliminate point clouds' distortion so as to optimize the scan match again. The calibration does not rely on any prior information. A series of real data are used to show the effectiveness of the proposed method.

**Keywords:** Hand-eye calibration · Lidar-IMU Calibration · Distortion correction

## 1 Introduction

With the continuous improvement of SLAM (Simultaneous Localization And Mapping), sensor fusion has been widely studied in the field of robotics and computer vision. In the past few years, the robotics community has proposed various multi-sensor fusion algorithms for localization and mapping. From visual-inertial navigation [1] to visual-lidar odometry and mapping [2], both techniques rely on accurate extrinsic calibration and synchronization between the sensing devices.

For the purpose of getting better result by combining the two sensors measurements, we must estimate the rotation and translation between Lidar and IMU. The hand-eye calibration (HEC) [3] is used to determine the relative pose between the gripper and the sensor mounted on the gripper. The word eye of the HEC expresses the lidar and the world hand represents the gripper. Through an external calibration and the measurement of IMU, the Lidar movement and the gripper movement can be acquired respectively. It is known that hand-eye calibration is virtually degenerated into an equation composed of the homogeneous matrices with the form of  $AX = XB$  which is proposed by Shiu and Ahnmad [4] firstly. The rotation and translation between IMU and camera is denoted by  $X$ , and the homogeneous matrices  $A, B$ , respectively describe the transformation of IMU and camera between two

different positions or postures of the IMU-Camera system. In this paper, the calculation of matrix  $A$  depends on the data collected by IMU and the algorithm of Attitude and Heading Reference System (AHRS) [5].

Over the past few years, visual-inertial extrinsic calibration has become more accurate and got less restrictive concerning the set-ups needed. Alves J. [6] and Lobo J [7] are examples of complex calibration rigs using actuators and external sensors, such as spinning tables and shaft encoders, to recover the sensor spatial transformation. Also, lidar-inertial extrinsic calibration has been developed fast, such as [8,9] and [10]. However, to our knowledge, there is no successful method which use point-to-line distance to jointly calibrate and localize the system using on-manifold optimization.

This paper proposes a calibration method that does not need any predefined targets. By using the Re-calibration method described in Sect. 3.3, it can get accurate and robust results in real data. The remainder of this paper is organized as follows: the next section describes the detailed exposition on hand-eye calibration equation and the initial solutions via linear decomposition algorithm; Sect. 3 describe the method to get transformation of two continuous scan; Sect. 4 contains several experimental results and contrast diagrams; Sect. 5 expresses our conclusions and pointers to future; Finally expresses our thanks for the support.

## 2 Problem Formulation

In Visual-Inertial System (VIO), the hand-eye calibration with camera and IMU can be translated to the solution of the equation  $AX = XB$ , where the matrix  $X$  represents the unknown rotation and translation between IMU and camera. Matrices  $A$  and  $B$ , denote the transformation of IMUs and cameras between two different positions or postures of the IMU-Camera system respectively. Besides, this can also be applied to Lidar-IMU calibration.

### 2.1 HEC Equation

HEC Equation can be written as:

$$AX = XB \quad (1)$$

The matrix  $A$  and  $B$ , respectively, denote the transformations from the current pose to the last pose of IMU and lidar, Matrix  $X$  represents the unknown rotation and translation between IMU and Lidar. In Eq. (1), Matrix  $B$  can be acquired through the AHRS, which is easy to be estimated. So we will not described it in this paper. The method to obtained matrix  $A$  is the key part in lidar-IMU calibration and it will be described in the next section.

## 2.2 HEC Linear Solution

Now we can focus on the solution of the HEC equation:  $A$ ,  $B$ ,  $X$  are all 44 matrixs. In order to analyze (1) more easily,  $A$ ,  $B$  and  $X$  can be expressed with rotation and translation:

$$A = \begin{bmatrix} R_A & t_A \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} R_B & t_B \\ 0 & 1 \end{bmatrix}, X = \begin{bmatrix} R_X & t_X \\ 0 & 1 \end{bmatrix} \quad (2)$$

where  $R_A$  represent the rotation part of  $A$ , and it is a 33 matrix,  $t_A$  represent the translation part of  $A$ , and it is a 31 vector. It is same for  $R_B$ ,  $t_B$ ,  $R_X$  and  $t_X$ . As a result,  $AX = XB$  can be transformed into:

$$R_A R_X = R_X R_B \quad (3)$$

and

$$(R_A - I_3)t_X = R_X t_B - t_A \quad (4)$$

According to (1), the cost function of hand-eye calibration is:

$$L(X) = \min \frac{1}{2} \sum_k \|A^k X - X B^k\|_F^2 \quad (5)$$

and it has another format:

$$L(R_X, t_X) = \min \frac{1}{2} \sum_k (\|R_{A^k} R_X - R_X R_{B^k}\|_F^2 + \|(R_{A^k} - I_3)t_X - R_X t_{B^k} + t_{A^k}\|_F^2) \quad (6)$$

where  $\|\cdot\|$  denotes the F-norm of matrix.

Then we treat (6) as the cost function. Its jacobian matrix is

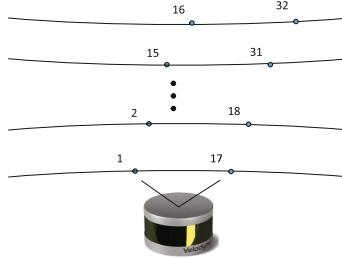
$$J = \frac{\partial L(R_X, t_X)}{\partial (R_X, t_X)} \quad (7)$$

In order to simplify (7), we use  $r_x, r_y$  and  $r_z$  to alternative  $R_X$ ,  $r_x, r_y$  and  $r_z$  respectively represent the rotation angles about the X-axis, Y-axis and Z-axis. Also  $t_X$  can be replaced by  $t_x, t_y, t_z$ . According to the Gauss-Newton algorithm, once the jacobian matrix is obtained,  $R_X$  and  $t_X$  will be optimezied easily.

## 3 Transformation Between Continuous Sweep

### 3.1 Feature Extraction

Figure 1 shows the working process of lidar. The black line represents lidars 16 scan lines from  $-15^\circ$  to  $15^\circ$  in vertical plane, its resolution is  $1^\circ$ . First, it generates a series of laser beam in the vertical plane (as show by point 1, 2,  $\dots$ , 15, 16), then the beam change its horizontal angle to begin a new scan (as show by point 17, 18,  $\dots$ , 31, 32).



**Fig. 1.** The working process of lidar (take VLP-16 as an example)

Similar to [11], we extract features with the smoothness criterion. Let  $P_k$  be the point cloud collected from sweep  $k$ , and  $P_{(k,i)}$  is a point in  $P_k$ , and let  $S$  be the set of consecutive points of  $i$  returned by the laser scanner in the same scan. Define a term to evaluate the smoothness of the local surface:

$$\eta = \frac{1}{|s| * \|P_{(k,i)}\|} \left\| \sum_{j \in S, j \neq i} (P_{(k,i)} - P_{(k,j)}) \right\| \quad (8)$$

$\eta$  is also called curvature. The term is normalized w.r.t. the distance to the lidar center. Usually  $s$  equals to 10. When calculating the curvature of point  $i$ , it is necessary to take 5 points on each side of  $i$ , and put the coordinates of these 11 points into (8). If one has a large curvature, it will be considered as a feature point.

### 3.2 Scan Matching

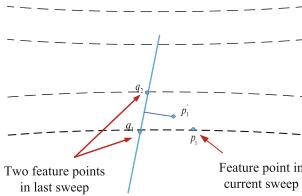
Suppose that we have gotten enough feature points from last section, the next step is to match the current scan with the last scan.

First, we need to transform all the point collected in last sweep to the end of this sweep. This set of points is called the source points. Next, when a new sweep finished, we should transform the point collected in the current sweep to the beginning of this sweep, and this points called target points.

Define  $P = \{p_1, \dots, p_k\}$  as a set of all the feature points in current sweep and  $Q = \{q_1, \dots, q_k\}$ , represent the set of all feature points, which has been transformed the feature points collected from last sweep to the beginning of the sweep. Figure 2 shows the matching operator. Point  $p_i$  represent one feature point in current sweep, and its coordinate is defined by  $P_{p_i} = \{x_{p_i}, y_{p_i}, z_{p_i}\}$ . Suppose that the transformation between last sweep and current sweep was define by  $R_A$  and  $t_A$ , point  $p'_i$  is the projection of point  $p_i$  to the start of the sweep. If point  $p_i$  is the last point in current sweep, point  $p'_i$ 's coordinates can be described as

$$P'_{p_i} = R_A P_{p_i} + t_A \quad (9)$$

Since we assume that the whole motion process is uniformed, linear interpolation can be used to estimate the pose of each point relative to the initial moment of the sweep. So for general feature points, we can also project it to the initial moment when the transformation of the whole sweep is known. We use the distance from the point to the line as the constraint of two adjacent sweep, as show in Fig. 2.



**Fig. 2.** Scan matching

Point  $q_1$  and  $q_2$  are match points to point  $p_1$ , and the method to find them is: both  $q_1$  and  $q_2$  are feature points in last sweep, point  $q_1$  belongs to the same scan line and has the smallest distance with point  $p_1$ , point  $q_2$  belong to the adjacent scan line and has the second smallest distance with point  $p_1$ . After the correlation line is determined, the constrain can be written as:

$$d_1 = \frac{\|(P_{p'_1} - P_{q_1}) \times (P_{p'_1} - P_{q_2})\|}{\|P_{q_1} - P_{q_2}\|} \quad (10)$$

$d_1$  represent the distance between point  $P_{p'_1}$  and line  $P_{q_1}P_{q_2}$ . If we can get enough feature points, we will get the transform between current sweep and last sweep by minimizing  $d_1$ .

### 3.3 Re-calibration

Up to now, we have got  $A$  in Eq. (1), and the method to obtain  $B$  is mature: we use Runge Kutta method to obtain the relative translation and rotation angles of IMU in the adjacent timestamps, and then transform the rotation angle into rotation matrix to represent the rotation. So the extrinsic matrix will be calculated by the algorithm described in Sect. 2. However, in the process of Lidar data collection, since the carrier of the Lidar is in motion, the data will contain the radar movement information of that sweep period.

As we all know, IMU's frequency can exceed 100 Hz. Though long working hours will produce a large cumulative error, the error is negligible at this small intervals. Lidars sweep frequency is 10 Hz, which is far less than IMU. After the first calibration, we have obtained the  $A$ ,  $X$  and  $B$ . In the re-calibration step, we changed our method to obtained  $A$ .

From Sect. 3.2, we suppose that the lidars motion is linear, but now we can accurately estimate the state of lidars movement during its sweep period with  $X$ . At the same time, IMUs data collection frequency is far more than Lidars frequency, it is easy to estimate the motion state of each point cloud.

At the beginning of second calibration, we use linear interpolation to align IMUs output with every point of Lidars output. And now lets define  $A_I$ ,  $B_I$  and  $X_I$  is obtained from the first calibration:

$$A_I X_I = X_I B_I \quad (11)$$

Next, we need to calculate  $B_{II}$ , which is different in definition compared with  $B_I$ .  $B_{II}$  not only contains the transformation between current IMU position, but also have the relative message between two adjacent outputs, it can be written as

$$B_{II} = B_{II}^L \oplus B_{II}^N \quad (12)$$

where  $B_{II}^L$  represents the linear part of  $B_{II}$ ,  $B_{II}^N$  represents the no-linear part of  $B_{II}$ ,  $\oplus$  represents a nonlinear combination between  $B_{II}^L$  and  $B_{II}^N$ .

Then we get

$$A_{II}^L = X_I B_{II}^L X_I^{-1} \quad (13)$$

$$A_{II}^N = X_I B_{II}^N X_I^{-1} \quad (14)$$

where  $A_{II}^L$  is the initial linear transformation obtained from IMU,  $A_{II}^N$  is the initial non-linear transformation obtained from IMU.

We use  $A_{II}^N$  to eliminate the nonlinear motion information of each point during lidars movement. Then, by removing the nonlinear motion information, all points can be regarded as the result of linear radar motion. After the scan match finished, we get a new result  $A'_{II}$ , then we combine the result with  $A_{II}^N$  as

$$A_{II} = A'_{II} \oplus A_{II}^N \quad (15)$$

where  $A_{II}$  is the transformation of lidar obtained after this step.

The rest of second calibration is the same as the first calibration, we put  $A_{II}$  and  $B_{II}$  to (1) and get a final result.

## 4 Experiment

In this section, our algorithm is tested on real data collected by VLP-16 lidar and Xsens-Mti-710 IMU.

We apply our algorithm to a series of data sets collected by different motion paths. Each dataset contains about 60s of data. During the sweep time of lidar, we let the platform experience as much movement as possible.

Since the vast scale errors exist in the external calibration, we analyze the error of rotation mainly. The error is denoted as

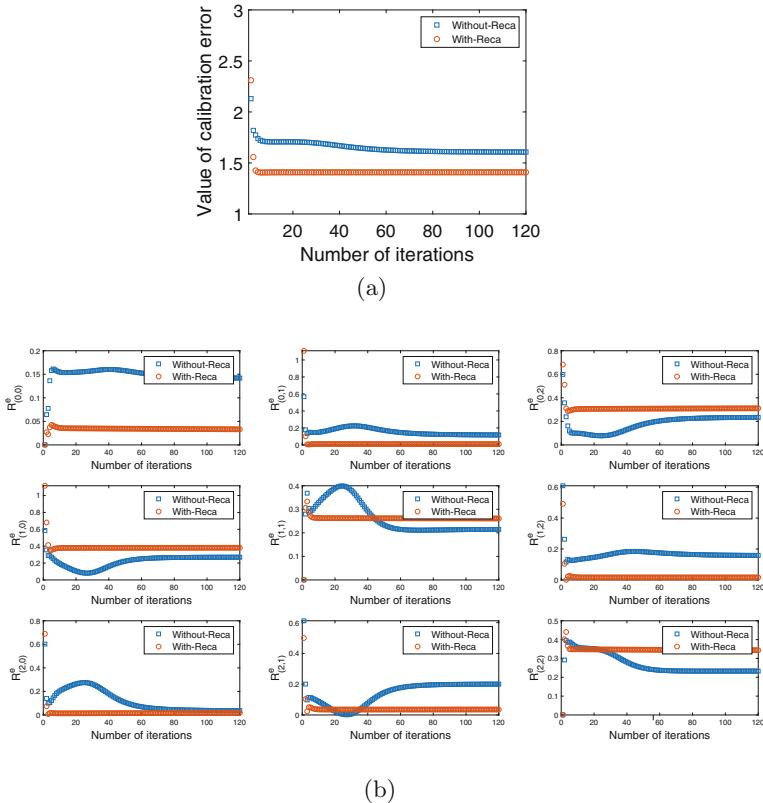
$$err = \|R_A R_X - R_X R_B\|_F^2 \quad (16)$$

In order to display the optimized effect more intuitively, we define  $e$  and  $R^e$ :

$$e = \sum_{i=1}^n \|R_{A_i}R_X - R_XR_{B_i}\|_F^2 \quad (17)$$

$$R_{(a,b)}^e = \sum_{i=1}^n (R_{A_i}R_X - R_XR_{B_i})_{(a,b)}^2 \quad (18)$$

We take all errors in each dataset as a whole, and  $n$  is set to 600. In Eq. (18), subscript  $(a, b)$  represents the index of elements in matrix. The initial rotation is set to zero. We notice the estimation of the matrix  $R_X$  causes less errors in all datasets after using the new optimal algorithm. Figure 3 shows a detail result of our algorithm.

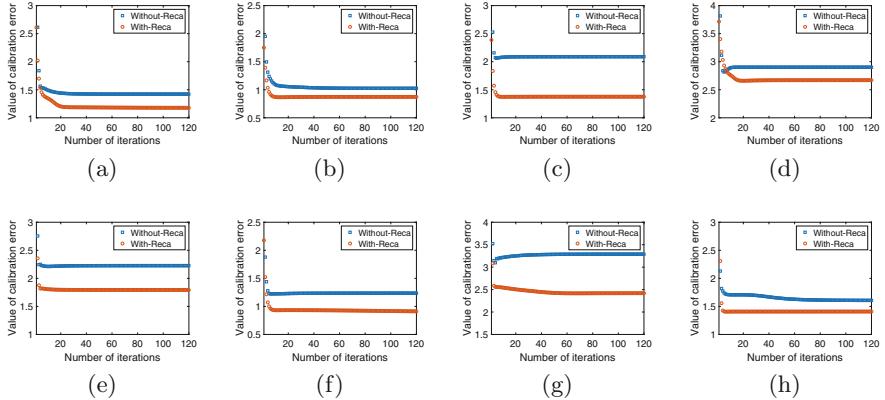


**Fig. 3.** (a) convergence trend of error  $e$ , (b) convergence trend of the square of elements in  $R^e$

The curve composed of blue box represents the error curve without recalibration and the curve composed of red circle represents the error curve after

re-calibration. We can see that not only the total error  $e$  is convergent, but also every element in  $R^e$  is convergent. Compared with the results without recalibration, the error of some elements increases, but the total error decreases obviously.

In order to quantify the improving accuracy and robustness of the algorithm, we conducted the same experiment on all data sets, and took eight groups of experimental results, as shown in Fig. 4.



**Fig. 4.** (a), (b), (c), (d), (e), (f), (g), (h) separately represent  $e$  of eight datasets.

The curve composed of blue box represents the error curve without re-calibration and the curve composed of red circle represents the error curve after re-calibration. In order to obtain a quantitative analysis of Fig. 4, The error before and after calibration is calculated in Table 1.

**Table 1.** Quantitative analysis of calibration

Dataset	a	b	c	d	e	f	g	h
Without-Reca	1.4219	1.0252	2.0907	2.9012	2.2271	1.2359	3.2889	1.606
With-Reca	1.1783	0.8732	1.3759	2.6706	1.7935	0.899	2.4195	1.4092
Dec	17.132%	14.826%	34.189%	7.948%	19.469%	27.2591%	26.434%	12.254%

In this table, the first row represents the eight datasets in Fig. 4, the second row represents the error of calibration without re-calibration, the third row represents the error of calibration with re-calibration and the last row represents the reduction rate of calibration between third row and second row. The table shows that the re-calibration improves the accuracy of Lidar-IMU calibration obviously.

## 5 Conclusion

This paper mainly discusses a new iterative method for HEC problem and applies it to Lidar-IMU system. The IMU information is used for motion estimation, which effectively reduces the affect of radar distortion on calibration results, we can get a more accurate radar and IMU external parameters. Meanwhile, the algorithm also owns the high rapidity and robustness during execution. This lays a solid foundation for the design of integrated navigation system using Lidar information and IMU information in the future.

However, in this paper, the point cloud feature extraction is based on the simple curvature, so the feature is not obvious enough. In the future work, more constraints can be considered to solve problem of feature extracting.

**Acknowledgements.** This paper is supported by the National Natural Science Foundation of China (Grant No. 41874034), the National Science and Technology Major Project of the National Key R&D Program of China (Grant No. 2016YFB0502102), the Beijing Natural Science Foundation (Grant No. 4202041), and the Aeronautical Science Foundation of China.

## References

1. Qin, T., Li, P., Shen, S.: Vins-mono: a robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Rob.* **34**(4), 1004–1020 (2018)
2. Zhang, J., Singh, S.: Visual-lidar odometry and mapping: low-drift, robust, and fast, pp. 2174–2181 (2015)
3. Strobl, K.H., Hirzinger, G.: Optimal hand-eye calibration, pp. 4647–4653 (2006)
4. Zhang, Z.: A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(11), 1330–1334 (2000)
5. Geiger, W., Bartholomeyczik, J., Breng, U., Gutmann, W., Hafen, M., Handrich, E., Huber, M.F., Jackle, A., Kempfer, U., Kopmann, H., et al.: MEMS IMU for AHRS applications, pp. 225–231 (2008)
6. Alves, J.B.D.M., Lobo, J., Dias, J.: Camera-inertial sensor modelling and alignment for visual navigation. *Mach. Intell. Robot. Control* **5**(3), 103–112 (2003)
7. Lobo, J., Dias, J.: Relative pose calibration between visual and inertial sensors. *Int. J. Robot. Res.* **26**(6), 561–575 (2007)
8. Zhang, Q., Pless, R.: Extrinsic calibration of a camera and laser range finder (improves camera calibration), pp. 2301–2306 (2004)
9. Velas, M., Spanl, M., Materna, Z., Herout, A.: Calibration of RGB camera with velodyne lidar (2014)
10. Ishikawa, R., Oishi, T., Ikeuchi, K.: Lidar and camera calibration using motion estimated by sensor fusion odometry. *arXiv, Computer Vision and Pattern Recognition* (2018)
11. Zhang, J., Singh, S.: Loam: lidar odometry and mapping in real-time (2014)



# Design of Semi-physical Simulation System for Multi-target Attack Air Defense Missile Weapon System

Shujun Yang<sup>(✉)</sup>, Jianqiang Zheng, Qinghua Ma, Shuaiwei Wang, Jirong Ma, Haipeng Deng, and Yiming Liang

Xian Institution of Modern Control Technology, Xian 201848, China  
22406135@qq.com

**Abstract.** Based on the requirements of semi-physical simulation of air defense weapon system, the scheme, design and development of comprehensive simulation of radar system, fire control system and missile system are introduced. A multi-target attack integrated hwil simulation system is constructed. The system can simulate the whole procedure of multi-target attack of air defense weapons system and provide a research platform for system technical index matching.

**Keywords:** Multi-target attack · Semi-physical simulation · Air Defense Missile Weapon System

## 1 Introduction

Simultaneous guidance of multiple missiles has become a basic feature of air defense weapon combat. In the complex battlefield, multiple targets are found and tracked, multiple missiles are launched to attack different targets, which can achieve the purpose of launching first and pre-emptively, and greatly improve the combat effectiveness and survival ability of air defense weapons. The whole multi-target attack process involves multiple components of air defense weapon system, including radar system, fire control system, air defense missile system [1], etc. Therefore, the hwil simulation system of air defense weapon is a large simulation system.

## 2 Hwil Simulation System

Multi-target attack requires simultaneous tracking of multiple targets in the air, fire control calculation and simultaneous guidance of multiple missiles. The multi-target attack air defense weapon system consists of three key sub-systems: radar system, launcher system and missile system [2].

## 2.1 Rader System

Multi-objective radar system includes a search radar, phased array radar and command transmission system, its function is: on the distance and anti-aircraft missiles with a range that meet the needs of tracking multiple targets at the same time, the state of each target, threat degree grade is given, the information such as position, velocity, angle deviation are given at the same time. After the missile launch, phased array radar in tracking multiple targets at the same time, the instruction will send guidance of multiple missiles by wireless transmission system, guided multi-missile fly to multi-target. The radar system is simulated by industrial computer.

## 2.2 Fire Control System

The function of Multi-objective fire control system is to process multiple targets provided by phased array radar in parallel, such as threat judgment, attack sequencing, fire allocation, calculation of missile permissible launch area, calculation of fire transfer criteria and multi-target attack sequencing [2], calculation of missile and target matching, calculation of multiple guidance instructions, etc.

## 2.3 Air Defense Missile System

According to binding flights fly to specify the target of the fire control system, receiving instructions transmission system transmission of guidance information, combined with information such as roll gyro, damping gyro, eventually forming system control instruction, control instruction to steering gear, the servo drive control rudder deflection ji, correcting deviation, missile flight of finally realizes effectively combat [4]. The flight process of missile is divided into initial stage and guidance stage. The initial stage is also known as the power flight stage, that is, from the missile Jane out to the end of the booster engine work and separated from the main stage missile; The guidance segment is to guide the missile into the narrow beam or photoelectric field of phased array radar until it hits the target. The transition between the two stages is completed by the transition of the guidance source.

# 3 Composition and Function Analysis

## 3.1 Basic Function and Composition

On the one hand, the correctness of the workflow design of the multi-target attack weapon system, the correctness of the multi-target servo turning model [3], and the correctness of the multi-target multiple guidance model are verified; On the one hand, the mutual relationship between air defense missile and air defense weapon system is studied, including the coordination of technical indicators and interface matching among various links of multi-target attack, and the design of

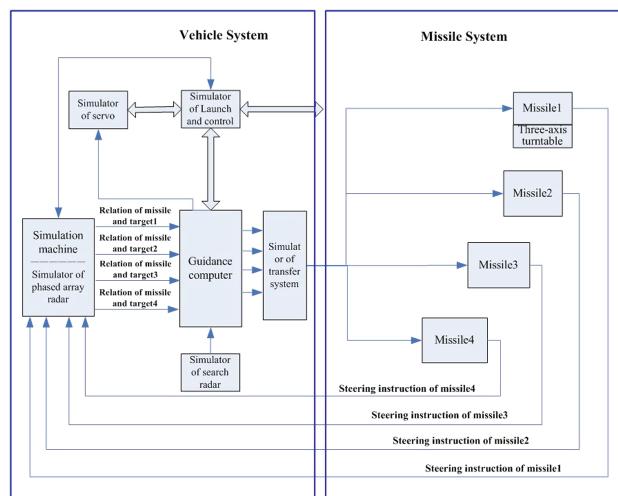
technical parameters of multi-target attack weapon system provides a platform for experimental research and verification.

Hwil simulation system of multi-target attack air defense weapon needs to simulate every link of the whole multi-target attack process and provide the same or similar test environment with the actual working condition [3]. In the system, the missile guidance and control system is used as the physical access simulation loop, and the guidance accuracy is used as the main index to evaluate the ability to carry out multi-target attack in the near real flight environment.

The hwil simulation system of multi-target attack air defense weapon needs to be simulated mainly including:

- Simulation of multi-target search and tracking process of search radar;
- Phased array radar provides multi-target characteristics, target movement and noise conditions;
- Simulation of development and control logic;
- Simulation of servo device;
- Simulation of air defense missile motion posture;
- Simulation of relative actuation characteristics.

According to the above analysis, the semi-physical simulation system of multi-target attack air defense weapon mainly consists of multi-target radar system simulator, development and control device simulator, servo device simulator, simulation computer, real-time network, internal rotation three-axis electric turntable, simulation test bench, data acquisition and recording equipment. The system composition block diagram is shown in Fig. 1.



**Fig. 1.** Block diagram of semi-physical simulation system of multi-target attack air defense weapon

### 3.2 Simulation Process of Multi-target

Attack the simulation process is consistent with the multi-target attack process of air defense missile, and the corresponding links are as follows [6]:

- a) The search radar simulator simulates multi-target search and air condition assessment to sort and number target tracks by threat level;
- b) The guidance computer carries out strike sequencing and fire diversion calculation on multiple targets sent by the search radar simulator;
- c) Determination and output of multi-target motion model completed by phased array radar simulator;
- d) the development and control simulator completes the self-inspection, binding parameters, bomb selection and issuing launch instructions of various ground equipment and missiles;
- e) The guidance computer completes the calculation of the launch area, the calculation of the multi-target attack task plan and the calculation of the missile target matching;
- f) The hair control simulator starts to judge the emission conditions and prompts to allow the emission;
- g) Press the missile to fire, and the transmitter and control simulator will send zero signal and complete ignition in time sequence;
- h) The guidance computer calculates the missile guidance instructions according to the target and missile information, and sends them to the missile through the instruction transmission simulator;
- i) The three-axis turntable simulates the missile attitude Angle change in the air according to the parameters provided by the simulator;
- j) The simulator conducts the next cycle calculation according to the missile data until the end of the simulation process and gives the simulation effect evaluation results.

## 4 The Key Technology

### 4.1 Multi-target Simulation Model

The key to the design of the simulation system is the clustering target model, which describes the behavior characteristics of the combat system [4]. The reliability of hwil simulation results of multi-target attack air defense weapons depends on the accuracy and precision of the description of target characteristics by the model. Therefore, to simulate the combat system, the system model is firstly established to objectively describe and quantify the battlefield target environment [5]. The main factors that determine the group-standard simulation model are:

- 1) classification of target motion model

According to the space condition of the target, it can be divided into uniform linear motion, uniform acceleration linear motion, subduction, turning motion and so on.

## 2) establishment of target motion model

To simplify the problem, it is assumed that all targets can be used as particle points and all target models are established in the same geographic coordinate system, and the uniform linear motion model, uniform acceleration linear motion model, compound curve motion model, subduction motion model and turning motion model can be established. In the sensor coordinate system, it is necessary to represent the random measurement errors of each independent measurement of the sensor, which is usually set as white noise.

## 4.2 Phased Array Radar Algorithm of Hand over to the Next Shift

The phased array radar beam rotation is flexible and can form multiple beams and fields of view in a short time. The multi-beam shift probability of air defense missile is an important index of missile guidance. In order to ensure strong robustness of the guidance shift, the principle of “strict in and strict out” is adopted. Under normal circumstances, it can ensure a successful shift and not repeat, and under abnormal circumstances, it is allowed to realize multi-field relay guidance for multiple shifts to improve the probability of successful shift.

In the process of hwil simulation, various abnormal conditions of phased array radar are analyzed in detail, and the error model and abnormal handling strategy are established to improve the guidance handover robustness. In addition, various shift situations are fully tested, the factors affecting the initial, middle and final shift of the missile are quantitatively studied, the specific gravity is determined, and the links needing improvement are found, so as to improve the combat effectiveness of the air defense missile and provide technical support.

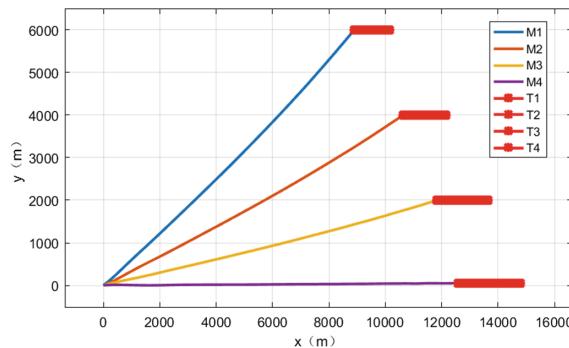
## 4.3 Real-Time Performance of Simulation System

System simulation is the basic characteristics and requirements of a material is introduced into the simulation circuit, simulation time scale required the same as the actual system time scale become real-time simulation, have very strong distribution by VMIC real-time system of data transmission, network protocol overhead is small, the hardware delays are also small, strong real-time performance, and adopt special VMIC real-time data communications interface to multiple target air defense weapon of choice for complex hardware-in-the-loop simulation system [1].

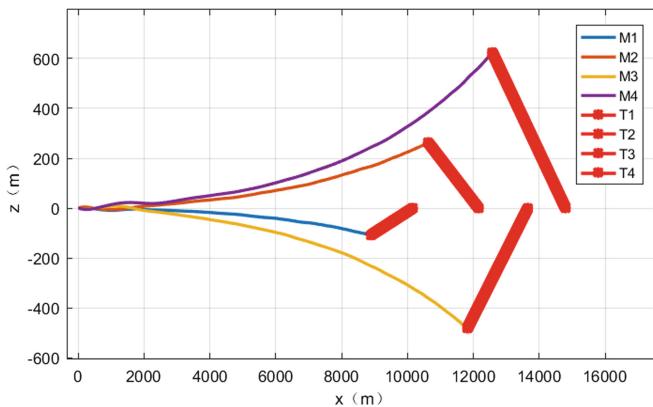
VMIC real-time network consists of reflective memory card, optical fiber and the HUB, is the high speed of information transmission channels between the simulation nodes, realize the data real-time transmission in the process of simulation test, on this basis, the real-time simulation computer control simulation time precision, real-time control in data-driven way each simulation equipment to the simulation time progressive scale synchronization.

## 5 The Results of Hwil Simulation

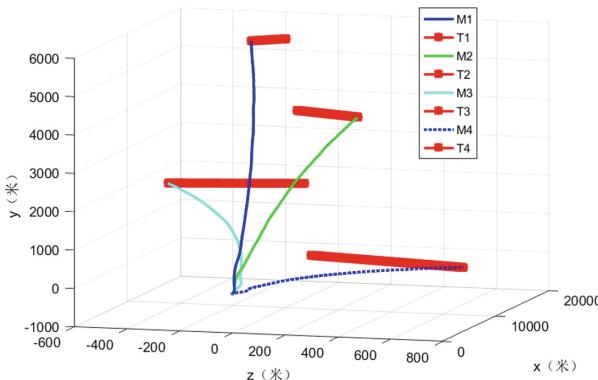
The multi-target attack semi-physical simulation system constructed in this paper is used to simulate the interception of four full-airspace cluster attack targets. Figures 2, 3 and 4 shows the position curves of the four targets attacked by four missiles. The simulation results meet the requirements of the weapon system and indexes of each measuring device.



**Fig. 2.** The vertical position curve of four targets and four missiles



**Fig. 3.** The route position curve of four targets and four missiles



**Fig. 4.** The space position curve of four targets and four missiles

## 6 Conclusion

The paper introduces a semi-physical simulation system of multi-target attack which combines radar system, fire control system and air defense missile system. The simulation results show that the system can realize the co-simulation of multi-target system, which provides an effective test method for the evaluation and appraisal of weapon system indexes, and has important guidance and reference value for the simulation research of similar multi-target attack weapons system.

## References

1. Liu, X.: Design and research of integrated hardware-in-the-loop simulation system for air to air multi-target interception and attack. *Obs. Control Technol.* **26**(1), 79–81 (2007)
2. Zhang, G.: How to deal with multi-target attack - one of the development directions of air defense system. *Fire Control Radar Technol.* **33**(1), 1–2 (2004)
3. Cheng, M.: Multi-objective simulation optimization of air defense combat system. PLA Air Force Engineering University (1991)
4. Jiang, H.: Simulation of fire group target assignment model for hybrid air defense missile. *J. Coll. Ordnance Eng.* **18**(1), 46–49 (2006)
5. Liu, W.: Scaling target echo modeling technology based on modified highlight model. *Torpedo Technol.* **17**(4), 20–24 (2009)
6. Lemma, A.N.: Multiresolution ESPRIT algorithm. *IEEE Trans. Signal Process.* **6**(6), 1722–1726 (1999)



# Application of Multi-network Fusion in Diagnosis of Chest Diseases

Shanshan Zhang<sup>(✉)</sup> and Hai Gao

School of Automation and Electrical Engineering,  
University of Science and Technology Beijing, Beijing 100083, China  
s20180607@xs.ustb.edu.cn, gaohai@ustb.edu.cn

**Abstract.** Image Classification is a hot research topic in the fields of Computer Vision and Medical Artificial Intelligence during the past few years. In this respect, X-ray is one of the most common radiological examinations in diagnosing chest diseases. However, there is still a great challenge for research. The reasons are as follows: 1) nuances in inter-class and large intra-class differences among fine-grained image sub-categories are existing, 2) dataset lacks accurate image annotations, 3) data is out-of-balance in most datasets. This paper presents a deep learning network fusion model that can automatically detect and classify Chest X-ray images. The model fuses three different neural network structures of Mobilenet-v2, Inception-v3 and A-Densenet-121 according to the weights to improve the chest disease diagnostic accuracy. The contribution of this work is to improve the structure of Densenet-121 and carry out network integration. To verify the reliability of the established network model, this paper conducted massive training and testing based on the ChestX-ray14 dataset. The experimental result shows that AUC is up to 0.8577, which outperforms the performance of most existing methods.

**Keywords:** Chest X-ray · Deep learning · Classification · Network fusion

## 1 Introduction

Since Chest X-ray has the properties of low cost, high speed and relatively less radiation, it has been one of the preferred diagnostic methods for patients [1]. It is widely used in preliminary screening and clinical practice of chest diseases. An X-ray image may contain a variety of diseases. More importantly, the diagnosis of diseases relies heavily on the expertise and experience of radiologists. Due to the complex correlation in different pathologies and the lack of radiologists, most computer-aided diagnosis (CAD) systems have been developed to help radiologists to enhance their diagnostic efficiency. CAD mainly provides simple visualization functions. Therefore, it is necessary for doctors to conduct a quantitative analysis of the features extracted from CAD. Considering that numerous and miscellaneous chest radiographs, the auxiliary effect of CAD will

become rather limited. As a result, radiologists still have large workload, which makes it essential to form a computer automatic diagnosis system.

In recent years, deep learning has made remarkable achievements in image processing, such as image classification, image segmentation [2] and target detection. The phenomenon has aroused interest in applying deep learning to medical images. Recent research results show that medical artificial intelligence has made progress in biomedical applications, such as tumor detection and nodule localization. For chest diseases, some studies have been done based on the existing public dataset. Li et al. proposed a solitary feature-based method for detecting lung nodules [3]. The method used stationary wavelet transform and convergence index filter to extract features and detected lung nodules on the JSRT dataset. In the previous work, a large number of parameters were generated and networks were always over-fitting on account of the use of deep network and training on small datasets.

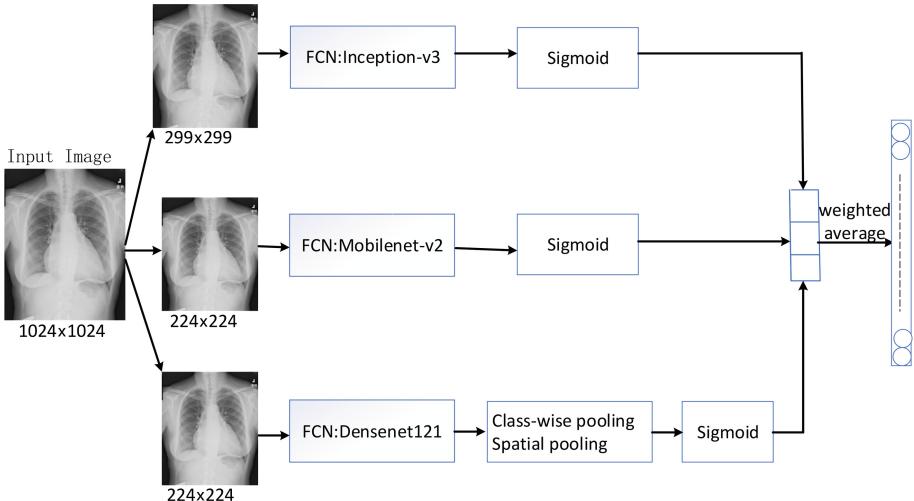
To solve the above problems, this paper proposes a fusion network that considers three aspects: depth, width and memory size. To reduce network parameters, Mobilenet-v2 [4] is selected. Meanwhile, Inception-v3 [5] and Densenet-121 [6] which can increase the network depth and width are quite necessary. In this article, the A-Densenet-121 network is proposed. The model uses Densenet-121 [6] as the backbone structure, and the attention module is added after it. Here are the processes of network training. Firstly, replace the classification layer of networks with sigmoid nonlinear function. Secondly, retrain these networks separately. Finally, average the outputs of networks by weights. Also, this paper adopts Ten-crop for data preprocessing.

## 2 Related Work

In research of deep learning, a well-prepared dataset is one of the essential factors. Owing to the privacy of patients' personal information and the time consuming of labeling data, datasets with large orders of magnitude and high labeling accuracy are rare. Most scholars utilized JSRT and Shenzhen dataset for experiments, but the small amount of data restricted their training on complex models. Recently, a large dataset ChestX-ray14 [7] is released. In the existing literature, a variety of methods have been proposed for the classification of ChestX-ray14 images. Wang et al. inserted a transition layer and a global pooling layer into CNNs, and compared the results of AlexNet, GoogleNet, VGGNet and ResNet running on ChestX-ray14 [7]. It is worth noting that previous studies have usually assumed that class labels are independent of each other, but they are related. Guan et al. proposed a category-wise residual attention learning framework for classification [8]. The model classifies multi-label chest radiographs by introducing the barrier of unrelated categories and increasing the weight of related categories. Ge et al. proposed two loss functions: multi-label Softmax Loss and Correlation Loss, which could be used in all the deep learning models and specialized in the images that are correlated among categories [9].

### 3 Approach

In this paper, the method we propose is an ensemble of three deep convolutional neural networks (DCNN), as shown in Fig. 1. Firstly, in order to make the network perform better, class-wise pooling and spatial pooling are added behind Densenet121's [6] fully convolutional network (FCN). Among them, class-wise pooling enables the network to obtain more discriminative features in the same category, and the spatial pooling selects the highest and lowest regional scores to make the classification more effective. Then, each network is trained in an end-to-end way to get the optimal weight parameters. After that, the output of Mobilenet-v2 [4], Inception-v3 [5] and A-Densenet-121 are fused to achieve the final result.



**Fig. 1.** The proposed fusion network structure.

#### 3.1 Neural Network Structure

**Mobilenet-v2.** Mobilenet-v2 [4] is a light weight convolutional neural network. Depthwise separable convolution is introduced into each bottleneck layer of the network. It consists of two parts: depthwise convolutions and pointwise convolutions. Compared with the traditional convolutional networks, this structure guarantees the accuracy, significantly decreases the computation and reduces the model size. Consequently, it is possible to deploy the convolution neural network with high real-time performance on mobile devices.

**Inception-v3.** Each input layer of the Inception network is connected with filters of different sizes, which increases the adaptability of the network to different scales and makes the feature expression ability stronger. Moreover, Inception network uses the  $1 \times 1$  convolution kernel so that it can shrinks network parameters through dimension reduction. The main improvement of Inception-v3 [5] is to split a large two-dimensional convolution into two smaller one-dimensional convolutions, such as the  $3 \times 3$  convolution can be replaced by a  $1 \times 3$  convolution followed by a  $3 \times 1$  convolution, which accelerates the operation and heightens the network nonlinearity.

**A-Densenet-121.** The A-Densenet-121 promotes network performance by adding class-wise pooling and spatial pooling. Class-wise pooling encodes the features generated by the fully convolutional network into  $M$  channels of each class through  $1 \times 1$  convolution, and then calculates the average value in the class channel by using the average pooling method, which reduces the size of the feature graph from  $w \times h \times MC$  to  $w \times h \times C$ . Since the maximum and minimum scoring areas are both important for the final prediction, the method of  $KMax + KMin$  is adopted for the spatial pooling in this paper. Besides, we introduce a factor  $\alpha$  to weigh the importance of both. The formula of spatial pooling is as follows:

$$s^c = \max_{h \in H_{k+}} \frac{1}{k^+} \sum_{i,j} h_{i,j} z_{i,j}^c + \alpha \min_{h \in H_{k-}} \frac{1}{k^-} \sum_{i,j} h_{i,j} z_{i,j}^c \quad (1)$$

Where  $z^c$  is feature map of class  $c$  after class-wise pooling.  $h \in H_k$  satisfies  $h_{i,j} \in (0,1)$  and  $\sum_{i,j} h_{i,j} = k$ . Spatial pooling selects the highest and lowest activation regions from  $z^c$ , the output  $s^c$  is the weighted average of scores of all the selected regions.

### 3.2 Fusion Method

To reduce the overfitting of networks, the output of each network that in this study was averaged by weights, as shown in formula 2. The weight parameters  $a$ ,  $b$  and  $c$  are used to control the influence of each model on the fusion network.

$$y^c = ay_1^c + by_2^c + cy_3^c \quad (2)$$

Where  $a$ ,  $b$ , and  $c$  satisfy  $a + b + c = 1$ .  $y_1^c$ ,  $y_2^c$ ,  $y_3^c$  respectively represent the output of the three networks, and  $y^c$  is the weighted average.

### 3.3 Loss Function

Since medical data is difficult to obtain and the incidence of various diseases is different, the majority of medical datasets is not uniformly distributed. The prediction effect of the model trained by unbalanced data is poor or completely unpredictable. Therefore, we adopt the weighted cross entropy loss function to

reduce the effect of data imbalance on training. The formula of loss function is as follows:

$$L = -\omega_1 \sum_{c=1}^M y_c \log p_c - \omega_0 \sum_{c=1}^M (1 - y_c) \log(1 - p_c) \quad (3)$$

$$\omega_1 = \frac{P + N}{P} \quad (4)$$

$$\omega_0 = \frac{P + N}{N} \quad (5)$$

where  $P$  and  $N$  represent the number of normal and abnormal in a batchsize, respectively.

## 4 Experiments

### 4.1 DataSet

This paper chooses ChestX-ray14 [7] as a designative dataset, which is provided by the National Institutes of Health (NIH). It contains 112,120 frontal-view X-ray images from 30,805 patients. Each image is annotated with 14 disparate thoracic pathology labels. Particularly, the researchers used the natural language processing (NLP) method to extract labels from radiological reports, which enabled the tag extraction accuracy to exceed 90%. According to the ratio of 7:2:1, the dataset is randomly divided into three parts: training, validation and test. There is no patient overlap among three sets.

### 4.2 Experimental Detail

**Train.** We initialize the network parameters through a pre-trained model on the ImageNet dataset. Since Inception-v3 [5] accepts input images with a resolution of 299, the input image is randomly cropped from the size of  $1024 \times 1024$  to  $299 \times 299$ . For Mobilenet-v2 [4] and A-Densenet-121, input images are cropped to  $224 \times 224$  in the same way. Besides, the image is normalized by using the mean and standard deviation on the ImageNet dataset. In order to improve the generalization ability, images are also flipped horizontally randomly. We train the model with batch size 16 and use Adam optimization function. Adam employs the default value (0.9,0.99) to calculate the coefficient of the running average value of the gradient and the square of the gradient. We initialize the learning rate at 0.0001. When the validation loss does not decrease within four steps, it will be decayed ten times from the current value.

**Test.** In the test, we first resize the image entered into Inception-v3 [5] to the resolution of  $341 \times 341$ , and then cut it to  $299 \times 299$  through Ten-crop. Similarly, images entered into the Mobilenet-v2 and the A-Densenet-121 are resized to  $256 \times 256$  and then cropped to  $224 \times 224$ . We search for the optimal weight from

running on the validation set, and apply these weights to the test set to obtain the final fusion result. Experiments show that the better classification result is obtained when the weights of Mobilenet-v2 [4], Inception-v3 [5] and A-Densenet-121 are 0.8, 0.1 and 0.1, respectively.

### 4.3 Our Experiments

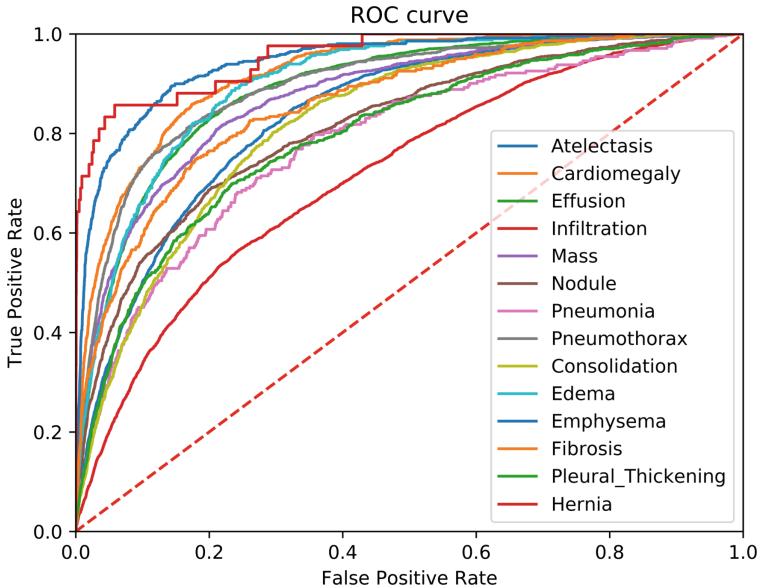
We use the area under the ROC curve (AUC) as the evaluation metric of classification performance. In Table 1, we present the result of the single network. Because of small volume and less computation, Mobilenet-v2 [4] has achieved a fine result in training. Inception-v3 [5] realizes parallel performance improvements by increasing the depth and width of the network, which makes the information extracted from the network richer. Therefore, Inception-v3 achieves the best classification result compared to other networks. Densenet-121 [6], proposed by CVPR2017, has been used in X-ray classification by some studies. In this paper, we introduce the attention network, A-Densenet-121, which is based on densenet121 and obtains a result of 0.8464. Since various networks may produce different overfitting, the average may make some opposite fitting cancel each other to reduce overfitting on the whole. Therefore, this paper adopts the method of network fusion to decrease overfitting.

**Table 1.** Per-class AUC of our method

Pathology	A-Densenet-121	Mobilenet [4]	Inception [5]	Ensemble
Atelectasis	0.8314	0.8173	0.8314	0.8372
Cardiomegaly	0.9151	0.9077	0.9117	0.9199
Effusion	0.8860	0.8803	0.8864	0.8885
Infiltration	0.7133	0.7130	0.7163	0.7189
Mass	0.8686	0.8477	0.8700	0.8741
Nodule	0.8000	0.7822	0.8109	0.8131
Pneumonia	0.7687	0.7704	0.7720	0.7818
Pneumothorax	0.8878	0.8823	0.8883	0.8968
Consolidation	0.8110	0.8107	0.8202	0.8222
Edema	0.8952	0.8856	0.8978	0.9003
Emphysema	0.9360	0.9289	0.9407	0.9435
Fibrosis	0.8459	0.8457	0.8584	0.8590
Pleural thickening	0.7788	0.7900	0.7901	0.7965
Hernia	0.9126	0.9175	0.9489	0.9558
Average	0.8465	0.8414	0.8531	0.8577

#### 4.4 Comparison Experiments

In this paper, the performance of network fusion is mainly studied in classification. Therefore, the A-Densenet-121 has a standard multi-label classification with  $M = 1$ . For spatial pooling, we select the maximum and minimum scoring areas from the feature map, and average the selected scores by introducing an impact factor  $\alpha$ . According to reference [10], we set the parameters as  $k^+ = k^- = 1$  and  $\alpha = 0.7$ . Compared with CheXNet [11], the performance of this network is improved by 0.62%. The average AUC obtained by merging the improved network with the other two networks was 0.8577. The resulting figure is shown in Fig. 2.



**Fig. 2.** ROC curve of Mobilenet-v2, Inception-v3 and A-Densenet-121 network fusion.

In Table 2, we compare the result of the network fusion with previous studies on ChestX-ray14 [7]. Yao et al. obtained a classification result of 0.8027 by using the dependency among labels [12]. Based on Densenet-121, CheXNet adopted the method of oversampling and obtained the result of 0.8414 [11]. For fair comparison, CheXNet was retrained according to our data processing method and the result was 0.8402. Huang et al. added class-wise pooling and top-k pooling after the FCN of Densenet-121, and got the result 0.8501 [13]. Our network ensemble method obtained an AUC value of 0.8577. Compared with the above four methods, the performance improvements of the proposed method are 5.50%, 1.75% and 0.76%, respectively. Above all, it proves that our method has better performance in the classification of chest diseases.

**Table 2.** Performance comparison of different architectures

Pathology	Yao [12]	CheXNet [11]	Huang [13]	Ensemble
Atelectasis	0.7720	0.8191	0.8297	0.8372
Cardiomegaly	0.9040	0.9138	0.9155	0.9199
Effusion	0.8590	0.8838	0.8878	0.8885
Infiltration	0.6950	0.7065	0.7115	0.7189
Mass	0.7920	0.8566	0.8619	0.8741
Nodule	0.7170	0.7879	0.8083	0.8131
Pneumonia	0.7130	0.7668	0.7809	0.7818
Pneumothorax	0.8410	0.8711	0.8795	0.8968
Consolidation	0.7880	0.8117	0.8115	0.8222
Edema	0.8820	0.8902	0.8992	0.9003
Emphysema	0.8290	0.9213	0.9387	0.9435
Fibrosis	0.7670	0.8292	0.8370	0.8590
Pleural thickening	0.7650	0.7796	0.7906	0.7965
Hernia	0.9140	0.9253	0.9492	0.9558
Average	0.8027	0.8402	0.8501	0.8577

## 5 Conclusion

This paper presents a new network fusion algorithm to detect and classify Chest X-ray images. This model uses a new A-Densenet-121 network and averages the output of each network by weights. The result that AUC is to 0.8577 proves the superiority of the method. In the future research, dependency among labels may be used for better classification. Besides, the attention module will be further optimized.

## References

1. Rakshit, S., Saha, I., Wlasnowolski, M., Maulik, U., Plewczynski, D.: Deep learning for detection and localization of thoracic diseases using chest X-ray imagery. In: International Conference on Artificial Intelligence and Soft Computing, pp. 271–282. Springer, Cham, June 2019
2. Dai, J., He, K., Li, Y.: Instance-sensitive fully convolutional networks. In: European Conference on Computer Vision, pp. 534–549. Springer, Cham, October 2016
3. Li, X., Shen, L., Luo, S.: A solitary feature-based lung nodule detection approach for chest X-ray radiographs. IEEE J. Biomed. Health Inform. **22**(2), 516–524 (2017)
4. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Adam, H.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint [arXiv:1704.04861](https://arxiv.org/abs/1704.04861) (2017)
5. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)

6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
7. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106 (2017)
8. Guan, Q., Huang, Y.: Multi-label chest x-ray image classification via category-wise residual attention learning. *Pattern Recogn. Lett.* **130**, 259–266 (2020)
9. Ge, Z., Mahapatra, D., Chang, X., et al.: Improving multi-label chest X-ray disease diagnosis by exploiting disease and health labels dependencies. *Multimedia Tools Appl.* **79**, 14889–14902 (2020)
10. Durand, T., Mordan, T., Thome, N., Cord, M.: Wildcat: weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 642–651 (2017)
11. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Lungren, M.P.: Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint [arXiv:1711.05225](https://arxiv.org/abs/1711.05225) (2017)
12. Li, Y., Eric, P., Dmitry, D., Ben, C., Devon, B., Kevin, L.: Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint [arXiv:1710.10501](https://arxiv.org/abs/1710.10501) (2017)
13. Huang, Z., Fu, D.: Diagnose chest pathology in X-ray images by learning multi-attention convolutional neural network. In: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), pp. 294–299. IEEE, May 2019



# Distributed Robust $H_\infty$ Containment Control for Fractional-Order Multi-agent Networks

Xiaolin Yuan<sup>1</sup>, Yongguang Yu<sup>1(✉)</sup>, and Lipo Mo<sup>2</sup>

<sup>1</sup> Department of Mathematics,

Beijing Jiaotong University, Beijing 100044, People's Republic of China  
ygyu@bjtu.edu.cn

<sup>2</sup> School of Mathematics and Statistics, Beijing Technology and Business University,  
Beijing 100048, People's Republic of China

**Abstract.** The objective of this paper is to study the distributed robust  $H_\infty$  containment control problem for fractional-order multi-agent networks (FOMANs) with modeling errors and external disturbances over a digraph. Instead of studying the original fractional-order closed-loop systems directly, the equivalent actual infinite-dimensional state-space model is established for studying. Then, a linear matrix inequality (LMI) condition is derived for achieving containment control under the robust  $H_\infty$  performance. Furthermore, with the help of traditionally Lyapunov stability theory, the convergence analysis of the closed-loop systems is completed. Finally, the results are illustrated by numerical examples.

**Keywords:** Robust  $H_\infty$  · Containment control · Fractional-order · Multi-agent networks

## 1 Introduction

The coordination control of multi-agent networks (MANs) has always been a research hot spot [1–3], where the containment control has been regarded as a fundamental control problem. The objective of containment control is to force all followers to move into the convex hull spanned by multiple leaders. In [4], the containment control of MANs with multiple stationary or dynamic leaders was investigated. In [5], the distributed tracking was investigated for the convex hull spanned by dynamic leaders whose speed was not measurable, and so on [6, 7]. Recently, fractional calculus attracts more attention in various fields because of its characteristics of memory and heredity. Notice that when agents moving in some complex environments, the dynamics of agent can be modeled by fractional-order differential equations more accurately, and hence the model of FOMANs was established [8]. And then, some necessary and sufficient conditions for containment control of FOMANs were deduced in [9]. The quasi-containment control problem of FOMANs was solved in [10], and so on [11].

Also, in actual environments, due to the uncertain factors such as the change of physical environment factors, external disturbances, and the dynamics of the system without modeling, the accurate information of the agent may not be directly obtained by the controller, which may affect the convergence performance of the system. Robustness refers to the characteristic that the control system maintains some other performance under certain parameter perturbations. Therefore, it is necessary to do relevant research to maintain the stability of the system and improve the robustness of FOMANs. At the end of the 20th century, Zames [12] chosen  $H$ -norm as the performance index and proposed the  $H_\infty$  control problem, namely the minimum sensitivity control problem.

Motivated by the above discussions, the distributed robust  $H_\infty$  containment control of FOMANs with model errors and external disturbances is studied. Compared with [9–11], where the containment control of FOMANs was investigated, however, the effect of modeling errors and external disturbances on the stability of the systems was not considered. Therefore, this paper makes some improvements. The main contributions of this paper are summarized as follows: (1) The robust  $H_\infty$  containment control of FOMANs is investigated; (2) To simplify the analysis, the original closed-loop systems is transformed into an equivalent state-space model; (3) A LMI condition is deduced for achieving containment under  $H_\infty$  performance; (4) The traditionally Lyapunov stability theory is used for convergence analysis.

## 2 Preliminaries

**Definition 1** [13]. Define the  $\alpha$ -order Caputo's fractional derivative for a function  $\varphi(t) \in C^1[[t_0, t], \mathbf{R}]$  as:

$${}_{t_0}^C D_t^\alpha \varphi(t) = \frac{1}{\Gamma(1-\alpha)} \int_{t_0}^t \frac{\varphi'(s)}{(t-s)^\alpha} ds,$$

where  $\alpha \in (0, 1)$ ,  $t_0 \leq t$ ,  $\Gamma(\cdot)$  is Gamma function.

**Lemma 1** [14]. The Caputo fractional differential equation  ${}_{t_0}^C D_t^\alpha \varphi(t) = g(\varphi(t))$  is equivalent to

$$\begin{cases} \frac{\partial \mu(\zeta, t)}{\partial t} = -\zeta \mu(\zeta, t) + g(\varphi(t)), \\ \varphi(t) = \int_0^\infty S(\zeta) \mu(\zeta, t) d\zeta, \quad \text{with } \varphi(0)\delta(\zeta) = S(\zeta)\mu(\zeta, 0), \end{cases}$$

where  $\alpha \in (0, 1)$ ,  $\delta(\cdot)$  is Dirac delta function,  $S(\zeta) = \frac{\sin \alpha \pi}{\pi} \zeta^{-\alpha}$  is the pseudo Laplace transform of impulsive response,  $\mu(\zeta, t)$  is the pseudo state variable.

**Lemma 2** [15]. For a symmetric block matrix  $m = \begin{pmatrix} m_1 & m_2 \\ m_3 & m_4 \end{pmatrix}$ , where  $m_1, m_4$  are invertible,  $m_2^T = m_3$ . If  $m_1 < 0$ ,  $m_4 - m_2^T m_1^{-1} m_2 < 0$  or  $m_4 < 0$ ,  $m_1 - m_2 m_4^{-1} m_2^T < 0$ , then  $m < 0$ . Vice versa.

### 3 System Model and Problem Formulation

This paper considers the FOMANs over a digraph  $\mathcal{G} = (\mathcal{N}, \varepsilon, \mathcal{A})$  with  $a \geq 1$  followers and  $b > 1$  leaders. Let  $\mathcal{N}_1 = \{1, 2, \dots, a\}$  and  $\mathcal{N}_2 = \{a+1, a+2, \dots, a+b\}$  respectively represents the set of all followers and leaders, and  $\mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2$ . The Laplacian  $L$  of  $\mathcal{G}$  can be partitioned into  $L = \begin{bmatrix} L_1 & L_2 \\ O_1 & O_2 \end{bmatrix}$ , where  $L_1 \in \mathbf{R}^{a \times a}$ ,  $L_2 \in \mathbf{R}^{a \times b}$ ,  $O_1 \in \mathbf{R}^{b \times a}$ ,  $O_2 \in \mathbf{R}^{b \times b}$ .

**Definition 2.** *The agent that can only send information is called leader, and the agent that can receive and send information is called follower.*

**Assumption 1.** In  $\mathcal{G}$ ,  $\forall i \in \mathcal{N}_1$ ,  $\exists j \in \mathcal{N}_2$ , such that there is a directed path from  $j$  to  $i$ .

**Lemma 3** [4]. *Under Assumption 1,  $L_1^{-1}$  exists, and  $\Re(\lambda_i(L_1)) > 0$ . Besides,  $-L_1^{-1}L_2 \geq 0$  and  $-L_1^{-1}L_2\mathbf{1}_b = \mathbf{1}_a$ .*

Suppose that the dynamics of each agent are

$$\begin{cases} {}_0^C D_t^\alpha \varphi_i(t) = (A + \Delta B(t))\varphi_i(t) + I_i(t) + \vartheta_i(t), & i \in \mathcal{N}_1, \\ {}_0^C D_t^\alpha \varphi_i(t) = (A + \Delta B(t))\varphi_i(t), & i \in \mathcal{N}_2, \end{cases} \quad (1)$$

where  $\alpha \in (0, 1)$ ,  $\varphi_i(t) \in \mathbf{R}^n$  and  $I_i(t) \in \mathbf{R}^n$  respectively represents the  $i$ th agent's state and control input;  $A \in \mathbf{R}^{n \times n}$  represents the system matrix;  $\vartheta_i(t) \in \mathbf{R}^n$  represents the external disturbance;  $\Delta B(t) \in \mathbf{R}^{n \times n}$  represents the modeling error, and  $\Delta B(t) = F_1 M(t) F_2$ , where  $F_1 \in \mathbf{R}^{n \times p}$ ,  $F_2 \in \mathbf{R}^{q \times n}$  are known,  $M(t) \in \mathbf{R}^{p \times q}$  is unknown and satisfy  $M^T(t)M(t) \leq I_q$ .

**Assumption 2.**  $\int_0^\infty \vartheta_i^T(t)\vartheta_i(t)dt < \infty$ .

**Definition 3.** [16]. *The convex hull spanned by all leaders is the minimal convex set containing all leaders, denoted by  $\text{Co}\{\varphi_j(t) \mid j \in \mathcal{N}_2\} = \{\sum_{j=1}^b k_j \varphi_{a+j} \mid k_j \in \mathbf{R}, k_j \geq 0, \sum_{j=1}^b k_j = 1\}$ .*

This paper takes the following controller:

$$I_i(t) = G \sum_{j \in N_i} a_{ij}(\varphi_j(t) - \varphi_i(t)), \quad i \in \mathcal{N}_1, \quad (2)$$

where  $N_i$  is the neighbor set of agent  $i$ ,  $G \in \mathbf{R}^{n \times n}$  is the undetermined control gain matrix. Take (2), then (1) could be written as

$$\begin{cases} {}_0^C D_t^\alpha \varphi_i(t) = (A + \Delta B(t))\varphi_i(t) + G \sum_{j \in N_i} a_{ij}(\varphi_j(t) - \varphi_i(t)) + \vartheta_i(t), & i \in \mathcal{N}_1, \\ {}_0^C D_t^\alpha \varphi_i(t) = (A + \Delta B(t))\varphi_i(t), & i \in \mathcal{N}_2. \end{cases} \quad (3)$$

### 3.1 Main Results

Let  $\xi_1(t) = (\varphi_1^T(t), \varphi_2^T(t), \dots, \varphi_a^T(t))^T$ ,  $\xi_2(t) = (\varphi_{a+1}^T(t), \varphi_{a+2}^T(t), \dots, \varphi_{a+b}^T(t))^T$ ,  $\vartheta(t) = (\vartheta_1^T(t), \vartheta_2^T(t), \dots, \vartheta_a^T(t))^T$ . Then write (3) in compact form as

$${}_0^C D_t^\alpha \xi_1(t) = (I_a \otimes (A + \Delta B(t))) \xi_1(t) - (L_1 \otimes G) \xi_1(t) - (L_2 \otimes G) \xi_2(t) + \vartheta(t),$$

$${}_0^C D_t^\alpha \xi_2(t) = (I_b \otimes (A + \Delta B(t))) \xi_2(t).$$

Define the following output function:

$$P(t) = \xi_1(t) + (L_1^{-1} L_2 \otimes I_n) \xi_2(t), \quad (4)$$

where  $P(t) = (p_1^T(t), p_2^T(t), \dots, p_a^T(t))$ ,  $p_i(t) \in \mathbf{R}^n$  represents the output of the  $i$  follower. According to Lemma 3,  $(-L_1^{-1} L_2 \otimes I_n) \mathbf{1}_{b \times n} = \mathbf{1}_{a \times n}$ , therefore if  $\lim_{t \rightarrow +\infty} P(t) = 0$ , i.e.,  $\lim_{t \rightarrow +\infty} \xi_1(t) = -\lim_{t \rightarrow +\infty} (L_1^{-1} L_2 \otimes I_n) \xi_2(t)$ , which implies that  $\lim_{t \rightarrow +\infty} \varphi_i(t) = \text{Co}\{\varphi_j(t) \mid j \in \mathcal{N}_2\}$ ,  $\forall i \in \mathcal{N}_1$ , then the containment control of the systems is achieved. Define  $\eta_i(t) = \sum_{j \in \mathcal{N}_i} a_{ij}(\varphi_i(t) - \varphi_j(t))$ ,  $i \in \mathcal{N}_1$ . Let  $\eta(t) = (\eta_1^T(t), \eta_2^T(t), \dots, \eta_a^T(t))^T$ , then write  $\eta_i(t)$  in compact form as

$$\begin{aligned} \eta(t) &= (L_1 \otimes I_n) \xi_1(t) + (L_2 \otimes I_n) \xi_2(t) \\ &= (L_1 \otimes I_n)[\xi_1 + (L_1^{-1} L_2 \otimes I_n) \xi_2(t)] = (L_1 \otimes I_n) P(t). \end{aligned} \quad (5)$$

According to Lemma 3,  $L_1$  is invertibility, therefore  $P(t) = (L_1 \otimes I_n)^{-1} \eta(t)$ . If  $\lim_{t \rightarrow +\infty} \eta(t) = 0$ , then  $\lim_{t \rightarrow +\infty} P(t) = 0$ .

For a given positive scalar  $\gamma \in (0, 1)$ , define the  $H_\infty$  performance index as

$$J(\vartheta) = \int_0^\infty [P^T(t) P(t) - \gamma^2 \vartheta^T(t) \vartheta(t)] dt. \quad (6)$$

**Definition 4.** The robust  $H_\infty$  containment control is said to be achieved if the containment control is achieved with  $\vartheta(t) = 0$ , and  $J(\vartheta) < 0$  with  $\vartheta(t) \neq 0$  and zero-valued initial conditions.

**Theorem 1.** Under Assumptions 1 and 2, with  $\eta(0)\delta(\zeta) = S(\zeta)\mu(\zeta, 0)$ , where  $\delta(\zeta)$ ,  $S(\zeta)$ ,  $\mu(\zeta, 0)$  will be defined later. Take the controller (2) with  $G = Q^{-1}$ , if there exist  $\varsigma > 0$  and symmetric definite matrix  $Q \in \mathbf{R}^{n \times n}$  satisfies the following LMI:

$$\begin{pmatrix} A & L_1 \otimes Q & I_a \otimes Q F_1 \\ L_1^T \otimes Q & -\gamma^2 I_{a \times n} & 0 \\ I_a \otimes F_1^T Q & 0 & -\varsigma I_{a \times p} \end{pmatrix} < 0, \quad (7)$$

where  $\Lambda = ((L_1 \otimes I_n)^{-1})^T (L_1 \otimes I_n)^{-1} + I_a \otimes (Q A + A^T Q) - 2L_1 \otimes I_n + I_a \otimes \varsigma F_2^T F_2$ , then the FOMANs (1) achieve robust  $H_\infty$  containment control.

*Proof.* Take  $\alpha$ -order Caputo's fractional derivative for  $\eta(t)$ , we have

$${}^C D_t^\alpha \eta(t) = (I_a \otimes (A + \Delta B(t)) - L_1 \otimes G)\eta(t) + (L_1 \otimes I_n)\vartheta(t). \quad (8)$$

According to Lemma 1, (8) is equivalent to

$$\begin{cases} \frac{\partial \mu(\zeta, t)}{\partial t} = -\zeta \mu(\zeta, t) + (I_a \otimes (A + \Delta B(t)) - L_1 \otimes G)\eta(t) + (L_1 \otimes I_n)\vartheta(t), \\ \eta(t) = \int_0^\infty S(\zeta) \mu(\zeta, t) d\zeta, \end{cases}$$

where  $S(\zeta) = \frac{\sin \alpha \pi}{\pi} \zeta^{-\alpha}$ . Correspondingly,

$$\begin{cases} \frac{\partial \mu^T(\zeta, t)}{\partial t} = -\zeta \mu^T(\zeta, t) + \eta^T(t)(I_a \otimes (A + \Delta B(t)) - L_1 \otimes G)^T + \vartheta^T(t)(L_1^T \otimes I_n), \\ \eta^T(t) = \int_0^\infty S(\zeta) \mu^T(\zeta, t) d\zeta. \end{cases}$$

Construct the following Lyapunov function

$$V(\mu(\zeta, t)) = \int_0^\infty S(\zeta) \mu^T(\zeta, t) (I_a \otimes Q) \mu(\zeta, t) d\zeta.$$

(i) First, consider the zero-input response, i.e.,  $\vartheta_i(t) = 0$ . Take derivative for  $V(\mu(\zeta, t))$  with respect to  $t$ , and together with  $S(\zeta)\zeta \geq 0$ ,  $G = Q^{-1}$ , then

$$\begin{aligned} \dot{V}(\mu(\zeta, t)) &= \int_0^\infty S(\zeta) [\frac{\partial \mu^T(\zeta, t)}{\partial t} (I_a \otimes Q) \mu(\zeta, t) + \mu^T(\zeta, t) (I_a \otimes Q) \frac{\partial \mu(\zeta, t)}{\partial t}] d\zeta \\ &= \int_0^\infty S(\zeta) [-\zeta \mu^T(\zeta, t) + \eta^T(t)(I_a \otimes (A + \Delta B(t)) - L_1 \otimes G)^T] (I_a \otimes Q) \mu(\zeta, t) d\zeta \\ &\quad + \int_0^\infty S(\zeta) \mu^T(\zeta, t) (I_a \otimes Q) [-\zeta \mu(\zeta, t) + (I_a \otimes (A + \Delta B(t)) - L_1 \otimes G)\eta(t)] d\zeta \\ &= -2 \int_0^\infty S(\zeta) \zeta \mu^T(\zeta, t) (I_a \otimes Q) \mu(\zeta, t) d\zeta \\ &\quad + \eta^T(t)(I_a \otimes (A + \Delta B(t)) - L_1 \otimes G)^T (I_a \otimes Q) \int_0^\infty S(\zeta) \mu(\zeta, t) d\zeta \\ &\quad + \int_0^\infty S(\zeta) \mu^T(\zeta, t) d\zeta (I_a \otimes (A + \Delta B(t)) - L_1 \otimes G)\eta(t) \\ &\leq \eta^T(t)[I_a \otimes (QA + A^T Q) - 2(L_1 \otimes I_n)]\eta(t) + 2\eta^T(t)[I_a \otimes QF_1 M(t)F_2]\eta(t) \\ &\leq \eta^T(t)[I_a \otimes (QA + A^T Q) - 2(L_1 \otimes I_n) + I_a \otimes \frac{1}{\varsigma} QF_1 F_1^T Q + I_a \otimes \varsigma F_2^T F_2]\eta(t) \\ &= \eta^T(t)[\Lambda - ((L_1 \otimes I_n)^{-1})^T (L_1 \otimes I_n)^{-1} + I_a \otimes \frac{1}{\varsigma} QF_1 F_1^T Q]\eta(t). \end{aligned}$$

According to Lemma 2, condition (7) is equivalent to

$$\begin{pmatrix} \Lambda + I_a \otimes \frac{1}{\varsigma} QF_1 F_1^T Q & L_1 \otimes Q \\ L_1^T \otimes Q & -\gamma^2 I_{a \times n} \end{pmatrix} < 0$$

and  $\Lambda + I_a \otimes \frac{1}{\varsigma} QF_1 F_1^T Q < 0$ . Since  $((L_1 \otimes I_n)^{-1})^T (L_1 \otimes I_n)^{-1} > 0$ , then  $\Lambda - ((L_1 \otimes I_n)^{-1})^T (L_1 \otimes I_n)^{-1} + I_a \otimes \frac{1}{\varsigma} QF_1 F_1^T Q < 0$ . Therefore, when condition (7) is satisfied, we have  $\dot{V}(\mu(\zeta, t)) < 0$ , therefore  $\lim_{t \rightarrow +\infty} \mu(\zeta, t) = 0$ . According to the relationship between real state and pseudo state [17], we have  $\lim_{t \rightarrow +\infty} \eta(t) =$

$\lim_{t \rightarrow +\infty} \int_0^\infty S(\zeta) \mu(\zeta, t) d\zeta = \int_0^\infty S(\zeta) \lim_{t \rightarrow +\infty} (\mu(\zeta, t)) d\zeta = \int_0^\infty S(\zeta) \cdot 0 d\zeta = 0$ , then  
 $\lim_{t \rightarrow +\infty} P(t) = \lim_{t \rightarrow +\infty} (L_1 \otimes I_n)^{-1} \eta(t) = 0$ , therefore the containment control of the systems is achieved.

(ii) Next, consider the zero-state response of the system (8), i.e.,  $\vartheta_i(t) \neq 0$ . Take derivative for  $V(\mu(\zeta, t))$  with respect to  $t$ , together with  $S(\zeta)\zeta \geq 0$  and  $G = Q^{-1}$ , then

$$\begin{aligned} \dot{V}(\mu(\zeta, t)) &\leq \eta^T(t)[I_a \otimes (QA + A^TQ) - 2(L_1 \otimes I_n) + I_a \otimes \frac{1}{\zeta} Q F_1 F_1^T Q \\ &+ I_a \otimes \zeta F_2^T F_2] \eta(t) + \eta^T(t)(L_1 \otimes Q) \vartheta(t) + \vartheta^T(t)(L_1^T \otimes Q) \eta^T(t) = \Xi(t) \Omega_1 \Xi^T(t), \end{aligned}$$

where  $\Xi(t) = [\eta^T(t), \vartheta^T(t)]$ ,  $\Omega_1 = \begin{pmatrix} \Lambda_1 + I_a \otimes \frac{1}{\zeta} Q F_1 F_1^T Q & L_1 \otimes Q \\ L_1^T \otimes Q & 0 \end{pmatrix}$ ,  $\Lambda_1 = I_a \otimes (QA + A^TQ) - 2L_1 \otimes I_n + I_a \otimes \zeta F_2^T F_2$ .

The performance index could be written as

$$\begin{aligned} J_T &= \int_0^T [P^T(t)P(t) - \gamma^2 \vartheta^T(t) \vartheta(t)] dt \\ &= \int_0^T [\eta^T(t)((L_1 \otimes I_n)^{-1})^T (L_1 \otimes I_n)^{-1} \eta(t) - \gamma^2 \vartheta^T(t) \vartheta(t) + \dot{V}(\mu(\zeta, t))] dt \\ &\quad - V(\mu(\zeta, T)) + V(0) \quad (V(0) = 0 \text{ obviously}) \\ &\leq \int_0^T \{\eta^T(t)[((L_1 \otimes I_n)^{-1})^T (L_1 \otimes I_n)^{-1} + I_a \otimes (QA + A^TQ) - 2L_1 \otimes I_n \\ &\quad + I_a \otimes \frac{1}{\zeta} Q F_1 F_1^T Q + I_a \otimes \zeta F_2^T F_2] \eta(t) + \eta^T(t)(L_1 \otimes Q) \vartheta(t) \\ &\quad + \vartheta^T(t)(L_1^T \otimes Q) \eta^T(t) - \gamma^2 \vartheta^T(t) \vartheta(t)\} dt \\ &= \int_0^T \Xi(t) \Omega \Xi^T(t) dt, \end{aligned}$$

where  $\Omega = \begin{pmatrix} \Lambda + I_a \otimes \frac{1}{\zeta} Q F_1 F_1^T Q & L_1 \otimes Q \\ L_1^T \otimes Q & -\gamma^2 I_{a \times n} \end{pmatrix}$ .

When condition (7) satisfied, we have  $\Omega < 0$  according to Lemma 2. On the one hand,  $J_T = \int_0^T [P^T(t)P(t) - \gamma^2 \vartheta^T(t) \vartheta(t)] dt < 0$ , let  $T \rightarrow +\infty$ , then  $\|P(t)\|_2^2 < \gamma^2 \|\vartheta(t)\|_2^2$ , i.e., the  $H_\infty$  performance is satisfied;

From (i) and (ii), it could deduce that the FOMANs (1) achieve robust  $H_\infty$  containment control. The proof of Theorem 1 is completed.

*Remark 1.* Here we make an analysis for the feasibility of LMI condition (7). According to Lemma 2, (7) is equivalent to  $\Omega < 0$ , and  $\Omega < 0$  is equivalent to

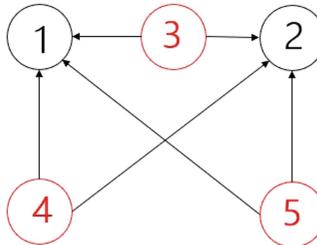
$$\begin{aligned} &\Lambda + I_a \otimes \frac{1}{\zeta} Q F_1 F_1^T Q + \frac{1}{\gamma^2} (L_1 \otimes Q)(L_1 \otimes Q)^T \\ &= -2L_1 \otimes I_n + ((L_1 \otimes I_n)^{-1})^T (L_1 \otimes I_n)^{-1} + I_a \otimes (QA + A^TQ) + I_a \otimes \zeta F_2^T F_2 \\ &\quad + I_a \otimes \frac{1}{\zeta} Q F_1 F_1^T Q + \frac{1}{\gamma^2} (L_1 \otimes Q)(L_1 \otimes Q)^T < 0. \end{aligned}$$

According to Lemma 3, all eigenvalues of  $-2L_1 \otimes I_n$  have negative real parts, i.e.,  $-2L_1 \otimes I_n$  is Hurwitz. Therefore, there must exists symmetric definite matrix  $Q \in \mathbf{R}^{n \times n}$  satisfies  $\Lambda + I_a \otimes \frac{1}{\zeta} Q F_1 F_1^T Q + \frac{1}{\gamma^2} (L_1 \otimes Q)(L_1 \otimes Q)^T < 0$ , which implies that condition (7) is feasible.

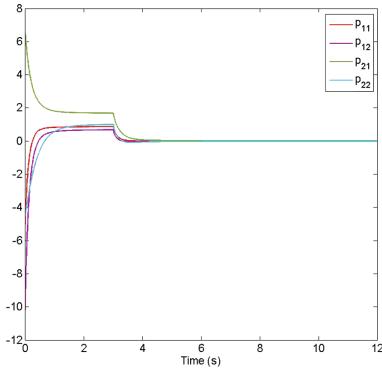
## 4 Simulations

This section gives simulations to verify the correctness of the theoretical results. Consider the FOMANs (1) with the protocol (2) over the digraph that depicted in Fig. 1, where  $\mathcal{N}_1 = \{1, 2\}$ ,  $\mathcal{N}_2 = \{3, 4, 5\}$ , Assumption 1 is satisfied obviously. From Fig. 1, we get  $L_1 = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$ ,  $L_2 = \begin{pmatrix} -1 & -1 & -1 \\ 0 & -1 & -1 \end{pmatrix}$ . Take  $\alpha = 0.95$ , iterative step size  $T = 0.01$ s,  $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ ,  $F_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ ,  $F_2 = (1 \ 1)$ , suppose that if  $\text{mod}(t/T, 5) = 1$ , then  $M(t) = 1$ ; if  $\text{mod}(t/T, 5) = 2$ , then  $M(t) = 1/2$ ; if  $\text{mod}(t/T, 5) = 3$ , then  $M(t) = 1/3$ ; if  $\text{mod}(t/T, 5) = 4$ , then  $M(t) = 1/4$ ; if  $\text{mod}(t/T, 5) = 5$ , then  $M(t) = 1/5$ . It could find  $\varsigma = 1$  and symmetric positive definite matrix  $Q = \begin{pmatrix} 0.5 & -0.15 \\ -0.15 & 0.6 \end{pmatrix}$  satisfying the condition (7), then construct  $G = Q^{-1} = \begin{pmatrix} 2.1622 & 0.5405 \\ 0.5405 & 1.8018 \end{pmatrix}$ . Consider the external disturbances  $\vartheta_1(t) = \begin{cases} (6, 6)^T, & 0 \leq t \leq 3, \\ (0, 0)^T, & \text{otherwise}, \end{cases}$   $\vartheta_2(t) = \begin{cases} (7, 7)^T, & 0 \leq t \leq 3, \\ (0, 0)^T, & \text{otherwise}. \end{cases}$ . The initial condition is  $\varphi_1(0) = (-3, -7.5)^T$ ,  $\varphi_2(0) = (7, 2)^T$ ,  $\varphi_3(0) = (5, -5)^T$ ,  $\varphi_4(0) = (2, 5)^T$ ,  $\varphi_5(0) = (-1, 8)^T$ . The superposition principle is used in this section. The trajectories of  $P(t)$  are depicted in Fig. 2, which show that  $\lim_{t \rightarrow +\infty} P(t) = 0$ .

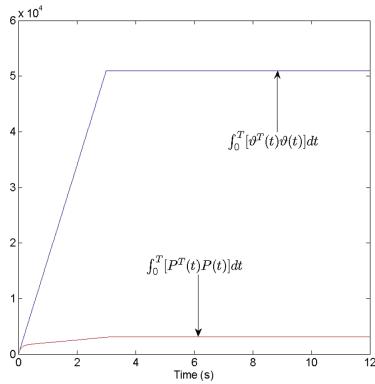
Figure 3 depicts the energy trajectories of  $P(t)$  and disturbances, which show that  $J(\vartheta) < 0$ , for given  $\gamma \in (0, 1)$ , i.e., the  $H_\infty$  performance is satisfied. The simulations results show that the FOMANs (1) achieve robust  $H_\infty$  containment control.



**Fig. 1.** The digraph.



**Fig. 2.** Trajectories of  $P(t)$  with external disturbances.



**Fig. 3.** The energy trajectories of  $P(t)$  and disturbances.

## 5 Conclusions

This paper devoted to the distributed robust  $H_\infty$  containment control problem of FOMANs with model errors and external disturbances over a digraph. By taking an equivalent transform, the original closed-loop systems were transformed into a state-space model. An LMI condition was deduced for achieving containment control under the robust  $H_\infty$  performance. And then, the convergence analysis was completed by taking the traditionally Lyapunov stability theory. Finally, the theoretical results were illustrated by simulations.

**Acknowledgement.** This work is supported by the National Natural Science Foundation of China (No. 61772063 and No. 61973329), the Beijing Natural Science Foundation (No. Z180005), and the Scientific Research Project of the Higher Education Institutions of Inner Mongolia, China (Grant No. NJZY20215).

## References

1. Mo, L., Lin, P.: Distributed consensus of second-order multiagent systems with non-convex input constraints. *Int. J. Robust Nonlinear Control* **28**, 3657–3664 (2018). <https://doi.org/10.1002/rnc.4076>
2. Mo, L., Yu, Y., Zhao, L., Cao, X.: Distributed continuous-time optimization of second-order multi-agent systems with nonconvex input constraints. *IEEE Trans. Syst. Man Cybern. Syst.* (2019). <https://doi.org/10.1109/TSMC.2019.2961421>
3. Ji, M., Ferrari-Trecate, G., Egerstedt, M., et al.: Containment control in mobile networks. *IEEE Trans. Autom. Control* **53**(8), 1972–1975 (2008). <https://doi.org/10.1109/TAC.2008.930098>
4. Cao, Y., Ren, W., Egerstedt, M.: Distributed containment control with multiple stationary or dynamic leaders in fixed and switching directed networks. *Automatica* **48**(8), 1586–1597 (2012). <https://doi.org/10.1016/j.automatica.2012.05.071>
5. Shi, G., Hong, Y., Johansson, K.: Connectivity and set tracking of multi-agent systems guided by multiple moving leaders. *IEEE Trans. Autom. Control* **57**(3), 663–676 (2012). <https://doi.org/10.1109/TAC.2011.2164733>
6. Liu, H., Xie, G., Wang, L.: Necessary and sufficient conditions for containment control of networked multi-agent systems. *Automatica* **48**, 1415–1422 (2012). <https://doi.org/10.1016/j.automatica.2012.05.010>
7. Xiong, Q., Lin, P., Chen, Z., et al.: Distributed containment control for first-order and second-order multiagent systems with arbitrarily bounded delays. *Int. J. Robust Nonlinear Control* **29**(4), 1122–1131 (2019). <https://doi.org/10.1002/rnc.4426>
8. Cao, Y., Li, Y., Ren, W., et al.: Distributed coordination of networked fractional-order systems. *IEEE Trans. Syst. Man Cybern. Part B-Cybern.* **40**(2), 362–370 (2010). <https://doi.org/10.1109/TSMCB.2009.2024647>
9. Liu, H., Xie, G., Yu, M.: Necessary and sufficient conditions for containment control of fractional-order multi-agent systems. *Neurocomputing* **323**, 86–95 (2019). <https://doi.org/10.1016/j.neucom.2018.09.067>
10. Yuan, X., Mo, L., Yu, Y.: Observer-based quasi-containment of fractional-order multi-agent systems via event-triggered strategy. *Int. J. Syst. Sci.* **50**(3), 517–533 (2019). <https://doi.org/10.1080/00207721.2018.1563222>
11. Liu, H., Xie, G., Gao, Y.: Containment control of fractional-order multi-agent systems with time-varying delays. *J. Franklin Inst.-Eng. Appl. Math.* **356**(16), 9992–10014 (2019). <https://doi.org/10.1016/j.jfranklin.2019.01.057>
12. Zames, G.: Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Trans. Autom. Control* **26**, 301–320 (1981). <https://doi.org/10.1109/tac.1981.1102603>
13. Podlubny, I.: Fractional differential equations of mathematics in science and engineering (1999). [https://doi.org/10.1007/978-3-642-39765-3\\_3](https://doi.org/10.1007/978-3-642-39765-3_3)
14. Gai, M., Cui, S., Liang, S., et al.: Frequency distributed model of Caputo derivatives and robust stability of a class of multi-variable fractional-order neural networks with uncertainties. *Neurocomputing* **202**, 91–97 (2016). <https://doi.org/10.1016/j.neucom.2016.03.043>
15. Jia, Y.: Alternative proofs for improved LMI representations for the analysis and the design of continuous-time systems with polytopic type uncertainty: a predictive approach. *IEEE Trans. Autom. Control* **48**(8), 1413–1416 (2003). <https://doi.org/10.1109/TAC.2003.815033>

16. Rockafellar, R.T.: Convex Analysis. Princeton University Press, New Jersey (1972)
17. Trigeassou, J.C., Maamri, N., Sabatier, J., et al.: A Lyapunov approach to the stability of fractional differential equations. *Sig. Process.* **91**(3), 437–445 (2011).  
<https://doi.org/10.1016/j.sigpro.2010.04.024>



# Refinement and Validation of Humoral Immunity Based on Event-B

Xuqing Shi, Shengrong Zou<sup>(✉)</sup>, Yudan Shu, and Li Chen

Institute of Information Engineering, Yangzhou University,  
Yangzhou, Jiangsu, China  
330416490@qq.com

**Abstract.** By the outbreak of Covid-19, we should focus more eyesight on the human immune system. Humoral immunity plays an important role in the immunologic mechanism. In this process, B cells and other immune cells cooperate each other to produce antibodies and eliminate antigens by series of interactions, activation, proliferation and differentiation. In this paper, we use the formal language Event-B to model the humoral immunity process on the development tool called Rodin. Humoral immunity process is abstract and has complexity in system design. Accidentally, the formal method is used to verify the correctness and consistency of the complex systems, which is an appropriate approach to model this immunity process by stepwise refinements and validation. We also present an instance to demonstrate the differences between the immunity responses after the invasion of influenza viruses and coronavirus respectively in the last refinement and validate it using proof obligations. Experimental results show that events in our model are all validated by the automatic certification tool on Rodin platform.

**Keywords:** Event-B · Refinements and validation · Coronavirus · Humoral immunity

## 1 Introduction

The correctness and security of softwares are always the important topic. Recently, the prevalence of Blockchain technology attracts many scholars and scientists from the field of Computer Science. We know its foundation is smart contract. The hackers put efforts into finding bugs in the code of the contract due to its publicity and this will cause serious accidents [1]. For example, in 2017, a bug happened in smart contract caused the loss of 50 million dollars [2]. In addition, in the field of chain design, the formal validation is equally essential to reduce the risk of large losses. There is an instance that when Intel Corporation published Pentium chains, a zero division error occurred. This code mistake costs Intel Corporation nearly 500 million dollars to retrieve Pentium chains. This accident brought huge economic and reputational losses [3].

Similar to the introduction above, with the development of the global outbreak of Covid-19, human immunology will be paid more attention. Humoral

immunity we concern about is an abstract process occurred among the interactions of diverse immune cells. We use Event-B to model it by stepwise refinements and verify it using proof obligations. Formal method can find amount of bugs and logical mistakes in this procedure. Assume that we validate the correctness of all the states in humoral immunity process, we can realize the software system using traditional programming languages such as java, python and etc. Then, this completed system can be applied for further research in medical community and used as teaching tools in biology classes.

## 2 Related Work

It's well known that when viruses or germ enter the body, the immune system in activated state will response and fight against them. As a result, the immune response which has two main lymphocytes B and T helps to clean the infected cells and the viruses. The response which B cells participate in is also called humoral immunity. For malarial infections, the humoral immunity response tends to be more effective as compared to the cellular immunity responses according to [4]. Tian et al. [5] has been inspired by humoral immunity and created a clustering model to recognize unknown antigens. From 2012 to 2017, authors in [6–9] have reported various virus dynamic models with humoral immunity or cellular immunity but due to the complexity and abstractness, these models exist drawbacks such as ignoring the absorption effect. The number of lymphocytes involved in immunity is more than  $10^{12}$  which exceeds the total number of nerve cells, and the immune system may be more complex than the nervous system. Motivated by the above studies, we consider adopting formal method to model humoral immunity and validate the correctness.

In the process of humoral immunity, antigens are divided into two groups, TD antigens and TI antigens. TD antigens require collisions and touch with phagocytic cells and helper T cells and to be absorbed and treated by them. Then B cells will receive the information from TD antigens and are activated to proliferate and differentiate plasma cells. Plasma cells will produce the related antibodies to combine with antigens and generate protein precipitation at last. Unlike TD antigens, TI antigens can directly stimulate B cells to secrete plasma cells in this process. In addition, some activated B cells can change into effector B cells which are also called memory cells during the proliferation and differentiation. When antigens with the same information as before invade the human body, they will discriminate these antigens swiftly and secrete plasma cells.

Based on the specification of humoral immunity above, we follow the refinement rules to model step by step, which means that the next model after the refinement is to add details based on the previous one. The model of each step is established by the corresponding requirements. At last, we will prove the consistency of this complex systems of immunity process, which means the logics of the model according to our requirements are reasonable.

The following are specific requirements and modeling descriptions. Each circulation in humoral immunity process can follow the same stepwise refinements.

### 3 Methodology

Event-B evolves from classical B language. It supports predicate calculus and theorem proving. It is a formal modeling method based on Action Systems. An Event-B model consists of two basic types of components, namely machine and context [10]. A context describes the static part of a model. In a context, relevant sets, constants, axioms and theorems are defined. On the contrary, a machine is used to describe the dynamic part, which includes relevant variables, invariants and events. An event has its guards and actions [11]. The guards represent the conditions under which the event will be fired. The actions mean what will happen if the event is fired. A context can be ‘seen’ by a machine, which means that the machine will use the static elements defined in the context [12]. The fundamental knowledge will be used in the following parts.

#### 3.1 The Initial Model

As shown in Table 1, we specific the requirements of the initial model. Briefly speaking, it describes the growth of B cells which is from pre-B cells to native B cells. Only native B cells can be interacted with other immune cells [13]. Then we start to model it.

Machine m0 is created which sees c0. M0 includes four events and we will introduce the main three of them. But before that, we need to define variables and the related invariants. As shown in Table 2, there are four variations. It is shown that cell which simulates a set of all the cells in the human body is contained in carrier set Cell defined in c0 and cell can partition into bcell(other cell sets will be joined in the next machines). Bcell\_hp and bcell\_state are all total surjections which denotes that one of the values of Hp and one State can correspond to multiple B cells and their domains and ranges are certainly equal to the sets revealed in the functions.

The following are events. In Fig. 1.(a), we can see the code of the event new\_bcell. Bcellobj is a formal parameter which denotes a B cell. Two guards are constrained conditions. Grd1 is the constrained condition which means that bcellobj is not the element of bcell. If grd1 is satisfied, then take actions. Act1 and act4 merge bcellobj into bcell and cell respectively. Then act2 assigns initialHp to the value of Hp of bcellobj and act3 assign prophaseState to the cell state of it. Then, we will introduce bcell\_grow\_to\_immatureState in Fig. 1.(b). In this event, the formal parameter is bcellobj. Guards constrain that bcellobj must satisfy some conditions. When satisfying these guards, take act1 to change the state into immatureState.

#### 3.2 The First Refinement Model

As shown in Table 1, we add the details of the process which describes how TD antigens and TI antigens stimulate B cells to activation respectively and it is divided into four steps. Then, we establish the model.

**Table 1.** Requirements analysis

Initial	Requirements description	First refinement	Requirements description
Req1.1	B cells are produced from the bone marrow and grow into pre-B cells	Req2.1	TD antigens are absorbed and treated by mng cells
Req1.2	Pre-B cells grow into immature B cells	Req2.2	mng cells present DNA information of TD antigens to helper T cells
Req1.3	Immature B cells grow into native B cells	Req2.3	Helper T cells present DNA information of TD antigens to native B cells and activate them
		Req2.4	TI antigens directly stimulate native B cells and activate them
Second refinement	Requirements description	third refinement	Requirements description
Req3.1	Most of the activated B cells proliferate and differentiate plasma cells	Req4.1	Memory cells discriminate antigens
Req3.2	A sub-fraction of the activated B cells proliferate and differentiate memory cells	Req4.2	Memory cells proliferate and differentiate plasma cells
Req3.3	Plasma cells produce antibodies	Req4.3	Antibodies and antigens combine into protein precipitation
Fourth refinement	Requirements description		
Req5.1	Influenza viruses invade human bodies		
Req5.2	Coronavirus invade human bodies		
Req5.3	B cells and memory cells discriminate influenza viruses and make them swallowed by neutrophile granulocyte and hyaline leukocyte		
Req5.4	T cells kill influenza viruses but kill pulmonary cells and upper respiratory tract cells at the same time		
Req5.5	Helper T cells present DNA information of TD antigens to native B cells and activate them		

**Table 2.** Variables in m0

m0	
Variables	Invariants
cell	$cell \subseteq Cell$
bcell	$partition(cell, bcell)$
bcell_hp	$bcell\_hp \in bcell \rightarrow N1$
bcell_state	$bcell\_state \in bcell \rightarrow State$

```

new_bcell: not extended ordinary >
ANY                                bcell_grow_into_immatureState: not extended ordinary >
  bcellobj      >
  WHERE
    . grd1: bcellobj#bcell not theorem >
  THEN
    . act1: bcell=bcell(bcellobj) >
    . act2: bcell_hp=bcell_hp(bcellobj)=initialHp >
    . act3: bcell_state=bcell_state(bcellobj)=prophaseState >
    . act4: cell=cell(bcellobj) >
END

```

(a)

```

bcell_grow_into_immatureState: not extended ordinary >
ANY                                bcellobj      >
  WHERE
    . grd1: bcellobj#bcell not theorem >
    . grd2: bcell_hp(bcellobj)=prophaseMaxHp not theorem >
    . grd3: bcell_state(bcellobj)=prophaseState not theorem >
  THEN
    . act1: bcell_state(bcellobj)=immatureState >
END

```

(b)

**Fig. 1.** The events defined in m0. (a) Describe the production of a B cell. (b) Describe that a B cell grows into immature state.

We establish m1 which sees c1 and refines m0. There are six events in m1 and we will concretely state three of them. As shown in Table 3, we define related variables and invariants. Cell, ag and aginfo are subsets of their parallel carrier sets in c1 and they have their corresponding elements which have slave relations with them according to the associated invariants. Td\_ag\_td\_aginfo and tiag\_tiaaginfo are both total surjections which denote that each TD antigen and TI antigen have their linked DNA information so domains are td\_ag and tiag and ranges are td\_aginfo and tiaginfo respectively. Mngcell\_td\_aginfo, helperTcell\_td\_aginfo, bcell\_td\_aginfo and bcell\_tiaaginfo are partial functions because their domains are changeable that cells are constantly produced and dead in the human body and not each phagocyte cell, helper T cell or bcell has opportunity to carry DNA information of antigens.

The following are events. The event new\_td\_ag has two formal parameters, td\_agobj and td\_aginfoobj. Grd1 and grd2 respectively constrain that they are not the elements of the linked sets. When grd1 and grd2 are contented, then take actions. Act1 and act2 merge them into the related sets. Act3 indicate that the antigen DNA information which td\_agobj carries is assigned to td\_aginfoobj. In Event-B method, function is not only relation but also set [14]. The event next event has three formal parameters, td\_agobj, mngcellobj and td\_aginfoobj. Grd1 and grd2 regulate their slave relations. Grd3 denotes that td\_aginfoobj is the value of td\_agobj under he map function. Grd4 restricts that mngcellobj is not in the domain of mngcell\_td\_aginfo because before this event is carried out, mngcellobj doesn't take any information. When act1 are permitted to take, mngcellobj will carry td\_aginfoobj.

**Table 3.** Variables in m1

m1	
Variables	Invariants
mngcell	$\text{partition}(\text{cell}, \text{bcell}, \text{mngcell}, \text{helperTcell})$
helperTcell	$\text{cell} \subseteq \text{Cell}$
aginfo	$\text{partition}(\text{aginfo}, \text{td\_aginfo}, \text{tiaginfo})$
td_aginfo	$\text{aginfo} \subseteq \text{AgInfo}$
ti_aginfo	$\text{partition}(\text{ag}, \text{td\_ag}, \text{tiag})$
ag	$\text{ag} \subseteq \text{Ag}$
td_ag	$\text{td\_ag} \cup \text{td\_aginfo} \in \text{td\_ag} \Rightarrow \text{td\_aginfo}$
ti_ag	$\text{tiag} \cup \text{tiaginfo} \in \text{tiag} \Rightarrow \text{tiaginfo}$
td_ag_td_aginfo	$\text{mngcell} \cup \text{td\_aginfo} \in \text{mngcell} \rightsquigarrow \text{td\_aginfo}$
ti_ag_ti_aginfo	$\text{helperTcell} \cup \text{td\_aginfo} \in \text{helperTcell} \rightsquigarrow \text{td\_aginfo}$
mngcell_td_aginfo	$\text{bcell} \cup \text{td\_aginfo} \in \text{bcell} \rightsquigarrow \text{td\_aginfo}$
helperTcell_td_aginfo	$\text{bcell} \cup \text{tiaginfo} \in \text{bcell} \rightsquigarrow \text{tiaginfo}$
bcell_td_aginfo	
bcell_ti_aginfo	

### 3.3 The Second Refinement Model

As shown in Table 1, the details of the process which describes how the activated B cells change into plasma cells and produce antibodies are added into the previous model. In addition, the appearance of memory cells will work soon. The second refinement requirements are divided into three steps. After stating the requirements, we will complete modeling.

We describe the dynamic machine m2. Similar as before, in Table 4, we define the variables plasmacell, memorycell, ab, td\_ab and ti\_ab, which imply the sets of these kinds of cells in the actual human body and the relevant invariants regulate the slave relations among them. Plasmacell\_td\_aginfo, plasmacell\_ti\_aginfo and etc. are all partial surjections because the entire cells in them are directly or indirectly produced by the B cells which have DNA information of antigens and they will also carry it. Thus, the domains are equal to their correlated cell sets. But doubtfully, B cells will not take all kinds of antigen information so the range is included by td\_aginfo or ti\_aginfo.

Later, we will state the two main events. The first event specifies the process that B cells with DNA information of TD antigens proliferate and differentiate. B cells will mostly change into plasma cells, and a little portion of them will translate into memory cells. To simulate this process, we set seven formal parameters in the event as before. Guards constrain that td\_aginfoobj is carried by bcellobj, and plasma cells and memorycellobj are not produced before taking actions. Hence, they are not the elements of the relevant cell sets. When satisfying guards, bcellobj disappears to change into plasma cells and

**Table 4.** Variables in m2

m2	
Variables	Invariants
Plasmacell	$\text{partition}(\text{cell}, \text{bcell}, \text{mngcell}, \text{helperTcell}, \text{plasmacell}, \text{memorycell})$
Memorycell	$\text{cell} \subseteq \text{Cell}$
ab	$\text{partition}(\text{ab}, \text{td\_ab}, \text{tiab})$
td_ab	$\text{ab} \subseteq \text{Ab}$
ti_ab	$\text{partition}(\text{ag}, \text{td\_ag}, \text{tiag})$
plasmacell_td_aginfo	$\text{plasmacell\_td\_aginfo} \in \text{plasmacell} \hookrightarrow \text{td\_aginfo}$
plasmacell_ti_aginfo	$\text{plasmacell\_ti\_aginfo} \in \text{plasmacell} \hookrightarrow \text{ti\_aginfo}$
memorycell_td_aginfo	$\text{memorycell\_td\_aginfo} \in \text{memorycell} \hookrightarrow \text{td\_aginfo}$
memorycell_ti_aginfo	$\text{td\_ab\_td\_aginfo} \in \text{td\_ab} \hookrightarrow \text{td\_aginfo}$
td_ab_td_aginfo	$\text{ti\_ab\_ti\_aginfo} \in \text{ti\_ab} \hookrightarrow \text{ti\_aginfo}$
ti_ab_ti_aginfo	$\text{memorycell\_ti\_aginfo} \in \text{memorycell} \hookrightarrow \text{ti\_aginfo}$

memorycellobj. In addition, the new produced cells carry the antigen information. The second event new\_td\_ab imitates the process of producing antibodies whose goal is to eliminate TD antigens. Formal parameters are td\_abobj, plasmacellobj and td\_aginfoobj. Guards regulate them. When guards are contented, td\_abobj is truely produced so it will join the set and it will carry the DNA information of the related antigens.

### 3.4 The Third Refinement Model

As shown in Table 1, this refinement consists of two procedures. One is that memory cells discriminate the antigens which the ones of the same type invaded before and they secret plasma cells to produce antibodies. The other is that antibodies combine with antigens and induce immunological effects. Subsequently, we start to build up the model.

In this refinement, we didn't append novel elements in the static context so we just need to create dynamic machine m3 which sees c2. In Table 5, we define the new entrants. Memorycell\_td\_ag and memorycell\_ti\_ag are both total functions which imply the interactions between the memory cells and antigens. Then, we introduce the events.

**Table 5.** Variables in m3

VARIABLES	INVARIANTS
memorycell_td_ag	$\text{memorycell\_td\_ag} \in \text{memorycell} \longrightarrow \text{td\_ag}$
memorycell_ti_ag	$\text{memorycell\_ti\_ag} \in \text{memorycell} \longrightarrow \text{ti\_ag}$

The first event specifies how memory cells indentify TD antigens. When guards are all satisfied, act1 will be taken that 'memorycellobj $\mapsto$ td\_agobj' will be

integrated into memorycell\_td\_ag. After that, We describe the immune responses during the process of combining antibodies with TD antigens. In this event, grd1 to grd4 make sure that they both belong to the related sets and possess antigen information. Grd5 to grd8 regulate that td\_abobj and td\_agobj carry the same type of DNA information of antigens. After fulfilling the guards, we take ac1 and act2 to extract td\_abobj and td\_agobj from the related sets because immune responses happen between them and they are combined and changed into precipitate.

### 3.5 The Fourth Refinement Model

As shown in Table 1, in the last refinement, for further research on Covid-19, we present an instance to specify the differences of the immune responses between normal influenza viruses and coronavirus when they invade human bodies [15–17]. This refinement includes five steps in the table. Then, we introduce our designed model.

**Table 6.** Variables in m4

m4	
Variables	<i>Invariants</i>
virus	$cell \subseteq Cell$
in_virus	$partition(cell, bcell, mngcell, helperTcell, plasmacell, memorycell, tcell, upp_cell, pul_cell)$
co_virus	$virus \subseteq Virus$
virusinfo	$partition(virus, in\_virus, co\_virus)$
in_virusinfo	$virusinfo \subseteq VirusInfo$
co_virusinfo	$partition(virusinfo, in\_virusinfo, co\_virusinfo, harmless\_cellinfo)$
harmless_cellinfo	$in\_virus\_in\_virusinfo \in in\_virus \Rightarrow in\_virusinfo$
in_virus_in_virusinfo	$co\_virus\_co\_virusinfo \in co\_virus \Rightarrow co\_virusinfo$
co_virus_co_virusinfo	$bcell\_in\_virusinfo \in bcell \rightsquigarrow in\_virusinfo$
bcell_in_virusinfo	$memorycell\_in\_virusinfo \in memorycell \rightsquigarrow in\_virusinfo$
memorycell_in_virusinfo	$bcell\_harmless\_cellinfo \in bcell \rightsquigarrow harmless\_cellinfo$
bcell_harmless_cellinfo	$memorycell\_harmless\_cellinfo \in memorycell \rightsquigarrow harmless\_cellinfo$
memorycell_harmless_cellinfo	$upp\_cell\_state \in upp\_cell \hookleftarrow State$
upp_cell_state	$pul\_cell\_state \in pul\_cell \hookleftarrow State$
pul_cell_state	

Figure 2 consists of five panels (a) through (e), each showing a table of event statistics. The tables have columns for Element Name, Total, Auto, Man., Rev., and Und.

- (a) m0:**

Element Name	Total	Auto	Man.	Rev.	Und.
Proof Oblig...	16	10	3	3	0
INITIALSATI...	3	3	0	0	0
inv1	0	0	0	0	0
inv2	2	2	0	0	0
inv3	3	1	0	2	0
inv4	4	1	3	0	0
new_bcell	3	1	1	1	0
bcell_grow	4	2	0	2	0
bcell_grow_L...	3	2	1	0	0
bcell_grow_L...	3	2	1	0	0
new_td_ag	7	7	0	0	0
- (b) m1:**

Element Name	Total	Auto	Man.	Rev.	Und.
m1	41	41	0	0	0
INITIALSATI...	13	13	0	0	0
inv1	4	4	0	0	0
inv2	2	2	0	0	0
inv3	2	2	0	0	0
inv4	0	0	0	0	0
inv5	0	0	0	0	0
inv6	0	0	0	0	0
inv7	0	0	0	0	0
inv8	0	0	0	0	0
inv9	3	3	0	0	0
inv10	3	3	0	0	0
inv11	4	4	0	0	0
inv12	4	4	0	0	0
collision_td_a...	2	2	0	0	0
collision_mng...	2	2	0	0	0
inv13	3	3	0	0	0
inv14	3	3	0	0	0
inv15	3	3	0	0	0
inv16	3	3	0	0	0
inv17	7	7	0	0	0
inv18	3	3	0	0	0
inv19	3	3	0	0	0
inv20	2	2	0	0	0
new_td_ab	3	3	0	0	0
new_td_ab	3	3	0	0	0
new_td_ab	1	1	0	0	0
- (c) m2:**

Element Name	Total	Auto	Man.	Rev.	Und.
m2	24	24	0	0	0
INITIALSATI...	14	14	0	0	0
inv1	3	3	0	0	0
inv2	0	0	0	0	0
bcellwithtd...	7	7	0	0	0
inv3	3	3	0	0	0
inv4	3	3	0	0	0
inv5	3	3	0	0	0
inv6	3	3	0	0	0
inv7	7	7	0	0	0
inv8	3	3	0	0	0
inv9	2	2	0	0	0
inv10	2	2	0	0	0
new_td_ab	3	3	0	0	0
new_td_ab	3	3	0	0	0
new_td_ab	1	1	0	0	0
- (d) m3:**

Element Name	Total	Auto	Man.	Rev.	Und.
m3	28	28	0	0	0
INITIALSATI...	1	1	0	0	0
inv1	1	0	0	0	0
inv2	5	5	0	0	0
memorycell_i...	3	3	0	0	0
memorycell_i...	4	4	0	0	0
memorycell_i...	5	5	0	0	0
memorycell_i...	5	5	0	0	0
td_ab_combi...	4	4	0	0	0
td_ab_combi...	5	5	0	0	0
- (e) m4:**

Element Name	Total	Auto	Man.	Rev.	Und.
m4	41	41	0	0	0
INITIALSATI...	11	11	0	0	0
inv1	1	0	0	0	0
inv2	0	0	0	0	0
inv3	0	0	0	0	0
inv4	5	5	0	0	0
new_in_virus	5	5	0	0	0
new_in_virus	3	3	0	0	0
inv5	0	0	0	0	0
inv6	3	3	0	0	0
inv7	4	4	0	0	0
inv8	2	2	0	0	0
inv9	2	2	0	0	0
inv10	2	2	0	0	0
inv11	3	3	0	0	0
inv12	4	4	0	0	0
inv13	2	2	0	0	0
inv14	2	2	0	0	0
inv15	4	4	0	0	0
inv16	4	4	0	0	0
inv17	1	1	0	0	0
inv18	1	1	0	0	0
inv19	6	6	0	0	0
inv20	5	5	0	0	0

**Fig. 2.** The experimental results for proof. (a) Show the results for verifying events in m0. (b) Show the results for verifying events in m1. (c) Show the results for verifying events in m2. (d) Show the results for verifying events in m3. (e) Show the results for verifying events in m4.

We build up the dynamic machine m4. As usual, we specify the variables and their invariants in Table 6. Virus, virusinfo and their internal elements are defined as same as ones in c4. In\_virus\_in\_virusinfo and co\_virus\_co\_virusinfo are both total surjections because each virus has DNA or RNA information and each kind of virus information can correspond to multiple viruses. Bcell\_in\_virusinfo, memorycell\_in\_virusinfo, bcell.harmless\_cellinfo and memorycell.harmless\_cellinfo are all partial functions because not each bcell or memory cell got in touch with viruses or harmless cells and carry their information. For that, domains and ranges of these functions are included by the relevant sets. Upp\_cell\_state and pul\_cell\_state are both partial surjections because each upper respiratory tract cell or pulmonary cell has a cell state(alive or dead), so their domains are certain.

After elaborating the variables and invariants, we can characterize the events. The event new\_in\_virus has two parameters and guards constrain that they are not the elements of the related sets before taking actions. When guards are satisfied, the two parameters will be combined into the sets. Finally, we introduce the event which describes how coronavirus avoid the examination from B cells or memory cells, and then cause fatal damage to the upper respiratory tracts cell, pulmonary cells and other useful cells which are playing important roles. This event has four parameters, which are bcellobj, co\_virusobj, upp\_cellobj and pul\_cellobj. Grd1 and grd2 represent that bcellobj is in activeState so it has

abilities to receive virus information and indirectly produce antibodies to fight against viruses. Grd5 to grd8 constrain that upp\_cellobj and pul\_cellobj are regularly working in the human body. The most important guard is grd4. This guard imitates that when bcell collides with co\_virusobj, it receives its information but this information is wrong which is mocked by co\_virusobj. Consequently, it can be reckoned as DNA information of the harmless cells and bcellobj will not handle it. When these guards are passed, co\_virusobj will do harm to upp\_cellobj and pul\_cellobj. Therefore, act1 and act2 assign deadState to them to show the consequences.

## 4 Experiments

Rodin is a complex tool platform which supports the application of the Event-B formal method. It provides core functionality for syntactic analysis and proof-based verification of Event-B models. The Rodin Platform is an Eclipse-based IDE for Event-B that provides effective support for refinement and mathematical proof [18]. We utilize the automatic proof plugin on Rodin to validate our established model and the events which cannot be automatically verified are proved manually. In Fig. 2, we display the results. In Fig. 2.(a), there are a total of 16 proof obligations in the initial model and 10 obligations of them are automatically verified. For the left 6 ones, we add some constrained conditions and hypotheses to make the obligations pass through the validation. In Fig. 2.(b), there are a total of 41 proof obligations in the first refinement model and they are all validated automatically. In Fig. 2.(c), there are a total of 34 proof obligations in the second refinement model and they are all validated automatically. In Fig. 2.(b), there are a total of 28 proof obligations in the third refinement model and they are all validated automatically. In Fig. 2.(b), there are a total of 41 proof obligations in the last refinement model and they are all verified by the automatic plugin.

## 5 Conclusion

In this paper, we extract requirements from the abstract process of humoral immunity and show an instance which reflects on the differences between the methods of the attacks which aim at harming the human bodies from influenza viruses and coronavirus. This instance is our try to hope to help push the research on Covid-19. Then, based on the requirements, we define models by stepwise refinements and prove its correctness and consistency. However, we know this is just a portion of the enormous and complex immune system. Thus, we will do further research and develop the superb system. In the future, we also want to use real data to validate our Event-B model to increase its rationality.

## References

1. Singla, V., Malav, I.K., Kaur, J., Kalra, S.: Develop leave application using blockchain smart contract. In: 2019 11th International Conference on Communication Systems Networks (COMSNETS), pp. 547–549 (2019)
2. Dirgantoro, K.P., Lee, J.M., Kim, D.: Generative adversarial networks based on edge computing with blockchain architecture for security system. In: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, pp. 039–042 (2020)
3. Sulskus, G., Poppleton, M., Rezazadeh, A.: An interval-based approach to modelling time in Event-B. In: Fundamentals of Software Engineering, vol. 9392, pp. 292–307 (2015)
4. Deans, J., Cohen, S.: Immunology of malaria. *Ann. Rev. Microbiol.* **37**, 25–50 (1983)
5. Tian, Y., Ren, P.: A clustering model inspired by humoral immunity. In: 2009 International Workshop on Intelligent Systems and Applications, Wuhan, pp. 1–4 (2009). <https://doi.org/10.1109/IWISA.2009.5072611>
6. Pradeep, B.G.S.A., Ali, H.: Global stability properties for a delayed virus dynamics model with humoral immunity response and absorption effect. In: 2017 International Conference on Electrical Engineering (ICEE), Lahore, pp. 1–6 (2017)
7. Wang, S.F., Zou, D.Y.: Global stability of in-host viral models with humoral immunity and intracellular delays. *Appl. Math. Model.* **36**, 1313–1322 (2012)
8. Wang, T., Hu, Z., Liao, F.: Stability and Hopf bifurcation for a virus infection model with delayed humoral immunity response. *J. Math. Anal. Appl.* **411**, 63–74 (2014)
9. Elaiw, A.M.: Global stability analysis of humoral immunity virus dynamics model including latently infected cells. *J. Biol. Dyn.* **9**(1), 215–228 (2015)
10. Siyuan, H., Hong, Z.: Towards transformation from UML to event-B. In: 2015 IEEE International Conference on Software Quality, Reliability and Security - Companion, Vancouver, BC, pp. 188–189 (2015)
11. El Mimouni, S., Bouhdadi, M.: Formal modeling of the simple text oriented messaging protocol using Event-B method. In: IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech 2015, pp. 1–4 (2015)
12. Abrial, J.R.: Modeling in Event-B: System and Software Engineering. Cambridge University Press, Cambridge (2010)
13. Ningthoujam, S., Chingkheinganba, T., Chakraborty, S.K.: Finding an effective distance between T-cell and B-cell using S/W ARQ in an immune system communication. *China Commun.* **17**(1), 174–185 (2020)
14. Preduț, S., Ipate, F., Gheorghe, M., Campean, F.: Formal modelling of cruise control system using Event-B and Rodin platform. In: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmarCity/DSS), Exeter, United Kingdom, pp. 1541–1546 (2018)
15. Wu, L., Horng, J., Huang, H., Chen, W.: Identifying discriminative amino acids within the hemagglutinin of human influenza A H5N1 virus using a decision tree. *IEEE Trans. Inf. Technol. Biomed.* **12**(6), 689–695 (2008). <https://doi.org/10.1109/TITB.2008.896871>
16. Dong, D., et al.: The role of imaging in the detection and management of COVID-19: a review. *IEEE Rev. Biomed. Eng.* (2020). <https://doi.org/10.1109/RBME.2020.2990959>

17. Pavesi, A., Tan, A.T., Chen, M.B., Adriani, G., Bertoletti, A., Kamm, R.D.: Using microfluidics to investigate tumor cell extravasation and T-cell immunotherapies. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, pp. 1853–1856 (2015). <https://doi.org/10.1109/EMBC.2015.7318742>
18. Hoang, T.S., Fürst, A., Abrial, J.: Event-B patterns and their tool support. In: 2009 Seventh IEEE International Conference on Software Engineering and Formal Methods, Hanoi, pp. 210–219 (2009). <https://doi.org/10.1109/SEFM.2009.17>



# Accelerated Distributed Algorithm for Solving Linear Algebraic Equations

Weikang Hu and Aiguo Wu<sup>(✉)</sup>

Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China  
[ag.wu@163.com](mailto:ag.wu@163.com)

**Abstract.** A novel distributed algorithm is proposed in this paper for solving any linear algebraic equations with a unique solution. Some convergence results have been obtained for this algorithm by investigating the spectral radius of an iterative matrix. The new algorithm is different from the previous results in nature since the information of current estimations as well as previous estimations are utilized in the latest update step. Both the theoretical analysis and the numerical experiment show that the proposed algorithm has a faster convergence rate if the relaxation parameter is chosen carefully.

**Keywords:** Distributed algorithms · Linear equations · Convergence rate

## 1 Introduction

It is known that many problems naturally arising from engineering application can be described mathematically through a system of linear equations, i.e.,  $Ax = b$  by notation. For this reason, efforts to develop numerical algorithms for solving  $Ax = b$  have been under way for a long time. However, due to the exploding requirement of computational resources, in the case with a large number of unknown variables, it is quite difficult to solve  $Ax = b$  via classical centralized algorithms such as Jacobi iteration, Gauss-Seidel iteration or Kaczmarz method. As a viable alternative of solving large-scale  $Ax = b$ , distributed algorithm has three central features: i) It has the ability to fully utilize the distributed nature of physical systems with less communication pressure [11]; ii) The heavy computational burden can be split over robust multi-agent networks [3]; iii) Only

---

This work was supported in part by the National Natural Science Foundation of China for Excellent Young Scholars under Grant No. 61822305; The Fundamental Research Funds for the Central Universities under Grant No. HIT.BRETIV.201907; Guangdong Natural Science Foundation under Grant Numbers 2020A1515011091 and 2019A1515011576; Shenzhen Municipal Project for International Cooperation with Project No. GJHZ20180420180849805; Shenzhen Municipal Basic Research Project for Discipline Layout with Project Numbers JCYJ20180507183437860 and JCYJ20170811160715620.

limited information needs to be shared among subsystems which is useful for privacy preserving purpose [7].

During the past several years, many effective distributed algorithms have been proposed to solve  $Ax = b$  due to the aforementioned advantages [14]. In [9], a projection-based distributed algorithm was described for undirected connected graph which restricts the updating term on the kernel of  $A$ . A major drawback of this algorithm is that the initial value of each agents must lie in the solution space of local subsystems. Later, the result in [9] was extended to repeatedly jointly strongly connected graphs [8]. By introducing the notion of  $D$ -connected, the non-redundant assumption in [8] was further relaxed by [5]. Apart from the extension of communication topology, the adding term proposed in [15] eliminates the required initialization step of [8]. In fact, the algorithm derived in [15] can be seen as a special version of the *projected consensus algorithm* proposed in [10] which solves general constrained consensus problem. The algorithm in [15] was further generalized in [13] by replacing the orthogonal projection matrix with a class of matrices possessing special spectrum property. By listing all event times in chronological order and introducing the so-called extended neighbor graph, the asynchronous cases of two algorithms developed in [8] and [15] were investigated, respectively [6]. Another modification of [9] was proposed in [2], which intended to speed up convergence without the proof of convergence rate, but with convincing experimental results. Two kinds of distributed algorithms with tunable parameters were designed in [12] and proved to converge to the solution of  $Ax = b$  if two parameters are chosen carefully.

Note that a common feature of all the previously mentioned distributed algorithms is that in each iterative step the latest estimation  $x(t+1)$  of individual agent is updated by only utilizing the information in the current step  $x(t)$ . However, the information in the last step  $x(t-1)$  is not exploited. In this paper, by taking advantage of the existing results developed in [15] and exploiting the previous estimation, we present a novel distributed algorithm with an over-relaxation parameter. In fact, the over-relaxation technique has been used in many iteration algorithms to speed up convergence. Recently, in [16] a successive over-relaxation based algorithm was proposed for solving coupled Lyapunov matrix equation. The algorithm presented here has the ability to solve any  $Ax = b$  with a unique solution for arbitrary initial condition. With the tunable parameter being appropriately chosen, a faster convergence rate can be achieved. The convergence property with different tunable parameters is obtained by investigating the spectral radius of an augmented iterative matrix.

The remainder of the paper is organized as follows: In Sect. 2, we formulate some preliminaries and propose a novel distributed algorithm. In Sect. 3, for our new algorithm, we give concrete analysis of its convergence as well as convergence rate. In Sect. 4, the performance with different tunable parameters is demonstrated by a numerical example. Finally, concluding remarks are given in Sect. 5.

Throughout this paper, for a matrix  $A \in \mathbb{R}^{n \times n}$ ,  $A^T$  and  $\rho(A)$  denote its transpose and spectral radius, respectively. The notation  $\otimes$  represents the Kronecker

product of two matrices. For two integers  $a \leq b$ , the notation  $\mathbb{I}[a, b]$  is defined as  $\mathbb{I}[a, b] = \{a, a+1, \dots, b\}$ .  $\text{col}\{\dots\}$  denotes a column stack of all elements in it.  $\text{diag}\{\dots\}$  refers to a block diagonal matrix. Moreover,  $I$  and  $0$  represent identity matrix and zero matrix, respectively, whose dimensions are compatible with the context.  $|.|$  and  $\|.\|$  refer to the absolute value and the Euclidean norm, respectively.

## 2 Preliminary and Algorithm

Consider a system of linear equations  $Ax = b$ , which can be partitioned as

$$A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix}_{n \times n}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}_{n \times 1}$$

where  $A_i \in \mathbb{R}^{m_i \times n}$ ,  $b_i \in \mathbb{R}^{m_i \times 1}$  and  $\sum_{i=1}^m m_i = n$ .

The problem of interest here is to solve  $Ax = b$  using a network of  $m > 1$  intelligent agents which are able to exchange information with certain other agents called its “neighbors”. We denote  $\mathcal{N}_i$  for the set of agent  $i$ ’s neighbor and we consider agent  $i$  as a neighbor of itself. Let  $d_i$  denote the cardinality of  $\mathcal{N}_i$ . Neighbor relations here are described by a undirected graph  $\mathcal{G}$  with  $m$  vertices and a set of undirected arcs. There is an arc  $(i, j)$  in  $\mathcal{G}$  just in case that agent  $i$  and  $j$  are neighbors. Each agent  $i$  updates a state vector  $x_i(t)$  taking values in  $\mathbb{R}^n$ , and we denote the information agent  $i$  receives from neighbor  $j$  at time  $t$  as  $x_j(t)$ . We suppose that each agent  $i$  only owns a part of the partitioned matrix  $[A \ b]$ , i.e.,  $[A_i \ b_i]$ .

As to the solutions for  $Ax = b$ , many distributed algorithms have been proposed in the existing literature. Here we are interested in accelerating the so-called *projected consensus algorithm* developed in [6, 10, 13, 15] for it has the following benefits: i) This algorithm has the ability to solve any solvable  $Ax = b$  with at least one solution exponentially fast for arbitrary initial condition. ii) Communications among agents require  $n$  dimensional estimation of each agent and nothing more. iii) It has the potential to converge for any repeatedly jointly strongly connected graph in an asynchronous manner. Note that we only focus on the synchronous case with connected communication topology which at once leads to the following lemma.

**Lemma 1.** [15]: *If the linear algebraic equation  $Ax = b$  has a unique solution  $x^*$  and the communication topology  $\mathcal{G}$  is connected, then the solution of  $Ax = b$  can be obtained by the following iterative algorithm for any initial value:*

$$x_i(t+1) = \bar{x}_i(t) - A_i^T (A_i A_i^T)^{-1} (A_i \bar{x}_i(t) - b_i), \quad i \in \mathbb{I}[1, m] \quad (1)$$

where

$$\bar{x}_i(t) = \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} x_j(t)$$

That is the iteration in (1) satisfies  $\lim_{t \rightarrow \infty} x_i(t) = x^*, i \in \mathbb{I}[1, m]$ .

Observe that in the above algorithm, the estimates  $x_i(t+1)$  for  $x^*, i \in \mathbb{I}[1, m]$  are updated by only using the estimates  $x_i(t)$  at the  $t$ -th step. In fact, in each iteration step, the previous estimates  $x_i(t-1), i \in \mathbb{I}[1, m]$  can be utilized. Inspired by the idea in [4] which studied an accelerated iterative algorithm for distributed averaging, we aim to renovate algorithm (1) with the information of current estimations as well as previous estimations. Doing this leads at once to the following algorithm

$$x_i(t+1) = \hat{x}_i(t) - A_i^T (A_i A_i^T)^{-1} (A_i \hat{x}_i(t) - b_i), \quad i \in \mathbb{I}[1, m] \quad (2)$$

where

$$\hat{x}_i(t) = \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} [(1 + \alpha)x_j(t) - \alpha x_i(t-1)]$$

In the next section, we will give concrete analysis on the convergence and convergence rate of algorithm (2) by investigating the spectral radius of an augmented iterative matrix. Before moving on, we present some lemmas derived from the existing literature which will be used in proving our main results.

**Lemma 2.** [1]: Given any square real matrix  $B$ ,  $\lim_{t \rightarrow \infty} B^t = 0$  if and only if  $\rho(B) < 1$ .

**Lemma 3.** [1]: Let  $X$  be any  $m \times n$  matrix and  $Y$  be any  $n \times m$  matrix. Then the non-zero eigenvalues of  $XY$  are the same as the non-zero eigenvalues of  $YX$ .

**Lemma 4.** [4]: Let  $B$  be any  $mn \times mn$  real matrix and  $\mathcal{B}$  be a  $2mn \times 2mn$  matrix given by

$$\mathcal{B} = \begin{bmatrix} (\alpha + 1)B & -\alpha I \\ I & 0 \end{bmatrix}$$

where  $\alpha$  is a real constant. For any  $\alpha$  and each eigenvalue  $\mu_i$  of  $B$ , there are two corresponding eigenvalues  $(\lambda_{i1}, \lambda_{i2})$  of  $\mathcal{B}$  which are two roots of the following equation:

$$\lambda^2 - \mu(\alpha + 1)\lambda + \alpha = 0 \quad (3)$$

### 3 Analysis

For the ease of analysis, we transfer algorithms (1) and (2) into state space form and make a change of variable by denoting  $e_i(t) = x_i(t) - x^*, i \in \mathbb{I}[1, m]$ . From (1), let  $Q_i = I - A_i^T (A_i A_i^T)^{-1} A_i$ , one has

$$e_i(t+1) = \frac{1}{d_i} Q_i \sum_{j \in \mathcal{N}_i} e_j(t), \quad i \in \mathbb{I}[1, m]$$

Toward this end, we denote  $e(t) = \text{col}\{e_1(t), \dots, e_m(t)\}$  and  $\mathcal{L} = \mathcal{D}^{-1}\mathcal{A}^T$  where  $\mathcal{A}$  is the adjacency matrix of  $\mathcal{G}$  and  $\mathcal{D} = \text{diag}\{d_1, \dots, d_m\}$ . With the above notations, we write algorithm (1) in a state space form

$$e(t+1) = Be(t) \quad (4)$$

where  $B = Q(\mathcal{L} \otimes I)$  and  $Q = \text{diag}\{Q_1, Q_2, \dots, Q_m\}$ .

In a similar manner, algorithm (2) can be rewritten as

$$z(t+1) = \mathcal{B}z(t) \quad (5)$$

where  $z(t) = \text{col}\{e(t), e(t-1)\}$  and  $\mathcal{B}$  is defined in Lemma 3.

It is clear that the iterative Eqs. (4) and (5) converge to 0 for any initial value if and only if  $\rho(\mathcal{B}) < 1$  by Lemma 2. Next, we will investigate the spectral radius of  $\mathcal{B}$  via characteristic Eq. (3). Since the  $2mn$  eigenvalues  $(\lambda_{i1}, \lambda_{i2}, i \in \mathbb{I}[1, mn])$  of  $\mathcal{B}$  are determined by the  $mn$  eigenvalues  $(\mu_i, i \in \mathbb{I}[1, mn])$  of  $B$  from Lemma 3, we need the following proposition which reveals the special property of  $\mu_i$ .

**Proposition 1.** *If the communication topology  $\mathcal{G}$  is connected and  $Ax = b$  has a unique solution, then the eigenvalues of  $B$  are all real and satisfy  $|\mu_i| < 1, i \in \mathbb{I}[1, mn]$ .*

*Proof.* By Lemma 1, for any initial value,  $\lim_{t \rightarrow \infty} e_i(t) = \lim_{t \rightarrow \infty} (x_i(t) - x^*) = 0$ ,  $i \in \mathbb{I}[1, m]$  and thus  $\lim_{t \rightarrow \infty} e(t) = 0$ . As the evolvement of  $e(t)$  is governed by the iterative Eq. (4), it must be true that  $\rho(B) < 1$  by Lemma 2. Note that for any connected topology  $\mathcal{G}$ ,  $\mathcal{L} \otimes I$  is symmetric. Since  $Q$  is a projection matrix, one has  $B = Q(\mathcal{L} \otimes I) = QQ(\mathcal{L} \otimes I)$ . It is obvious that the non-zero eigenvalues of  $B$  are the same as the non-zero eigenvalues of  $Q(\mathcal{L} \otimes I)Q$  by Lemma 3. Since  $Q(\mathcal{L} \otimes I)Q$  is clearly symmetric,  $\mu_i$  are real for any  $i \in \mathbb{I}[1, mn]$ .

In the sequel, we suppose that  $|\mu_1| \geq |\mu_2| \geq \dots \geq |\mu_{mn}|$  and thus  $\rho(B) = |\mu_1|$ . The main results of this paper are summarized in the following theorem.

**Theorem 1.** *If the linear algebraic equation  $Ax = b$  has a unique solution  $x^*$  and the communication topology  $\mathcal{G}$  is connected, then the solution of  $Ax = b$  can be obtained by algorithm (2) for any initial value*

1. if and only if  $-1 < \alpha < 1$
2. as fast as algorithm (1) if  $\alpha = 0$  or  $\alpha = \mu_1^2$
3. slower than algorithm (1) if  $-1 < \alpha < 0$  or  $\mu_1^2 < \alpha < 1$
4. faster than algorithm (1) if  $0 < \alpha < \mu_1^2$
5. fastest when  $\alpha = \frac{1-\sqrt{1-\mu_1^2}}{1+\sqrt{1-\mu_1^2}}$

With the help of Proposition 1, Theorem 1 is a direct subsequence of the following two lemmas.

**Lemma 5.** If the communication topology  $\mathcal{G}$  is connected and  $Ax = b$  has a unique solution, then  $\rho(\mathcal{B})$  can be expressed by the following function

$$\rho(\mathcal{B}) = \begin{cases} \frac{1}{2}|\mu_1(\alpha + 1)| + \frac{1}{2}\sqrt{\mu_1^2(\alpha + 1)^2 - 4\alpha}, & \alpha \leq \alpha_{11} \\ \sqrt{\alpha}, & \alpha_{11} < \alpha < \alpha_{12} \\ \frac{1}{2}|\mu_1(\alpha + 1)| + \frac{1}{2}\sqrt{\mu_1^2(\alpha + 1)^2 - 4\alpha}, & \alpha \geq \alpha_{12} \end{cases} \quad (6)$$

where

$$\alpha_{11} = \frac{1 - \sqrt{1 - \mu_1^2}}{1 + \sqrt{1 - \mu_1^2}}, \quad \alpha_{12} = \frac{1 + \sqrt{1 - \mu_1^2}}{1 - \sqrt{1 - \mu_1^2}} \quad (7)$$

*Proof.* If for some  $\mu_i$  and  $\alpha$ , the roots of Eq. (3) are complex, then

$$|\lambda_{i1}| = |\lambda_{i2}| = \sqrt{\alpha} \quad (8)$$

Or if for some  $\mu_i$  and  $\alpha$ , the roots of Eq. (3) are real, then

$$\lambda_{i1,2} = \frac{1}{2}\mu_i(\alpha + 1) \pm \frac{1}{2}\sqrt{\mu_i^2(\alpha + 1)^2 - 4\alpha} \quad (9)$$

Since the convergence and convergence rate of algorithm (2) are only determined by the certain eigenvalue with maximum magnitude, i.e., the spectral radius of  $\mathcal{B}$ , we are interested in  $\rho_i = \max\{|\lambda_{i1}|, |\lambda_{i2}|\}$  for real roots case. From (9), we can get that

$$\rho_i = \frac{1}{2}|\mu_i(\alpha + 1)| + \frac{1}{2}\sqrt{\mu_i^2(\alpha + 1)^2 - 4\alpha} \quad (10)$$

Since clearly  $\frac{\partial \rho_i}{\partial |\mu_i|} > 0$ , we can obtain that  $\rho(\mathcal{B}) = \rho_1$  when the roots of Eq. (3) are real. From (3), whether  $\lambda_{11}$  and  $\lambda_{12}$  are real or complex determined by the sign of the following quadratic function:

$$f(\alpha) = \mu_1^2(\alpha + 1)^2 - 4\alpha$$

Since  $|\mu_i| < 1, i \in \mathbb{I}[1, mn]$ , it can be checked that  $f(\alpha) = 0$  has the two real roots as defined in (7). Now we can conclude that for any  $\mu_1 < 1$ ,  $\lambda_{11}$  and  $\lambda_{12}$  are complex if and only if  $\alpha_{11} < \alpha < \alpha_{12}$ . That is, from Eq. (8),  $\rho(\mathcal{B}) = \sqrt{\alpha}$  when  $\alpha_{11} < \alpha < \alpha_{12}$ .

**Lemma 6.** The function  $\rho(\mathcal{B})$  defined in (6) has the following monotonicity

$$\frac{\partial \rho(\mathcal{B})}{\partial \alpha} : \begin{cases} < 0, \alpha \leq \alpha_{11} \\ > 0, \alpha > \alpha_{11} \end{cases} \quad (11)$$

Due to the space limitation, the proof of Lemma 6 is omitted. We are now ready to prove Theorem 1.

*Proof.* By Lemma 5,  $\rho(\mathcal{B})$  is a continuous function of  $\alpha$ ,  $\alpha_{11} \in [0, 1)$  and when  $\alpha = \pm 1$ ,  $\rho(\mathcal{B}) = 1$ . We can further get the following relations by Lemma 6

$$\rho(\mathcal{B}) : \begin{cases} < 1, -1 < \alpha < 1 \\ \geq 1, \alpha \leq -1 \text{ or } \alpha \geq 1 \end{cases} \quad (12)$$

$$\rho(\mathcal{B}) : \begin{cases} > \rho(B), -1 < \alpha < 0 \text{ or } \mu_1^2 < \alpha < 1 \\ = \rho(B), \alpha = 0 \text{ or } \alpha = \mu_1^2 \\ < \rho(B), 0 < \alpha < \mu_1^2 \end{cases} \quad (13)$$

The first relation (12) implies that  $-1 < \alpha < 1$  is the necessary and sufficient condition for algorithm (2) to converge to the solution of  $Ax = b$  for arbitrary initial condition. And the second relation (13) shows the conditions when algorithm (2) converges faster or slower than algorithm (1). Finally, by relation (11), it is clear that  $\alpha = \alpha_{11}$  is the minimum point of  $\rho(\mathcal{B})$  which corresponds to the fastest convergence speed.

## 4 Simulation

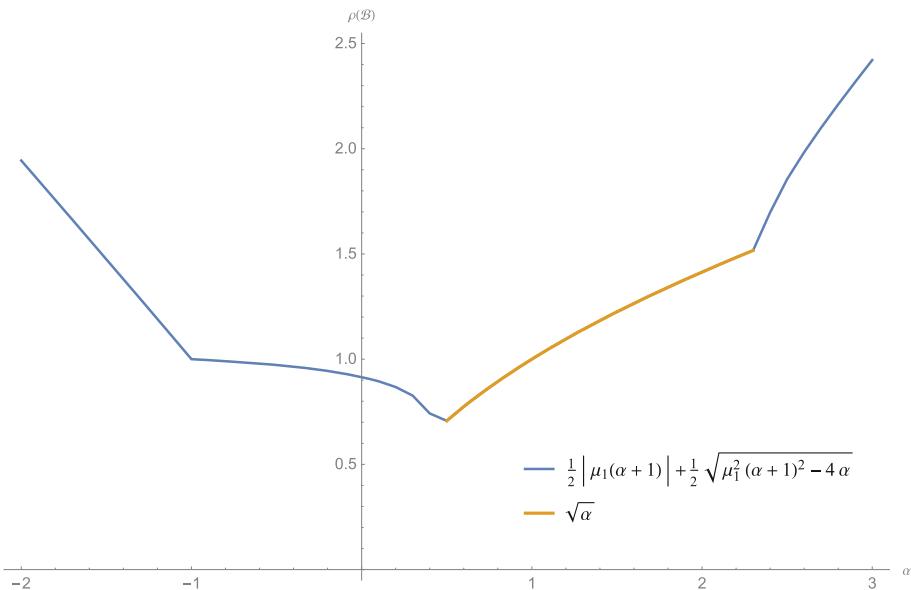
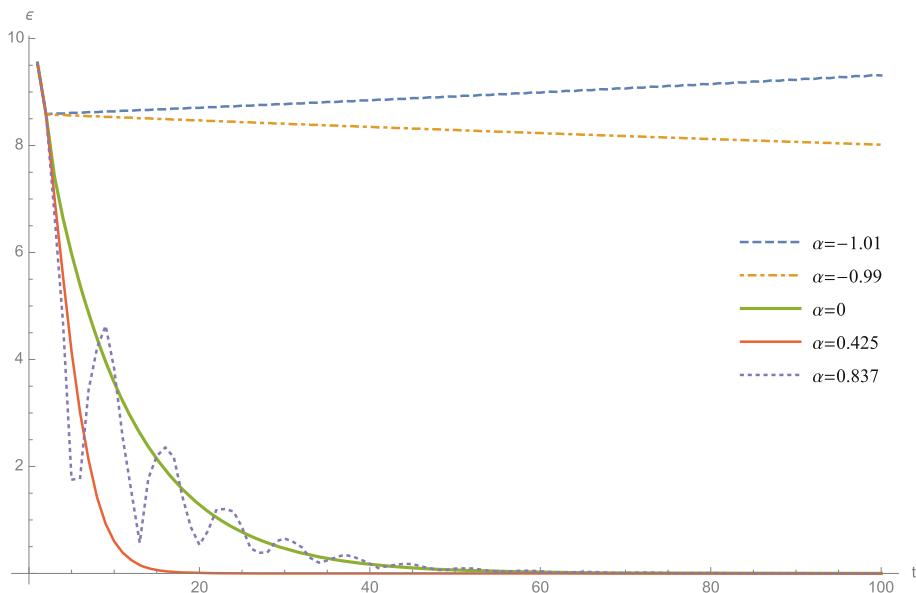
In this section, we verify the results presented in Theorem 1 by a simple numerical example. Consider a connected ring network with 6 agents and each agent owns a distinct row of the following linear algebraic equation

$$\begin{bmatrix} 100 & 1 & 1 & 1 & 1 & 2 \\ 1 & 99 & 1 & 1 & 1 & 2 \\ 1 & 1 & 98 & 1 & 1 & 2 \\ 1 & 1 & 1 & 97 & 1 & 2 \\ 1 & 1 & 1 & 1 & 96 & 2 \\ 1 & 1 & 1 & 1 & 1 & 95 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 126 \\ 223 \\ 318 \\ 411 \\ 502 \\ 585 \end{bmatrix}$$

Firstly, we plot the spectral radius function  $\rho(\mathcal{B})$  of  $\alpha$  in Fig. 1 by Eq. (6). It can be seen that the monotonicity of  $\rho(\mathcal{B})$  is in accordance with relation (11). By Theorem 1 two special parameters are calculated of which one ( $\alpha = 0.425$ ) is supposed to let algorithm (2) have the fastest convergence speed and the other one ( $\alpha = 0.837$ ) may promise the same convergence speed as algorithm (1). We simply take  $x_i(0) = 0, i \in \mathbb{I}[1, 6]$  since algorithm (2) works for arbitrary initial condition. Here we use

$$\varepsilon(t) = \frac{1}{m} \sum_{i=1}^m \|x_i(t) - x^*\|$$

to measure the average estimation error of our iterative algorithm. Next, the performance of the proposed algorithm (2) is shown in Fig. 2, from which we can see that the fastest convergence rate is achieved when  $\alpha = 0.425$  (more than 3 times faster than the case when  $\alpha = 0$  or  $\alpha = 0.837$ ). It should be noted that the proposed algorithm (2) is reduced to algorithm (1) when  $\alpha = 0$ .

**Fig. 1.** Curve of  $\rho(\mathcal{B})$  as a function of  $\alpha$ **Fig. 2.** Performance of algorithm (2) with different  $\alpha$

## 5 Conclusion

In this paper, a novel distributed algorithm has been presented to solve  $Ax = b$ . The central features of this algorithm are threefold: i) It has the ability to solve any  $Ax = b$  with a unique solution for arbitrary initial condition in a distributed manner. ii) Both current estimations and previous estimations are used in the update equation. iii) Faster convergence rate can be achieved if the relaxation parameter is chosen carefully.

One of the problems of this algorithm is that it is difficult to come up with a provably correct method to determine the relaxation parameter  $\alpha$  for achieving faster convergence rate only using local information. Nevertheless, the underlying idea (making use of the information of previous estimations) actually sheds the light of accelerating many other existing distributed algorithms. Another issue is to reduce the restriction of the communication topology, which will be discussed in our future work.

## References

1. Bernstein, D.S.: Matrix Mathematics: Theory, Facts, and Formulas. Princeton University Press, Princeton (2009)
2. Cihan, O.: Rapid solution of linear equations with distributed algorithms over networks. IFAC-PapersOnLine **52**(25), 467–471 (2019)
3. Hu, W., Huang, W., Huang, Y., Chen, S., Wu, A.G.: On reachable set estimation of multi-agent systems. Neurocomputing **401**, 69–77 (2020)
4. Liu, J., Morse, A.S.: Accelerated linear iterations for distributed averaging. Ann. Rev. Control **35**(2), 160–165 (2011)
5. Liu, J., Morse, A.S., Nedić, A., Başar, T.: Exponential convergence of a distributed algorithm for solving linear algebraic equations. Automatica **83**, 37–46 (2017)
6. Liu, J., Mou, S., Morse, A.S.: Asynchronous distributed algorithms for solving linear algebraic equations. IEEE Trans. Autom. Control **63**(2), 372–385 (2017)
7. Mo, Y., Murray, R.M.: Privacy preserving average consensus. IEEE Trans. Autom. Control **62**(2), 753–765 (2016)
8. Mou, S., Liu, J., Morse, A.S.: A distributed algorithm for solving a linear algebraic equation. IEEE Trans. Autom. Control **60**(11), 2863–2878 (2015)
9. Mou, S., Morse, A.S.: A fixed-neighbor, distributed algorithm for solving a linear algebraic equation. In: 2013 European Control Conference (ECC), pp. 2269–2273. IEEE (2013)
10. Nedic, A., Ozdaglar, A., Parrilo, P.A.: Constrained consensus and optimization in multi-agent networks. IEEE Trans. Autom. Control **55**(4), 922–938 (2010)
11. Pasqualetti, F., Carli, R., Bullo, F.: Distributed estimation via iterative projections with application to power network monitoring. Automatica **48**(5), 747–758 (2012)
12. Tang, Y., Mei, J.: Distributed algorithms for solving a linear equation under a directed graph. In: 37th Chinese Control Conference (CCC), pp. 7193–7198. IEEE (2018)
13. Wang, L., Fullmer, D., Morse, A.S.: A distributed algorithm with an arbitrary initialization for solving a linear algebraic equation. In: American Control Conference (ACC), pp. 1078–1081. IEEE (2016)

14. Wang, P., Mou, S., Lian, J., Ren, W.: Solving a system of linear equations: from centralized to distributed algorithms. *Ann. Rev. Control* **47**, 306–322 (2019)
15. Wang, X., Mou, S., Sun, D.: Improvement of a distributed algorithm for solving linear equations. *IEEE Trans. Ind. Electron.* **64**(4), 3113–3117 (2016)
16. Wu, A.G., Sun, H.J., Zhang, Y.: An SOR implicit iterative algorithm for coupled Lyapunov equations. *Automatica* **97**, 38–47 (2018)



# CNN-Based Automatic Diagnosis for Knee Meniscus Tear in Magnetic Resonance Images

Hao Zhou<sup>1,2</sup>, Liyan Zhang<sup>1,2</sup>, Bing Zhang<sup>1,2</sup>, Juan Wang<sup>3</sup>,  
and Chengyi Xia<sup>1,2(✉)</sup>

<sup>1</sup> Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300384, People's Republic of China  
[cxyia@email.tjut.edu.cn](mailto:cxyia@email.tjut.edu.cn), [xialooking@163.com](mailto:xialooking@163.com)

<sup>2</sup> School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, People's Republic of China

<sup>3</sup> School of Electrical and Electronic Engineering, Tianjin University of Technology, Tianjin 300384, People's Republic of China

**Abstract.** The meniscus tear is common for athletes and the elderly, and because of its irresistible nature, early diagnosis and treatment is particularly important. Magnetic Resonance (MR) Imaging is noninvasive and has a high diagnostic accuracy of 98%, which has been considered as an ideal method to diagnose meniscus tear. However, it is a time consuming process for radiologists to diagnose meniscus tear by comparing dozens of MR images and to diagnose the tear grade. Here, the convolutional neural network (CNN) is used to accomplish the aim of automatic diagnosis of tear grade. At first, we apply the Hough transformation to preprocess the data, during which we shrink the image to about 1/10 of its original size so that the meniscus local candidate frame (LCF) is formed. Then, to validate the method, 3062 actual clinical MR images were used and an accuracy of 89.5% is achieved. The experimental result has demonstrated the validity of our proposed method, which can meet the needs of radiologists for computer-aided diagnosis of meniscus tear classification.

**Keywords:** Computer-aided diagnosis · Meniscus tear · Convolutional neural network · Magnetic resonance imaging · Local candidate frame

## 1 Introduction

The meniscus is a 2-crescent-shaped fibrous cartilage located on the medial and lateral articular surfaces of the tibial plateau. The sagittal face of the meniscus is triangular, the lateral side is thick and the medial side is thin, the upper part is slightly concave to fit the femoral condyle, and the lower part is flat to meet the tibial plateau. Such a structure makes the femoral condyle form a deep depression on the tibial plateau, which increases the stability of the spherical

femoral condyle and tibial plateau. The meniscus can help to improve the joint surface anastomosis, and hence reduce the friction and absorb the impact force. The typical sagittal view of MR image of knee joint is shown in Fig. 1. At present, the annual prevalence of meniscus tears is as high as 12% to 14%, and it is ranked as the second most common knee injury [1]. Reicher et al. [2] divided meniscus tear into 4 grades, which include tear grade 1 to 4, is regarded as the most authoritative classification method till nowadays. Due to the irreversible nature of meniscus tears, low grade tears can easily deteriorate to high level tears without timely and effective treatment. Thus, the early detection and diagnosis is particularly important for the patients with knee joint injury.

Knee meniscus detection methods mainly include Magnetic Resonance Imaging and arthroscopy detection methods. Since the mid-1980s, the reliability and safety of Magnetic Resonance Imaging (MRI) have been proved in clinical application [3], and MRI is also considered to be the best noninvasive mechanism for detecting the intra-articular structure of knee and diagnosing meniscus tear [4]. However, compared with MRI, arthroscopy is one type of the immersion detection, which needs to pierce the skin and has certain risks for patients. Therefore, arthroscopy is just viewed as the most ideal surgical treatment scheme [5,6], while MRI is the first choice for most patients because of its non-invasion properties and lower risks.



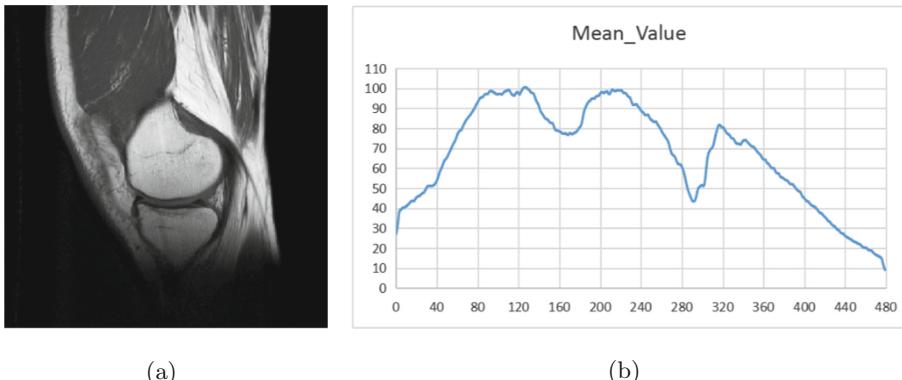
**Fig. 1.** The sagittal view of MR image of knee joint.

## 2 Local Candidate Frame Definition of Meniscus

Local candidate frame (LCF) is defined to shrink the whole image to a suitable size including the target area, which is the focus of the image analysis for knee meniscus and can help to reduce the processing time and increase the accuracy.

## 2.1 Method Review

By defining a fixed size window in the whole meniscus image, Ramakrishna et al. [7] utilized the mean method to select the region of interest. After selecting a fixed size window in a T1-weighted image, they can calculate the mean value of each row of pixels and sort them. Taking the row with the minimum value of the mean as the center of the region of interest, and then we can include 50 rows up and 49 rows below the centered one, total 100 rows are defined as the area of interest of meniscus [7]. However, it is often difficult to determine the size and location of the fixed window in the experiment, which can only be marked manually by radiologists or experts. This manual process is not up to the spirit of computer-aided diagnosis automation, which wastes a lot of time and increases the workload of doctors to a certain extent. But if such a window of fixed size is not defined, the entire meniscus image is directly sorted according to the mean value of the rows, and it will be found that the mean sorting method is not appropriate to select or set the region of interest of meniscus as illustrated in Fig. 2.



**Fig. 2.** (a) A sagittal T1-weighted image of the knee joint (Image size is 480 \* 480 pixels). (b) Row-wise average intensity profile of the (a) mean image. The abscissa indicates that the image is divided into 480 rows according to the image size and the ordinate represents the mean value of the row image.

## 2.2 The Introduction of Hough Transformation

To solve the above-mentioned problem, Hough transformation is introduced here, and it is a typical approach for the feature detection widely used in the image analysis and computer vision [8–10]. The basic principle of Hough transformation can be briefly described as follows: Based on the duality of points and lines, a designated curve in the original image space is transformed into a point in the parameter space through the curve expression [11–13]. By use of this method, how to detect a designated curve in the original space is converted into the problem of finding the peak value in the other parameter space.

### 2.3 Detection of Circles in the Meniscus LCF Using the Hough Transformation

The Hough transformation detects the circles in the same way as it detects the straight lines. The expression of the circle is described as:

$$(x - a)^2 + (y - b)^2 = r^2 \quad (1)$$

where  $(x, y)$  represents the coordinates of points on the circle,  $(a, b)$  denotes the coordinates of the center of the circle, and  $r$  represents the radius of the circle, and the problem is converted to the parameter pairs  $(a, b, r)$  that pass through the  $(x, y)$  pixel most. Here we will find that the parameter space of  $(a, b, r)$  is extremely large and the computation is extremely large, and thus we use the Hough gradient method to solve the transformation of a circle. If we calculate the gradient of a circle, the gradient of all points on the circle is oriented towards the center of the circle.

In addition, the meniscus is located between the femur and tibia, and the shape of the femur is close to a circle. Therefore, the position of the femur can be found by Hough transformation detecting the circle so that the position of the meniscus LCF is determined. In order to locate the meniscus, we propose a meniscus LCF extraction algorithm described as the Algorithm 1.

---

#### Algorithm 1. Meniscus LCF extraction algorithm.

---

**Input:**

Image with meniscus,  $img$ ;

**Output:**

A image of meniscus LCF with  $LCF_{width} = 200$  and  $LCF_{height} = 120$ ;

1: Read the  $img$  and convert it to COLOR\_BGR2GRAY mode.

2: Initialize parameters:  $\text{maxRadius} = 100$ ,  $\text{threshold} = 100$ .

3: **while**  $\text{threshold} > 1$  **do**

4:   circle=HoughCircles( $img$ ,  $\text{threshold}$ ,  $\text{maxRadius}$  )

5:   add circle( $x, y, r$ ) to circlesSet

6:    $\text{maxRadius} -= 5$ ,  $\text{threshold} -= 2$

7:   Judge whether the radius is less than 40.

8: **end while**

9: **for** circle  $\in$  circlesSet **do**

10:   Calculate Eu\_dst(circle\_centre,image\_centre) with Equation (2)

11:   Add Eu\_dst to circle

12: **end for**

13:  $(x, y, r, \text{Eu\_dst}) = \text{circlesSet}[\text{argmin}_n(\text{circle}(\text{eu\_dst}))]$

14: croppedLCF = image[int(y):int(y)+120, int(x)-100:int(x)+100]

---

The circle detection algorithm based on the Hough transformation is not strong enough to stand on its own if we do not know exactly how much the radius of circle in the image is. If we do not know the exact radius, we need the large-scale check and post-processing verification. Luckily, Hough transformation

detecting circles algorithm is pretty quick so that we can guess the radius by iteration. First, we read an image and use cv2 to convert it to COLOR\_BGR2GRAY for Hough transformation detection. At the beginning of the detection, we need to initialize two parameters as follows:

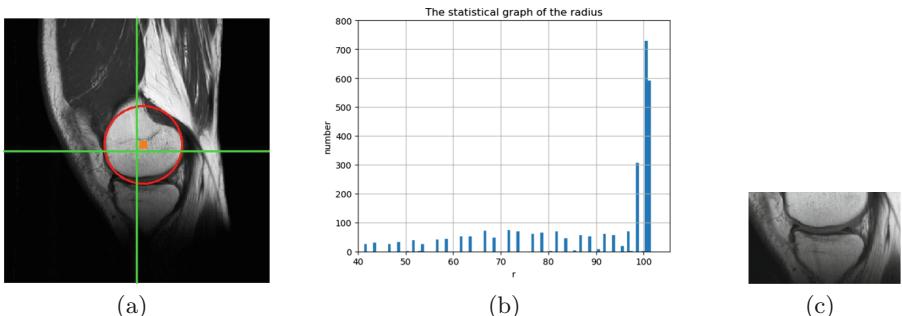
- 1) MaxRadius is used to represent the radius of the largest circle, which is decreased by 5 in each iteration. By observing the femur, we decide not to consider the circle with radius less than 40.
- 2) Threshold is used to represent the accumulator threshold of the circle in the detection stage, which is decreased by 2 in each iteration. The smaller the value is, the more false circles may be found. Circles, corresponding to the larger accumulator values, will be got first [14].

After the above work is completed, a set of circles can be obtained. When the number of circles in this set is greater than 2, an optimal circle needs to be selected. By calculating the Euclidean distance between the center of all circles and the center of the whole image, we can take the minimum index to determine the fittest circle. The Euclidean distance can be measured as:

$$Eu\_dst = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (2)$$

where  $(x_1, y_1)$  represents the coordinates of the center of the circle and  $(x_2, y_2)$  denotes the coordinates of the center of the image.

Then, we obtained the fittest femur shape circle as shown in Fig. 3(a) and their center coordinates  $(x, y)$  and radius  $r$  of the circle. Figure 3(b) provides the statistics of the radii obtained after our detection algorithm examines all images. As presented in Fig. 3(b), the meniscus is mostly located in the lower half of the circle. After the statistics of all circle radii,  $r = 100$  (pixels) was selected as the baseline radius so as to unify the size of meniscus LCF. In addition,  $r + 20$  (pixels) is used as the height of meniscus LCF and  $2 * r$  (pixels) is used as the width of meniscus LCF. The resulting meniscus LCF is shown in Fig. 3(c).



**Fig. 3.** (a) The fittest circle for femur. (b) The statistics of radii. (c) The meniscus LCF.

### 3 Using CNN to Predict the Grade of Meniscus Tear

#### 3.1 Data Sets

MRI data of meniscus in the experiment were picked up from real medical clinical examination, which were provided by the first affiliated hospital of Anhui Medical University, and were graded by four professional doctors who were divided into two groups. Each group of doctors verified the judgment of the other group to build the final data sets. The initial data included images with a size of 480 \* 480 pixels, and also has images with a size of 370 \* 370 pixels. In order not to lose the image information and unify the image size, we resized the size of all images to 480 \* 480 pixels. The number of meniscus tearing images of each grade is included in Table 1, and the size of the corresponding meniscus LCF image is 200 \* 120 pixels.

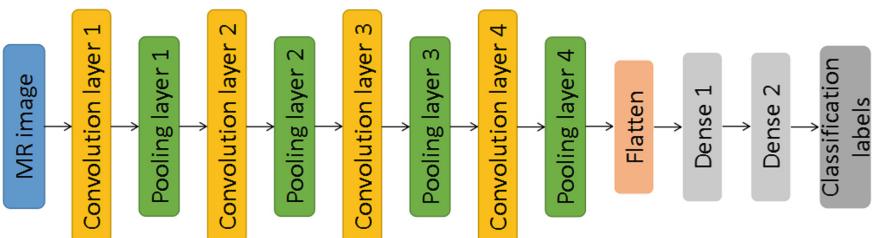
**Table 1.** The number of meniscus tearing images of each grade.

	Tear grade 1	Tear grade 2	Tear grade 3	Tear grade 4	Total
Number	611	1223	873	355	3062

#### 3.2 CNN Model

In this study, convolutional neural network is used to automatically classify the tear grade. The corresponding neural network consists of four convolution layers and four pooling layers. After each convolution operation, the rectified correction unit (ReLU) is used as the activation function to improve the classification effect [15]. The flatten layer is used to “flatten” the data to form a vector. Two dense layers, in which one takes the ReLU as the activation function and the other one uses the softmax as the activation function, will sample vectors and output plane classification labels. The structure of the CNN network model is depicted in Fig. 4 and the specific model setup is provided in Table 2.

Here, all the experiments were run on a workstation installing the Windows 10 operating system, the related algorithms are programmed with python. The hardware environment of the workstation is Intel Core @3.60 GHz CPU, 4 pieces of NVIDIA GeForce RTX 2080 Ti graphics card and 64 GB RAM.



**Fig. 4.** The structure of the CNN network for image analysis.

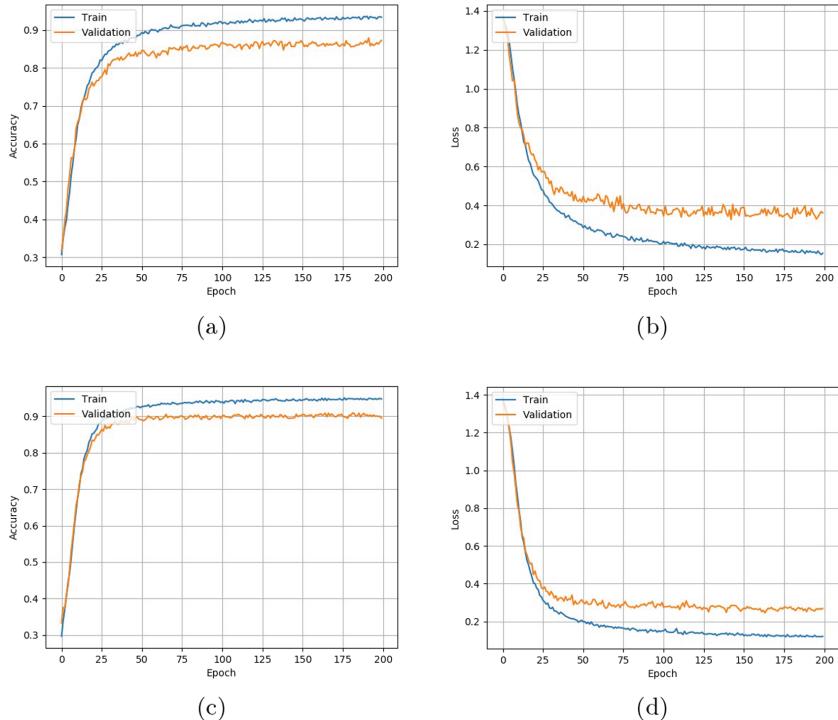
**Table 2.** The specific information of the CNN model.

Layer	Kernel size	Stride	Channel
Conv 1	5 * 5	1	32
Pool 1	2 * 2	1	32
Conv 2	3 * 3	1	64
Pool 2	2 * 2	2	64
Conv 3	3 * 3	1	96
Pool 3	2 * 2	2	96
Conv 4	3 * 3	1	96
Pool 4	2 * 2	2	96
Flatten	–	–	–
Dense 1	–	–	512
Dense 2	–	–	4

### 3.3 Analysis of Results

The data sets of the entire MR images and meniscus LCF obtained by Algorithm 1 have been separately put into the convolutional neural network for training, and we set the training epochs to be 200. Figure 5 presents the accuracy and loss information of training and validation of data sets regarding the whole meniscus image and LCF. Among them, Fig. 5(a) and (b) show the accuracy and loss of the entire meniscus image data set after training, respectively, while Fig. 5(c) and (d) display the accuracy and loss of the meniscus LCF data set after training, respectively. By comparing Fig. 5(a) and (c), it can be observed that the LCF data set converges faster than the data set of the whole image, where LCF data set converges after about 60 epochs while the entire MR image data set converges after about 125 epochs.

Finally, it takes about 6 h to train 200 epochs for the data set of the whole meniscus MR images, and the eventual accuracy arrives at 87.3%. While the current algorithm comes to the date set of meniscus LCF, it takes about 55 min under the same parameter and hardware setup, and the final accuracy comes up to 89.5%. Meanwhile, we compare the accuracy of meniscus tear detection of our algorithm with those obtained by other methods, which are shown in Table 3. In the literature studies mentioned in Table 3, the accuracy of others only considers whether there is tearing, while our result is an accurate detection of meniscus tear grade from 1 to 4, which is more detailed than the results of previous studies. The current results indicate that our method has reached the highest accuracy for the largest MR image data sets, which will further aid in the computer-aided diagnosis for the radiologists.



**Fig. 5.** The loss-value curves and accuracy curves of training set and validation set. (a) The CNN accuracy curves of the date set of entire MR image. (b) The CNN loss curves of the date set of entire MR image. (c) The CNN accuracy curves of the date set of meniscus LCF. (d) The CNN loss curves of the date set of meniscus LCF.

**Table 3.** Accuracy of the meniscus tear detection studies in the literature.

Author(s)	Size of Data	Accuracy
Saygili et al. [16]	1031 Slices	84.97%
Köse et al. [17]	500 Slices	88.3%
Boniatis et al. [18]	55 Slices	89.1%
This study	3062 Slices	89.5%

## 4 Conclusion

In summary, we define the local candidate frames of meniscus by using Hough transformation to detect circles, and find that the local candidate frames of meniscus only account for about 1/10 of the size of the original MR image. Compared with the processing of the whole image, the time and calculation cost are greatly reduced. Furthermore, convolutional neural network has been introduced to classify meniscus tear grade accurately, and its classification accuracy

has reached 89.5%, which is the highest in the state-of-art methods at present. The experimental result has proved the effectiveness of the proposed method, which can meet the needs of radiologists for computer-aided diagnosis of meniscus tear classification.

In the future work, we plan to further segment the meniscus to obtain the more accurate position of the meniscus LCF. Then, we can locate the tear lesions from this more precise location. At last, we expect to perform the tear classification more quickly and accurately.

**Acknowledgments.** This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61773286, and Tianjin Municipal Natural Science Foundation under grant No. 18JCYBJC87800.

## References

1. Logerstedt, D.S., Scalzitti, D.A., Bennell, K.L., et al.: Knee pain and mobility impairments: meniscal and articular cartilage lesions revision 2018. *J. Orthop. Sports Phys. Therapy* **48**(2), A1–A50 (2018)
2. Reicher, M.A., Hartzman, S., Duckwiler, G.R., et al.: Meniscal injuries: detection using MR imaging. *J. Radiol.* **159**(3), 753–757 (1986)
3. Oei, E., Nikken, H.G., et al.: MR imaging of the menisci and cruciate ligaments: a systematic review. *J. Radiol.* **226**(3), 837–848 (2003)
4. MacFarlane, L.A., Yang, H., Collins, J.E., et al.: Associations among meniscal damage, meniscal symptoms and knee pain severity. *J. Osteoarthritis Cartilage* **25**(6), 850–857 (2016)
5. Rangger, C., Klestil, T., et al.: Influence of magnetic resonance imaging on indications for arthroscopy of the knee. *J. Clin. Orthopaedics Relat. Res.* **330**(330), 133–142 (1996)
6. Rappeport, E.D., Mehta, S.: MR imaging before arthroscopy in knee joint disorders? *J. Acta Radiol.* **37**, 602–609 (1996)
7. Ramakrishna, B., Liu, W., Saiprasad, G., et al.: An automatic computer-aided detection system for meniscal tears on magnetic resonance images. *J. IEEE Trans. Med. Imaging* **28**(8), 1308–1316 (2009)
8. Kuo, W.J., Lin, C.C.: Two-stage road sign detection and recognition. In: *IEEE International Conference on Multimedia and Expo*, pp. 1427–1430 (2007)
9. Cai, J., Zhou, X., Li, Y., et al.: Recognition of mature oranges in natural scene based on machine vision. *J. Trans. Chin. Soc. Agric. Eng.* **24**(1), 175–178 (2008)
10. Jonker, P., Caarls, J., Bokhove, W.: Fast and accurate robot vision for vision based motion. In: *Lecture Notes in Computer Science*, pp. 149–158 (2001)
11. Duda, R.O., Hart, P.E., et al.: Use of the Hough transformation to detect lines and curves in pictures. *J. Commun. ACM* **15**(1), 11–15 (1972)
12. Ballard, D.H.: Generalizing the Hough transform to detect arbitrary shapes. *J. Pattern Recognit.* **13**(2), 111–122 (1981)
13. Mei, D., Chen, D.: Optic disc segmentation method based on low rank matrix recovery theory. In: *Chinese Control and Decision Conference* (2018)
14. Sengupta, S., Lee, W.S.: Identification and determination of the number of immature green citrus fruit in a canopy under different ambient light conditions. *J. Biosyst. Eng.* **117**, 51–61 (2014)

15. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning, vol. 3, pp. 804–814 (2010)
16. Saygili, A., Albayrak, S.: An efficient and fast computer-aided method for fully automated diagnosis of meniscal tears from magnetic resonance images. *J. Artif. Intell. Med.* **97**, 118–130 (2019)
17. Köse, C., Gençalioğlu, O.: An automatic diagnosis method for the knee meniscus tears in MR images. *Expert Syst. Appl.* **36**, 1208–1216 (2009)
18. Boniatis, I., Panayiotakis, G., Panagiotopoulos, E.: A computer-based system for the discrimination between normal and degenerated menisci from magnetic resonance images. In: Imaging Systems and Techniques, pp. 335–339 (2008)



# Application of an Effective Fault Localization Prioritization Method to Stereo Matching Software

Jinfeng Li<sup>1(✉)</sup>, Yan Zhang<sup>1</sup>, and Jilong Bian<sup>2</sup>

<sup>1</sup> College of Computer and Information Technology, Mudanjiang Normal University,  
Mudanjiang 157011, China  
[allylili0453@163.com](mailto:allylili0453@163.com)

<sup>2</sup> College of Information and Computer Engineering, Northeast Forestry University,  
Harbin 150040, China

**Abstract.** During the process of software development, there will always exist some faults. At this time, we need to locate the specific faults and fix them. Fault Localization is a complicated and time-consuming process. This paper utilizes the executed test cases to establish the prediction model of test case execution results through neural network and sorts the unexecuted test cases according to the coverage and prediction results. Executes the test cases according to this order and records the execution status. Substitutes the execution information into the formula of Tarantula and calculates the suspiciousness values of program elements. Developers check program elements in non-increasing order of suspiciousness values until a fault is found. Finally, the method proposed in this paper is applied to stereo matching software, and it is found that the efficiency is higher than that of random method.

**Keywords:** Neural network · Fault localization · Stereo matching

## 1 Introduction

During the software development period, continuous software testing is required to ensure software reliability. Many literatures have studied how to use existing test cases or generate new test cases to find faults in software, namely fault detection. According to the influence of different execution orders of test cases on the effectiveness of fault detection, many literatures propose fault detection prioritization methods to detect software faults as early as possible. When faults are detected in the software, it is necessary to determine the location of faults according to the executed test cases or some additional test cases, i.e. fault localization. Some literatures propose formulas to calculate suspiciousness values of program elements according to the execution information of test cases. Developers check each element in non-increasing order according to the suspiciousness values until a fault is found. The goals of fault detection and fault localization

are different, so the requirements for test cases are different. After locating the faults, developers turn to fix the faults. It is always accompanied by numerous fault detection–fault localization–fault fix in the entire process of software development.

In this paper, an effective fault localization prioritization method is proposed. Applying neural network model according to the executed test cases to predict the running results of unexecuted test cases, i.e. the true and false values of test case execution result. Then apply a vector-based method to sort test cases. The method is applied to stereo matching software, which demonstrates that the method suggested in this paper can locate faults faster.

## 2 Related Work

When a software fault is detected, we ought to find fault location as soon as possible, that is, fault localization, so as to facilitate developers to fix it. Fault localization is a complicated process. When we detect faults, the count of executed test cases maybe limited and cannot provide sufficient information for our fault localization. At this time, we need to further execute more test cases to extract more information. How to select test cases that can provide more useful information for fault localization from unexecuted test cases has become an important problem. In order to settle this problem, some documents have put forward their own solutions.

Sejun Kim [1] et al. put forward an valid fault aware test case prioritization method combined with fault localization technology. They think that test cases with high coverage are not always effective. The faults found in previous versions have been removed during debugging, so the areas containing faults in previous versions are less likely to contain faults again in subsequent versions. This means that the detection capability of test cases covering these faults may be decreased. Then the ordering method of test cases based on coverage information may be invalid. The authors introduce a new test case prioritization technology, which combines fault localization technology with coverage and historical fault information. The method uses the historical fault detection information of test cases to adjust the priority of test cases that find defects. This method can decrease the overall cost of executing the whole test suite and can detect faults earlier in the testing process. Wenhao Fu [2] et al. proposed a method for ordering test cases that effectively improves fault localization. Program elements are divided into two groups: high suspicious elements and common elements, and select the test case that can maximize the changing of suspicious ranks of high suspicious elements and minimize the changing of common elements to execute. Sorting test cases with this standard can help reduce the amount of debugging work. Xiaoyi Zhang [3] et al. proposed a black-box prioritizing method with input information. One method is to select the most different test case from the executed test case from the unexecuted test cases to execute. The other method is to select the most different test case from the successfully executed test cases from the unexecuted test cases to execute. The third method is to randomly choose 10 test cases

of successfully executed and choose the test case that is the most different from the 10 test cases from the unexecuted test cases to execute. Use these methods to maximize the accuracy of fault localization. Xiaoyi Zhang [4] et al. proposed a random fault localization prioritization method based on coverage. First, randomly select several test cases from the unexecuted test cases as candidates, and select test cases from the candidates that can distinguish statements with the same suspiciousness value to execute. If the distinction degree between the two test cases is in common, then select the test case that is most likely to fail. This method can quickly locate defective statements and reduce the number of test cases. Shin Yoo [5] et al. put forward a fault localization prioritization method on the basis of information theory, and calculated the information entropy value for each test case. Test cases with smaller information entropy will provide more information for fault localization, so they will be executed preferentially.

Although there are many fault localization methods, there is still room for improvement in fault localization capability, and these methods are not suitable for all programs and all fault types. The fault localization method proposed in this paper can locate faults faster in stereo matching software.

### 3 Methodology

#### 3.1 Basic Ideas

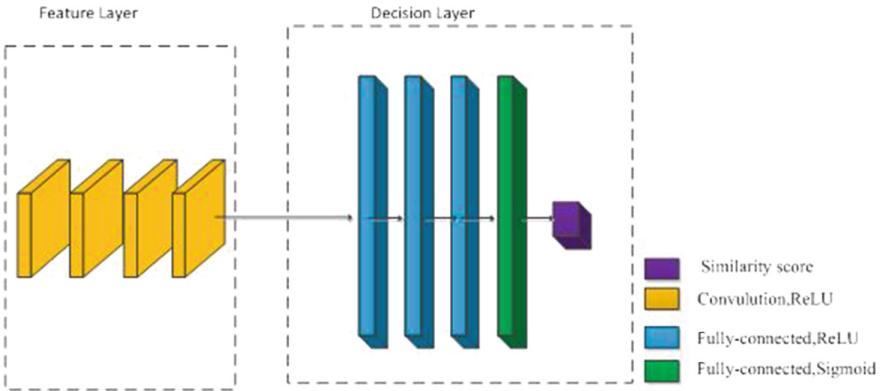
When we detect faults in software, we usually need to immediately switch to the fault localization. The test cases that have been executed may be too few to locate faults effectively. Under the circumstances, we need to execute more test cases to obtain testing information for effective fault localization. So, how to prioritize the unexecuted test cases to maximize the early fault localization?

In order to solve this problem, this paper proposes an effective fault localization method. Passed test cases and failed test cases both play a certain role in fault localization, and failed test cases can be used to limit the scope of elements with faults, so we give high priority to those failed test cases with fewer coverage. Passed test cases indicate the range of elements do not have faults, so the passed test cases with higher coverage are preferentially selected to execute. Before executing test cases, we do not know the execution results of test cases in the current version, so we use the executed test cases to build a prediction model with neural network to forecast the execution result of unexecuted test cases. Alternate execution of failed test cases with fewer coverage and passed test cases with higher coverage. Finally, the formula of Tarantula is applied to calculate suspiciousness value for each element. Developers check elements according to the suspiciousness values in non-increasing order until a fault is located.

#### 3.2 Prediction Model Based on Neural Network

At present, deep neural network has achieved excellent results in image recognition, objection detection and other fields. For this reason, this paper wants to

apply the deep neural network to test case prediction. Test cases are divided into passed test cases and failed test cases, so this model belongs to binary classification model. The deep model is divided into feature layer and decision layer in Fig. 1. In the feature layer, four convolution layers are selected, the size of each convolution kernel is  $3 \times 3$  and the number of features is 112. Four linear layers are selected in the decision layer. The input of the first linear layer is 112, the output is 224. The input and output of the 2nd and 3rd linear layers are 224. The input of the 4th linear layer is 224, and the output is 2. Finally, the loss function selects the cross entropy loss. This paper utilizes the model to train on the executed test cases, after that uses the model to predict the running results of unexecuted test cases in the software testing process.



**Fig. 1.** Deep model

### 3.3 Vector-Based Test Case Prioritization

Element vector refers to the set of elements covered by a test case. As shown in Table 1,  $e$  represents a program element and the program contains 10 elements.  $t$  is a test case. When a fault is detected in the program, the test suite includes 10 unexecuted test cases. “√” means that the test case covers the corresponding program element. “F” in the status line indicates that the neural network model predicts that the test case failed to execute and “P” indicates that the execution is passed. The element vector set of test case  $t_1$  is  $\{e_1, e_2, e_3, e_5, e_6\}$  in Table 1. The test cases with the same vector are divided into the same group. The test suite is split into five groups,  $\{t_1, t_7, t_8\}$ ,  $\{t_2, t_9\}$ ,  $\{t_3, t_6\}$ ,  $\{t_4, t_{10}\}$  and  $\{t_5\}$ , each group covers 5, 4, 5, 6 and 4 program elements respectively, as shown in Table 2.

Passed test cases and failed test cases also play a certain role in fault localization. The failed test cases and passed test cases are selected alternately to execute. The program elements covered by the failed test cases may contain defective elements, and test cases with the same execution result in one group have identical coverage area and make similar contributions to fault localization.

In order to narrow the scope of defective elements, a failed test case is randomly selected from each test case group in non-decreasing order of element vector. Only after all groups have selected a failed test case, then they can re-select in the same group in non-decreasing order of element vector. Passed test cases limit the scope of fault-free elements, so a passed test case is randomly selected from each test case group in non-increasing order of element vector. The selection method is the same as the failed, except that in the order of non-increasing of element vectors.

In this example, the 2nd group and 5th group of test cases have the least element vector. One group can be randomly selected, if the 2nd group is selected, from which the failed test case  $t_9$  is selected as the first test case to be executed. Then select a passed test case from group with the largest element vector. The 4th group has the largest element vector, but does not contain the passed test cases. Therefore, one group is randomly selected from the 1st group and the 3rd group with the largest element vector, which contains passed test cases. The 1st group is selected and select the passed test case  $t_1$  to execute. The 5th group with the smallest element vector does not include failed test cases, so one group is randomly selected from the 1st group and the 3rd group with the smaller element vector. If the 1st group is selected, in which there are two failed test cases, and the test case  $t_7$  is randomly selected. Next, the passed test case  $t_3$  from the 3rd group and the failed test case  $t_4$  from the 4th group are selected. The 2nd group and the 4th group have the same element vector, we randomly select the passed test case  $t_2$  from the 2nd group. The failed test cases have been selected from all test case groups, so we can select again in non-decreasing order. The final test case execution sequence is  $t_9, t_1, t_7, t_3, t_4, t_2, t_8, t_5, t_{10}, t_6$ .

Execution in this order can be stopped whenever necessary according to the actual situation. The execution results are substituted into Tarantula, and the suspiciousness values of the program elements are calculated.

**Table 1.** Information of test cases

Test cases	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$	$t_7$	$t_8$	$t_9$	$t_{10}$
$e_1$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$e_2$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$e_3$	✓	✓		✓	✓		✓	✓	✓	✓
$e_4$		✓							✓	
$e_5$	✓				✓		✓	✓		
$e_6$	✓			✓			✓	✓		✓
$e_7$			✓	✓		✓				✓
$e_8$			✓	✓		✓				✓
$e_9$			✓			✓				
$e_{10}$										
Status	P	P	P	F	P	P	F	F	F	F

**Table 2.** Group information of test cases

Group number	One	Two	Three	Four	Five
Test cases	$\{t_1, t_7, t_8\}$	$\{t_2, t_9\}$	$\{t_3, t_6\}$	$\{t_4, t_{10}\}$	$\{t_5\}$
Result status	P F F	P F	P P	F F	P
Element vector	5	4	5	6	4

### 3.4 Tarantula Method

Spectrum-based fault localization (SBFL) [6] technology is a fault localization technology based on coverage. Substitute the information of test cases into the suspiciousness value calculation formula, and the suspiciousness values indicate the possibility of elements with faults. Developers check program elements in non-decreasing order according to suspiciousness values until a fault is found. Tarantula [7], Ochiai [8] and Jaccard [9] are three commonly used fault localization calculation formulas based on spectrum. Here we choose Tarantula method, whose idea is that program elements covered by much more failed test cases and fewer passed test cases are more likely to include faults and assign higher suspiciousness values. The program element here can be a single statement or a method. The formula is as follows:

$$\text{Tarantula}(e) = \frac{\frac{n_{10}(e)}{n_{10}(e)+n_{00}(e)}}{\frac{n_{10}(e)}{n_{10}(e)+n_{00}(e)} + \frac{n_{11}(e)}{n_{11}(e)+n_{01}(e)}}$$

$P$  is the program under test consisting of element  $e$ ,  $e \in P$ .

$n_{11}(e)$  represents the number of all passed test cases covering element  $e$ ;

$n_{10}(e)$  represents the number of all failed test cases covering element  $e$ ;

$n_{01}(e)$  represents the number of all passed test cases not covering element  $e$ ;

$n_{00}(e)$  represents the number of all failed test cases not covering element  $e$ .

### 3.5 Algorithm

Input:

- (1) The program  $P=\{e_1, e_2, \dots, e_M\}$  contains  $M$  elements;
- (2) The test suite  $T=\{t_1, t_2, \dots, t_N\}$  includes  $N$  test cases;
- (3)  $CM=\{a_{ij}\}$  is a binary matrix that represents the coverage information of program elements by test cases.  $a_{ij}=1$  indicates that the test case  $i$  covers the program element  $j$ ; otherwise, it is 0.

Output: Program elements are ordered in non-increasing sequence of suspiciousness values.

Process:

- Step 1: Execute sufficient test cases including failed test cases and passed test cases;
- Step 2: Training with some executed test cases by the neural network, extracting features to build a model for predicting the test case results, and verifying it with the rest of the executed test cases;
- Step 3: Substitute the unexecuted test case into the neural network model to predict the running results;
- Step 4: Group unexecuted test cases according to element vectors and sort groups in non-decreasing order;
- Step 5: Sort failed test cases, select one failed test case from one group at each time. Select failed test cases from each group again after all groups have been selected one round until all failed test cases are selected. Passed test cases are selected in non-increasing order.
- Step 6: Alternate execution of failed test cases and passed test cases until termination conditions are met;
- Step 7: Substitute the running results into Tarantula formula and calculate the suspiciousness value of each program element. The program elements are ordered in the non-increasing sequence of suspiciousness values and assigned serial numbers. The serial numbers of elements with the same suspiciousness value are averaged.

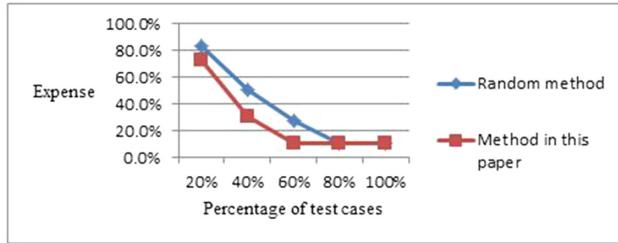
### 3.6 Evaluation Metric

We select *Expense* [10] to assess the validity of the fault localization method. Its value is the percentage of the program elements checked when the fault is found. Test cases are sorted according to the test case execution results predicted by the neural network model and the vector size covered by the test cases. Execute program according to the order of test cases, and substitute the execution information into Tarantula to calculate the suspiciousness values of program elements. Sort program elements in non-increasing sequence of suspiciousness values. When several program elements have the same suspiciousness value, the serial numbers of these program elements take the average value. *Expense* value is the ratio of the serial number of the program element with fault to the count of all program elements. The smaller the value, the less program elements need to be inspected when a fault is found, and the more effective the method is. The formula is as follows:

$$\text{Expense} = \frac{\text{rank of faulty element}}{\text{number of executable element}}$$

## 4 Experiment

Stereo matching is an vital research field in computer vision, which is extensively used in three-dimensional reconstruction, robot vision and other fields.

**Fig. 2.** Expense values of two methods**Table 3.** Comparison of test case execution and expense value

Percentage of test cases	Random method	Method in this paper
20%	83.3%	72.5%
40%	50.6%	31.3%
60%	27.8%	11.1%
80%	11.1%	11.1%
100%	11.1%	11.1%

The stereo matching process is divided into four steps: cost calculation, cost accumulation, disparity calculation and disparity refinement. This paper mainly tests the disparity calculation program. We manually transplant faults and one defective program contains one fault and generates five defective programs. The test suite contains 1,000 test cases, and the coverage information of the program by the test cases is known. Here we randomly select 100 test cases for each defective program to execute, substitute the execution information of 80 test cases into neural network for training, generate prediction model and verify it with the remaining 20 test cases. The average accuracy can reach 90%, and the more test cases involved in training, the higher the accuracy. The remaining unexecuted 900 test cases are substituted into the prediction model to forecast the execution results of test cases. The unexecuted test cases are ordered according to the method proposed in this article base on the prediction results and the coverage information, and substituted into the program to execute. Substitute the actual execution information of the test cases into Tarantula to calculate the suspicious value of each statement. Check the statements in the non-increasing sequence of the suspiciousness value. Perform five times for each defective program according to above steps and the average value *Expense* is calculated. The results are compared with the results of randomly executed test cases shown in Fig. 2 and Table 3. We discover that compared with the random method, the method proposed in this paper is more efficient and can find faults earlier when executing the same number of test cases.

## 5 Conclusion

When developers detect a program fault, they will switch to fault localization. Fault localization is a very complicated process. This paper proposes an effective fault localization prioritization method for stereo matching software. Construct a model through neural network based on the information of executed test cases to predict the running results of unexecuted test cases, and then sort the failed test cases and the passed test cases according to the coverage vector of the test cases. The test cases are executed according to this order, and terminate at any time according to execution requirements. The execution information is substituted into Tarantula to calculate the suspicious values of program elements. In stereo matching software, this method can locate faults faster than random method.

This method has certain requirements on test cases and needs sufficient test cases when construct prediction model. Passed test cases and failed test cases need to have some features. In the future, we can try whether this method can be used in other fields.

**Acknowledgment.** This work was supported by the Foundation of Heilongjiang Education Department (No.1354MSYYB003, No.1355ZD005), the Natural Science Foundation of Heilongjiang Province (No. F2018002).

## References

1. Kim, S., Baik, J.: An effective fault aware test case prioritization by incorporating a fault localization technique. In: Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, pp. 1–10. Association for Computing Machinery, New York (2010). <https://doi.org/10.1145/1852786.1852793>
2. Fu, W., Yu, H., Fan, G., Ji, X.: Test case prioritization approach to improving the effectiveness of fault localization. In: 2016 International Conference on Software Analysis, pp. 60–65. Testing and Evolution (SATE), Kunming (2016). <https://doi.org/10.1109/SATE.2016.17>
3. Zhang, X., Towey, D., Chen, T.Y., Zheng, Z., Cai, K.: Using partition information to prioritize test cases for fault localization. In: IEEE 39th Annual Computer Software and Applications Conference, Taichung, pp. 121–126. IEEE Press (2015). <https://doi.org/10.1109/COMPSAC.2015.35>
4. Zhang, X., Towey, D., Chen, T.Y., Zheng, Z., Cai, K.: A random and coverage-based approach for fault localization prioritization. In: Chinese Control and Decision Conference (CCDC), Yinchuan, pp. 3354–3361. IEEE Press (2016). <https://doi.org/10.1109/CCDC.2016.7531562>
5. Yoo, S., Harman, M., Clark, D.: Fault localization prioritization: comparing information-theoretic and coverage-based approaches. In: ACM Transactions on Software Engineering & Methodology, pp. 1–29. Association for Computing Machinery (2013). <https://doi.org/10.1145/2491509.2491513>
6. Jones, J.A., Harrold, M.J., Stasko, J.: Visualization of test information to assist fault localization. In: Proceedings of the 24th International Conference on Software Engineering, Orlando, pp. 467–477. IEEE Press (2002). <https://doi.org/10.1145/581396.581397>

7. Jones, J.A., Harrold, M.J.: Empirical evaluation of the tarantula automatic fault-localization technique. In: Proceedings of the 20th IEEE/ACM International Conference on Automated Software Engineering, pp. 273–282 (2005). <https://doi.org/10.1145/1101908.1101949>
8. Abreu, R., Zoeteweij, P., van Gemund, A.J.C.: An evaluation of similarity coefficients for software fault localization. In: 12th Pacific Rim International Symposium on Dependable Computing, Riverside, CA, pp. 39–46. IEEE Press (2006). <https://doi.org/10.1109/PRDC.2006.18>
9. Abreu, R., Zoeteweij, P., Golsteijn, R., van Gemund, A.J.C.: A practical evaluation of spectrum-based fault localization. *J. Syst. Softw.* **82**, 1780–1792 (2009). <https://doi.org/10.1016/j.jss.2009.06.035>
10. Jiang, B., Zhang, Z., Chan, W.K., Tse, T.H., Chen, T.Y.: How well does test case prioritization integrate with statistical fault localization. *Inf. Softw. Technol.* **54**, 739–758 (2012). <https://doi.org/10.1016/j.infsof.2012.01.006>



# Denoising of X-Ray Pulsar Signal Based on Variational Mode Decomposition

Yong Zhao<sup>1</sup>, Yingmin Jia<sup>1(✉)</sup>, and Qiang Chen<sup>1,2</sup>

<sup>1</sup> The Seventh Research Division and the Center for Information and Control, School of Automation Science and Electrical Engineering, Beihang University (BUAA), Beijing 100191, China

{zhaoyp1996,ymjia}@buaa.edu.cn

<sup>2</sup> China Academy of Space Technology, Beijing 100094, China

**Abstract.** X-ray pulsar navigation as a new type of autonomous navigation has very important application prospects and research value. The pulsar signal is a non-stationary signal with low signal-to-noise ratio, so it needs to be denoised. In this paper, a denoising method for X-ray pulsars based on variational mode decomposition is proposed. Simulation results show that the proposed method can improve the signal-to-noise ratio better than existing methods based on wavelet transform and empirical mode decomposition.

**Keywords:** Variational mode decomposition · X-ray pulsar · Denoising

## 1 Introduction

Pulsars are neutron stars that rotate at high speeds and emit multi-band electromagnetic waves. They have extremely stable periodicity, especially the rotation rate of the millisecond pulsar's rotation period reaches  $10^{-19}\text{--}10^{-21}\text{s/s}$ , known as the most stable astronomical clock in nature. X-rays are high-energy photons, which concentrate most of the pulsar's radiant energy. It is easy to miniaturize equipment detectors and processing, and can be detected by spacecraft flying outside the atmosphere, thus using this signal for autonomous navigation.

Sheikh [1] systematically demonstrated the principle of X-ray pulsar navigation (XPNAV) rationality and feasibility. In recent years, scholars have conducted extensive research on the key technologies and theories of XPNAV. The pulsar is very far away from the earth, the signal radiates from the pulsar and is attenuated by the propagation process. The X-ray pulsar signal received at the detector is already a single photon sequence. The X-rays of the cosmic background and other reasons make the detector receive pulsar X-ray signals with low signal noise, so it is difficult to directly use X-ray pulsar signals for navigation. Therefore, it is necessary to perform denoising processing on the X-ray pulsar signal. In addition, the denoising of the X-ray pulsar signal is helpful for the study of pulsars.

Wavelet Transformation was applied to pulsar signal denoising by [2]. Subsequently, many scholars denoised X-ray pulsar signals based on wavelet transform (WT) [3, 4]. The processing process of this kind of method is roughly as follows: firstly, perform wavelet decomposition on the X-ray pulsar signal which has been epoch folded, then perform threshold processing on the wavelet coefficients obtained by decomposition, and finally perform wavelet reconstruction to obtain the denoised signal. X-ray pulse signal is a typical non-stationary signal, and wavelet has good analysis ability for non-stationary signal. However, denoising based on WT has the problem of choice of wavelet basis function and poor adaptability because there are many wavelet decomposition and threshold processing methods. In addition, some scholars denoise pulsar signals based on empirical mode decomposition (EMD) [5, 6]. EMD makes up for the shortcomings of WT, has good adaptability, and can reflect the local frequency characteristics of the signal very well, but it lacks strict mathematical proof in theory, and there are problems such as endpoint effects and modal aliasing. Therefore, the decomposition result cannot effectively separate the useful signal from the noise, which affects the denoising effect.

Konstantin Dragomiretskiy et al. [7] proposed Variational Mode Decomposition (VMD), which is a new adaptive signal processing method and has obvious advantages for the processing of nonlinear and non-stationary signals. This method has high operation efficiency, can overcome the modal aliasing problem in EMD, achieve accurate separation of signals, and use its own Wiener filtering characteristics to obtain better noise filtering effects [8].

This paper proposes a new method of denoising for X-ray pulsar signals, and introduces the VMD into pulsar signal denoising. Compared with the existing denoising methods based on wavelet transform and empirical mode decomposition, the proposed method can achieve better denoising effects. It can be refined and improved on the basis of the proposed method, and this method can also be used for reference in dealing with other signal denoising problems.

## 2 X-Ray Pulsar Signal Preprocessing

Since the pulsar is far away from the earth, the X-ray pulsar signal received by the detector is the time series of photon arrival. It is shown as:

$$t_0 \leq t_1 \leq t_2 \leq \cdots \leq t_{N-1} \quad (1)$$

where  $t_i, i \in Z$  is the arrival time of the  $i$ th photon.

X-ray pulsar profiles are generally used for navigation and pulsar studies. The method proposed in this paper also deals with the pulsar profile. To obtain the pulsar profile from the X-ray pulsar photon sequence, three steps are generally required, namely converting the time of arrival (TOA) of photons to the Solar System Barycenter (SSB), pulsar period estimation, and epoch folding.

## 2.1 Converting the TOA of Photons to the SSB

Under the framework of general relativity, time can be divided into inherent time and coordinate time [9]. The X-ray pulsar photon arrival time is inherent time in the local coordinate system measured by the spacecraft. In order to obtain navigation observations and facilitate the study of pulsars, the inherent time must be converted to coordinate time under the SSB coordinate frame.

How to convert the TOA of photons to the SSB was introduced by [1] and [9]. HEASoft [10] software developed by the High Energy Astrophysics Science Archive Research Center (HEASARC) also provides the function of converting the TOA of photons to the SSB.

## 2.2 Pulsar Period Estimation

Pulsar period estimation is the problem of obtaining the pulsar period through the photon arrival time series. The pulse period is essential for epoch folding. Common pulse period estimation methods include chi-square period estimation method and Fourier analysis method. Chi-square period estimation method is also used by HEASoft. Its basic principle is: assuming that  $t_i'$  is a photon sequence converted to ssb, then according to a certain assumed period  $T_i$ , the absolute phase of the sequence relative to  $T_0$  is calculated and the remainder is obtained in  $[0,1]$ . Finally, the phase interval is divided into  $m$  equal bin blocks, and the number of photons  $N_i$  in each bin block interval is counted. The best estimate period can be calculated by

$$T = \arg \max_{T \in [T_{\min}, T_{\max}]} \chi^2 \quad (2)$$

$$\chi^2 = \sum_{i=1}^m \frac{(N_i - \bar{N})^2}{\bar{N}} \quad (3)$$

where  $\bar{N} = \frac{N}{m}$ ,  $[T_{\min}, T_{\max}]$  is the possible period range.

## 2.3 Epoch Folding

Epoch folding is currently the most basic profile acquisition algorithm. This method combines pulsar ephemeris data and reference epochs to calculate the phase of the arrival time of each photon relative to the reference epoch, and reduces it to each bin block within a phase period. Count the number of photons in each bin block to obtain the cumulative pulse profile.

It can be seen that when the pulsar flow is known, as the observation time becomes longer, the amount of observation data increases, the number of photons in each phase bin block increases, and the profile signal-to-noise ratio (SNR) will increase. However, too long observation time will affect the real-time and accuracy of XPNAV. In addition, it will cause a huge amount of observation data and face huge pressure on data storage and transmission. When the observation

time is short, the amount of observation data is small, and the SNR of pulse profile is low. The pulse contour with a low SNR will affect the accuracy of pulse TOA estimation, which in turn affects the accuracy of XPNAV. Therefore, it is necessary to reduce the noise of the pulsar folding profile.

### 3 Denoising Method for X-Ray Pulsars Signal Based on VMD

The purpose of pulsar signal denoising is to obtain a pulsar profile with a high SNR, which is used for pulsar navigation and research on pulsars. The pulsar profile obtained by folding the epoch is denoted as  $f$ .

$$f = x + d \quad (4)$$

where  $x$  is the useful pulsar profile signal, and  $d$  is noise.

Now we use VMD to denoise  $f$ . The VMD algorithm [7,8] determines the center frequency and bandwidth of each Intrinsic Mode Function by iteratively searching for the optimal solution of the variational model, and realizes the effective separation of the signal from low frequency to high frequency. The process is essentially the process of solving the variational problem.

First, construct a variational problem that assumes that each mode is a finite bandwidth with a center frequency, and seeks  $K$  modes (the sum of each mode is equal to the input signal  $f$ ) so that the sum of the estimated bandwidths of each mode is the smallest. The model is:

$$\begin{aligned} \min_{\{h_k\}, \{\omega_k\}} & \left\{ \sum_{k=1}^K \left\| \frac{\partial}{\partial t} \left[ (\delta(t) + j \frac{1}{\pi t}) * h_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ \text{s.t. } & \sum_{k=1}^K h_k = f \end{aligned} \quad (5)$$

where  $\{h_k\}, \{\omega_k\}$  respectively correspond to the decomposed  $K$  modal components and center frequency,  $\delta(t)$  is Dirac function,  $*$  is convolution operation.

Second, solve the variational problem of construction. The quadratic penalty factor can ensure the reconstruction accuracy of the signal in a noisy environment, and the Lagrange penalty operator can make the constraint conditions more strict. Therefore, by introducing the quadratic penalty factor and the Lagrange penalty operator  $\lambda(t)$ , the constrained variational problem described in Eq. (5) is transformed into an unconstrained variational problem. The augmented Lagrangian expression is obtained as follows:

$$\begin{aligned} \mathcal{L}(\{h_k\}, \{\omega_k\}, \lambda) = & \alpha \sum_{k=1}^K \left\| \frac{\partial}{\partial t} \left[ (\delta(t) + j \frac{1}{\pi t}) * h_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \\ & + \left\| f(t) - \sum_{k=1}^K h_k(t) \right\|_2^2 + \left\langle \lambda(t), f(t) - \sum_{k=1}^K h_k(t) \right\rangle \end{aligned} \quad (6)$$

Use the alternating direction multiplier method to further solve the variational problem, by iteratively updating  $h_k^{n+1}$ ,  $\omega_k^{n+1}$  and  $\lambda_k^{n+1}$  seeking to the “saddle

point" expressed by augmented Lagrange to obtain the optimal constraint variational model solution. The minimum value of  $h_k^{n+1}$  can be described as:

$$\begin{aligned} h_k^{n+1} = \arg \min_{h_k \in X} & \left\{ \alpha \left\| \frac{\partial}{\partial t} \left[ (\delta(t) + j \frac{1}{\pi t}) * h_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right. \\ & \left. + \left\| f(t) - \sum_i h_i(t) + \frac{\hat{\lambda}(t)}{2} \right\|_2^2 \right\} \end{aligned} \quad (7)$$

According to Parseval/Plancherel Fourier equidistant transformation, Eq. (7) can be converted to the frequency domain, and the minimum value problem can be solved:

$$\hat{h}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{h}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha (\omega - \omega_k)^2} \quad (8)$$

Using the same principle to solve the minimum value problem of  $\omega_k^{n+1}$ , the problem is converted to the frequency domain to solve, and the center frequency is obtained as:

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{u}_k(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k(\omega)|^2 d\omega} \quad (9)$$

The specific process of the VMD algorithm is as follows:

---

**Algorithm 1.** VMD algorithm

---

**Step 1** Initialize  $h_k^1$ ,  $\omega_k^1$ ,  $\hat{\lambda}^1$  and  $n$

**Step 2** Update  $h_k$  and  $\omega_k$  in frequency domain according to (8) and (9)

**Step 3** Keep updating  $\lambda$  according to  $\hat{\lambda}^{n+1}(\omega) = \hat{\lambda}^n(\omega) + \tau(\hat{f}(\omega) - \sum_k \hat{h}_k^{n+1}(\omega))$

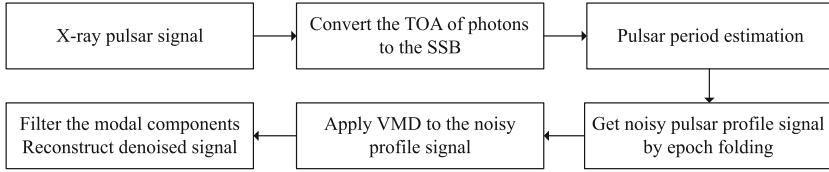
**Step 4** Stop iterating until  $\sum_{k=1}^K \frac{\|\hat{h}_k^{n+1} - \hat{h}_k^n\|_2^2}{\|\hat{h}_k^n\|} < e$ , otherwise return to **Step 2**.

---

After applying VMD to the noisy pulsar profile signal  $f$ , we can get  $K$  modal components  $h_1, h_2, \dots, h_K$  arranged in order from high frequency to low frequency, the dominant effect of noise on each component gradually decreases, and the dominant effect of signal on each component continues to strengthen. Different pulsars have different pulsar profiles. Critical modal components can be determined from historical data, thereby filtering out modal components that are mainly noisy. Subsequently, the signal is reconstructed to obtain the noise-reduced pulsar profile signal.

The process of denoising of X-ray pulsar signal based on VMD is shown by Fig.1.

We can use SNR, root mean square error (RMSE) and Pearson correction coefficient (PCC) to measure the noise reduction effect. Where  $x$  is noise-free signal,  $\hat{x}$  is denoised signal,  $M$  is the length of signal. Obviously, the largerer the SNR and the smaller the RMSE, the better the denoising effect. The PCC



**Fig. 1.** Process of Denoising of X-ray Pulsar Signal Based on VMD

is widely used to measure the degree of linear correlation between two variables, with values between  $-1$  and  $1$ . For the problems studied in this paper, the larger the PCC, the better the noise reduction effect.

$$SNR = 10 \log \frac{\sum_{i=1}^M x(i)^2}{\sum_{i=1}^M [x(i) - \hat{x}(i)]^2} \quad (10)$$

$$PCC = \frac{M \sum_{i=1}^M x(i)\hat{x}(i) - \sum_{i=1}^M x(i) \sum_{i=1}^M \hat{x}(i)}{\sqrt{M \sum_{i=1}^M x(i)^2 - \left(\sum_{i=1}^M x(i)\right)^2} \sqrt{M \sum_{i=1}^M \hat{x}(i)^2 - \left(\sum_{i=1}^M \hat{x}(i)\right)^2}} \quad (11)$$

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^M \frac{(x(i) - \hat{x}(i))^2}{x(i)^2}} \quad (12)$$

## 4 Simulation

NICER launched aboard a SpaceX Falcon 9 rocket on June 3, 2017 is an International Space Station payload devoted to the study of neutron stars through soft X-ray timing.

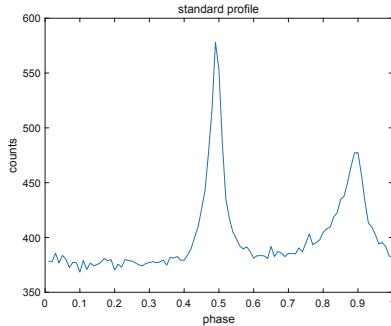
NICER is more advanced than the Rossi X-ray Timing Explorer (RXTE) whose data has been widely studied and used by scholars. This chapter uses the data of NICER to simulate and verify the proposed method.

In this paper, we compare the epoch folding, wavelet transform, empirical mode decomposition and the proposed method.

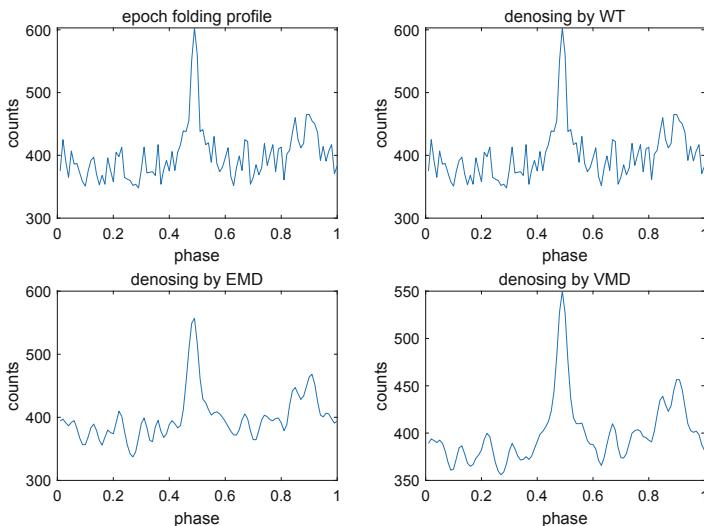
This article uses the data of the observation number 1011010301 ('cl' file) of NICER for simulation. This part of the data contains the arrival time of 1.68 million photons, and the observation time is nearly 100 min. Observation target is PSR\_B0531+21. Firstly, convert the TOA of photons to the SSB and estimate the pulse period. As mentioned above, as the observation time increases, the number of photons increases, and the SNR of the folded pulsar profile increases. So, in

this paper, the profile generated by folding all photon data is used as standard profile (noise-free signal) which is shown in Fig. 2, and the profile generated by folding 40,000 photon data are used as noise signals.

Figure 3 shows the simulation results of denoising based on epoch folding, wavelet transform, EMD and VMD. Table 1 shows the SNR, RMSE and PCC of denoised signal by different method. It can be seen that compared with other methods, the proposed method can better suppress noise, obtain the highest SNR and PCC, obtain the lowest RMSE, which illustrates the advantages of the proposed method.



**Fig. 2.** Standard profile.



**Fig. 3.** Denoising based on epoch folding, wavelet transform, EMD and VMD.

**Table 1.** SNR, RMSE & PCC of denoised signal by different method

Method	Epoch folding	WT	EMD	VMD
SNR	25.91	26.76	27.98	<b>30.35</b>
RMSE	0.0506	0.0457	0.0393	<b>0.0295</b>
PCC	0.875	0.890	0.909	<b>0.945</b>

## 5 Conclusions

In this paper, the problem of pulsar signal denoising is studied. An X-ray pulsar denoising method based on variational mode decomposition is proposed. Simulation results show that the proposed method can improve the signal-to-noise ratio of pulsar signals better than wavelet transform and empirical mode decomposition. On the basis of the method mentioned in this article, other methods can be combined to further improve the denoising effect. It should be studied further in the future.

**Acknowledgements.** This work was supported by the National Key Research and Development Plan (2017YFB0503300), the National Basic Research Program of China (973 Program: 2012CB821200, 2012CB821201) and the NSFC (61327807, 61521091, 61520106010, 61134005).

## References

- Sheikh, S.I.: The Use of Variable Celestial X-ray Sources for Spacecraft Navigation (2005). <http://hdl.handle.net/1903/2856>
- Zhu, X., Liao, F., Tang, Y.: Pulsar signal denoising based on wavelet transformation. *Acta Astronomica Sinica* **47**(3), 328–335 (2006). <https://doi.org/10.3321/j.issn:0001-5245.2006.03.011>
- Xiuping, L., et al.: X-ray pulsar signal de-noising for impulse noise using wavelet packet. *Aerosp. Sci. Tech.* **64**, 147–153 (2017). <https://doi.org/10.1016/j.ast.2017.01.024>
- Hu, H., Zhao, B., Sheng, L., et al.: Poisson noise removal for X-ray pulsar integrated pulse profile. *Acta Optica Sinica* **31**(8), 21–27 (2011). <https://doi.org/10.3788/AOS201131.0804002>
- Wang, L., Zhang, S., Lu, F.: Pulsar signal denoising method based on empirical mode decomposition and independent component analysis. In: Proceeding of the 2018 Chinese Automation Congress, pp 3218–3221 (2018). <https://doi.org/10.1109/CAC.2018.8623656>
- Zhao, P., Wang, W., Gong, B. et al.: Pulsar signal denosing based on empirical mode decomposition spatial correlation filter. *China Sci. Pap.* **10**(5), 592–596, 607 (2015). <https://doi.org/10.3969/j.issn.2095-2783.2015.05.021>
- Dragomiretskiy, K., Zosso, D.: Variational mode decomposition. *IEEE Trans. Signal Process.* **62**(3), 531–544 (2014). <https://doi.org/10.1109/TSP.2013.2288675>

8. Xu, F., Chang, J., Liu, B., et al.: De-noising method research for lidar echo signal based on variational mode decomposition. *Laser Infrared* **48**(11), 1443–1448 (2018)
9. Lirong, S.: Resaearch on Signal Processing Method of X-ray Pulsar-based Navigation. Xidian University (2017)
10. HEASARC. <https://heasarc.gsfc.nasa.gov/>



# Signal Estimation of Fatigue-Magnetic Properties of 25CrMo4 Based on Stein Algorithm

Zhenfa Bi<sup>(✉)</sup> and Guobao Yang

School of Railway Transportation, Shanghai Institute of Technology, Shanghai, China  
bizhenfa@sit.edu.cn, 2932129501@qq.com

**Abstract.** With the rapid development of high-speed railway, the safety detection and performance evaluation of train wheelsets have become the focus of technicians. In order to investigate this problem, a fatigue-magnetic performance estimation method based on Stein Algorithm for high-speed train wheelsets material 25CrMo4 is proposed. The characteristic value of the single dimensional magnetic memory signal is transformed into the characteristic value of magnetic memory signal in space domain by fitting the magnetic signal acquired by metal magnetic memory testing method and solving the partial and total derivatives of curved surface, therefore, the performance of fatigue specimens is evaluated. The results represent that as the fatigue cycle of the sample gradually raises from 3 million to 10 million, the average value of the tangential component of the magnetic signal increases to 273.76 A/m and the maximum increase in the maximum value is 80.22%, the mean value of the normal component increases to -19.50 A/m; after fitting, the gradient of the tangential component of the magnetic signal and the zero line of the normal component reflect that the stress concentration is near  $Y = 23$  mm. Any slice of the magnetic signal in the spatial domain can more completely and comprehensively reflect the fatigue state of the specimen, and can provide a basis for the fatigue-magnetic performance evaluation of the specimen.

**Keywords:** 25CrMo4 · Magnetic memory detection · Stein algorithm · Spatial fitting · Performance evaluation

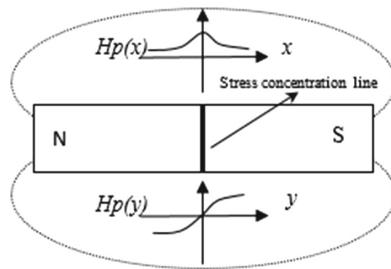
## 1 Introduction

The changing trend of rail vehicles towards high speeds and heavy loads puts forward higher requirements for the safe operation of trains, among which the safety detection and performance evaluation of train wheelsets are the core issues. The metal magnetic memory detection technology [1] is used to collect the magnetic memory signal of the plate of high-speed train wheelsets material 25CrMo4 under fatigue tension, the magnetic memory signals are spatially interpolated based on the Stein Algorithm and the same eigenvalue analysis method is used to evaluate the fatigue-magnetic performance of the plate. The analysis of the magnetic

memory signal distribution and its eigenvalues of the sample under fatigue can provide a reference for the online detection and life evaluation of high-speed train wheelsets.

## 2 Metal Magnetic Memory Detection Principle and Stein Estimation

In the geomagnetic environment, under the action of external force, the internal region of ferromagnetic components will rearrange irreversibly, and magnetic leakage field will be produced on its surface. In addition, the irreversible change of domain state will continue after the workload is eliminated [2,3]. As shown in Fig. 1, it is generally considered that the tangential component  $H_p(x)$  of the leakage magnetic field has a maximum value, and the normal component  $H_p(y)$  changes its sign and has a zero point [4,5]. By analyzing the distribution of the magnetic flux leakage signal on the surface of the test piece to reflect the change of its internal organization state.



**Fig. 1.** Principle diagram of magnetic memory detection

There is a basic requirement in the process of experimental (observation) data processing, that is, the data in the unknown space is fitted from the experimental (observation) data, which is generally used for regular prediction [6]. In actual research, due to the huge amount of spatial data, it is very difficult to obtain data in the entire experimental area. Generally, the spatial fitting method is used. The existing spatial interpolation methods include least squares fitting, ridge estimation, stein estimation, and spline fitting. Among them, the fitting effect of Stein estimation under multivariate is better than other fitting methods such as least square method [7,8]. Stein estimate is to compress the experimental data on the basis of the least square method, and then fit the compressed data.

For the model:  $Y = Xa + \varepsilon$ ,  $\hat{a}$  is its least squares estimate, then define  $\hat{a}(c) = c\hat{a}$  as the stein estimate of  $a$ .  $0 \leq c \leq 1$  is called the compression coefficient. In the data fitting, the basis for judging the validity of the stein estimate is that the existence of  $0 < c < 1$  makes the mean square error of  $\hat{a}(c)$  the smallest.

The mean square error of  $\hat{\mathbf{a}}(c)$  is shown in Eq. (1):

$$\begin{aligned}
 MSE(\hat{\mathbf{a}}(c)) &= \text{tr}Cov(\hat{\mathbf{a}}(c)) + \|E(\hat{\mathbf{a}}(c))\|^2 \\
 &= c^2\sigma^2\text{tr}(\mathbf{X}'\mathbf{X})^{-1} + (c-1)^2\|\mathbf{a}\|^2 \\
 &= c^2\sigma^2 \sum_{i=1}^n \lambda_i^{-1} + (c-1)^2\|\mathbf{a}\|^2 \\
 &= q(c)
 \end{aligned} \tag{1}$$

$q(c)$  derivative about  $c$ , and make  $q'(c) = 0$ , the best solution to  $c$  is:

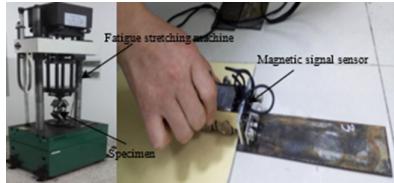
$$q'(c) = 0 \tag{2}$$

In the formula, at  $c^*$ ,  $q'(c) = MSE(\hat{\mathbf{a}}(c))$  reaches the minimum.

### 3 Metal Magnetic Memory Method

#### 3.1 Fatigue Tensile Test

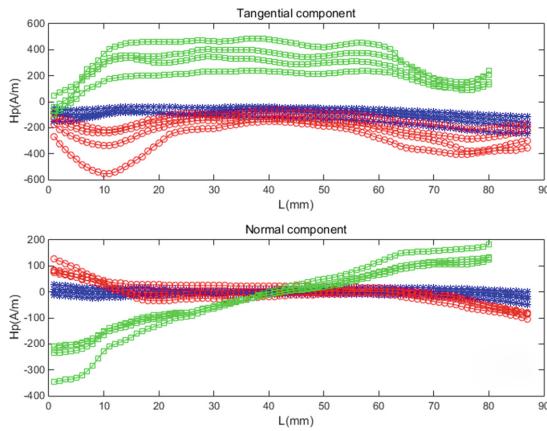
The specimens of the high speed train wheel-set material 25CrMo4 are divided into 3 groups, numbered as specimen 1, specimen 2 and specimen 3. Specimen 1 is a plate with 3 million fatigue times, specimen 2 is a plate with 5 million fatigue times, and specimen 3 is a plate with 10 million fatigue times. The fatigue test equipment and magnetic memory signal acquisition process are shown in Fig. 2.



**Fig. 2.** Fatigue stretching machine and magnetic signal acquisition

#### 3.2 Preliminary Processing of Magnetic Memory Signals

After performing tensile tests on the specimen at different fatigue times, a metal magnetic memory detector is used to collect magnetic memory signals along four parallel channels on the surface of the test piece. The raw data obtained is shown in Fig. 3.



**Fig. 3.** Raw data of magnetic memory signal

After sorting the data in the above figure, the numerical characteristics of the magnetic memory signals of the test piece under different fatigue life are obtained, as shown in Table 1.

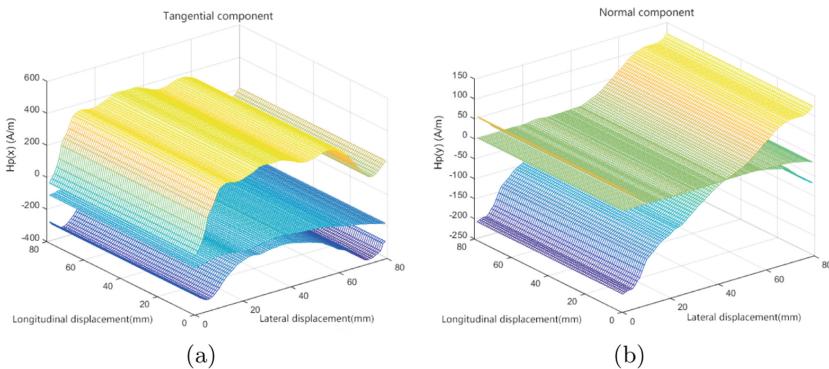
**Table 1.** Numerical characteristics of test pieces under different fatigue life

Component direction	Numbering	Minimum value (A/m)	Maximum (A/m)	Mean (A/m)
Tangential	Specimen 1	-245	-38	-100.27
	Specimen 2	-554	-66	-214.56
	Specimen 3	-104	487	273.76
Normal	Specimen 1	-49	28	-2.72
	Specimen 2	-105	126	-3.47
	Specimen 3	-346	182	-19.50

It can be seen from Table 1 that the tangential component of the magnetic memory signal changes obviously with the increase of the fatigue life, while the normal component changes relatively little. The floating range of the tangential component of the magnetic memory signal of specimen 1 is  $-245 \text{ A/m} \sim -38 \text{ A/m}$ , the average value is  $-100.27 \text{ A/m}$ , and the variation range of the magnetic signal of the normal component is  $-49 \text{ A/m} \sim 28 \text{ A/m}$ , the average value is  $-2.72 \text{ A/m}$ ; when the fatigue loading is adjusted to specimen 2, the numerical range of the tangential component of the magnetic memory signal changes greatly, and the range of variation is  $-554 \text{ A/m} \sim -68 \text{ A/m}$ , The mean value increases to  $-208 \text{ A/m}$ ; the normal component magnetic signal changes less, the floating range is  $-105 \text{ A/m} \sim 126 \text{ A/m}$ , and the mean value increases to  $-3.47 \text{ A/m}$ ; when the fatigue loading is adjusted to specimen 3, the change of the tangential component of the magnetic memory signal is more obvious, it

changes from  $-104 \text{ A/m}$  to  $487 \text{ A/m}$ , the average value increases to  $273.76 \text{ A/m}$ , and the change of the normal component magnetic signal is more Obviously, the value range is  $-346 \text{ A/m} \sim 182 \text{ A/m}$ , and the average value increases to  $-19.50 \text{ A/m}$ .

In order to study the distribution of magnetic memory signal on the surface of the specimen and judge the existence of stress limitation more accurately, it is necessary to fit and expand the collected signal in the transverse and longitudinal directions, and calculate and draw the distribution surface of the magnetic memory signal on the surface of the specimen under different life cycles, the results are shown in Fig. 4. It can be seen from the figure that the tangential component of the magnetic signal has a maximum value and a peak appears; the zero-crossing point of the normal component shows a zero-value line.



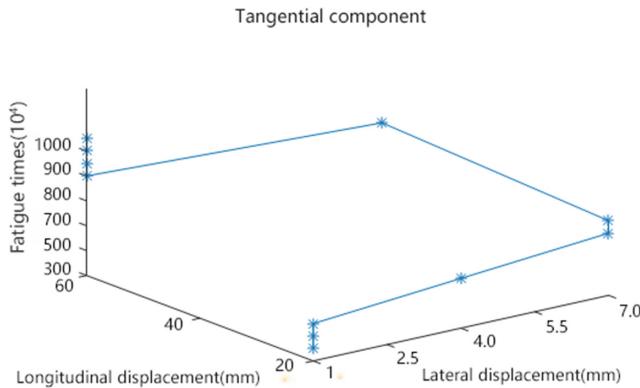
**Fig. 4.** The distribution of magnetic memory signals on the surface of the test piece

## 4 Features and Applications of Magnetic Memory Signals in the Spatial Domain

### 4.1 Extreme Value of Signal Tangential Component

From the principle of metal magnetic memory detection, it can be seen that the extreme point of the tangential component of the magnetic memory signal can be used to determine the location of the stress concentration of the specimen, but there may be errors in the extreme value analysis on the one-dimensional signal. Between the probes of the memory signal detection sensor is the undetected area, which can cause the loss of the true extreme value of the magnetic memory signal and cannot provide an accurate stress concentration location. The distribution of the tangential components of the magnetic memory signal on the surface of the fitted specimen under different life cycles is shown in Fig. 6.

The coordinates of the extreme point of the normal component of the magnetic memory signal under different life cycles in the figure are extracted and displayed in Table 2.



**Fig. 5.** The distribution of extreme point of tangential component of magnetic signal

**Table 2.** Signal extreme point distribution

Fatigue time ( $10^4$ )	Abscissa (mm)	Ordinate (m/m)	Signal value (A/m)
300	1.2	23	-86.11
400	1.2	23	-55.98
500	1.2	23	-55.55
600	4.6	25	-54.7
700	7	27	-53.51
800	7	27	-51.89
900	7	60	-49.91
1 000	7	60	-47.78

It can be seen from Table 2 that when the number of fatigue increases gradually from 3 million to 10 million, the extreme value of the tangential component of the magnetic signal near the coordinates  $X = 5$  mm and  $Y = 23$  mm gradually increases, the largest increase is 80.22%.

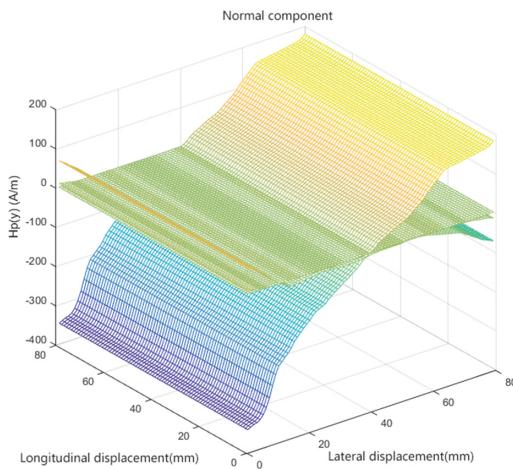
## 4.2 Zero Value of Normal Signal Component

In the analysis of one-dimensional magnetic signals, the distribution of normal components is characterized by the existence of zero points, and the collected signals do not fully reflect the distribution of the surface signals of the specimen. Calculate and plot the zero point of the magnetic signal normal component fitting the surface, the result is shown in Fig. 7. It can be seen that the zero points under different life cycles are connected into a line, which is the intersection of the signal surface and the  $H_p(y) = 0$  plane.

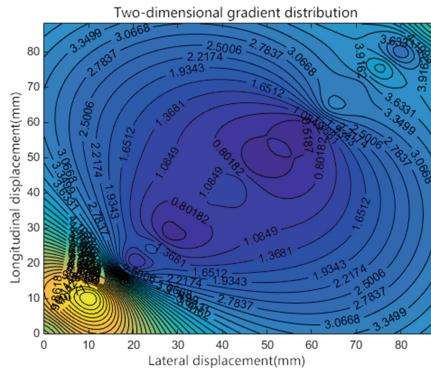
The zero line of the normal component of the magnetic memory signal and the extreme point of the tangential component of the magnetic memory signal can provide more accurate positioning for the stress concentration area of the specimen.

### 4.3 Magnetic Memory Signal Gradient Distribution

After fitting the magnetic memory signal of the test piece, the gradient value analysis of the one-dimensional magnetic memory signal can also be introduced into the space surface, that is, the partial derivative of the normal component of the magnetic memory signal in the Y direction is calculated, and the results are shown in 8.



**Fig. 6.** Distribution of zero line of signal normal component

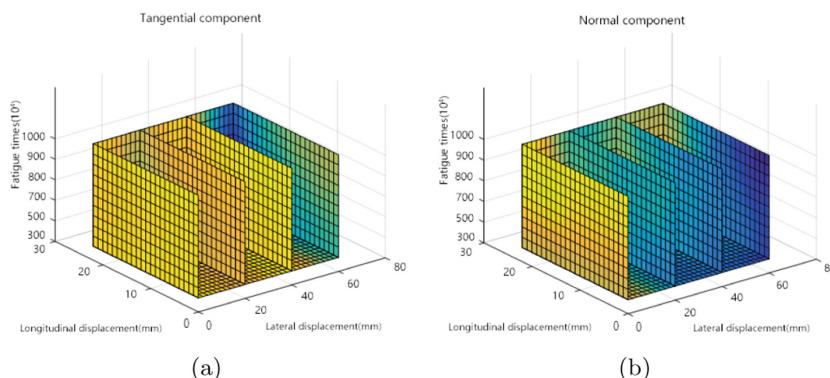


**Fig. 7.** Magnetic signal normal component gradient distribution

It can be seen from the figure that the two-dimensional gradient distribution of normal component of magnetic memory signal can reflect the violent change of magnetic signal more clearly than one-dimensional signal.

#### 4.4 Magnetic Signal Distribution Under Full Life Cycle

The magnetic memory signals on different components are further expanded to obtain the distribution of the magnetic signals under the full life cycle. The distribution of magnetic memory signals at any spatial position can be viewed, and the location of the magnetic memory signals can also be located according to the current distribution status of the magnetic signals. According to the distribution rule of the magnetic signal and the result of the signal fitting, the distribution diagram of the magnetic memory signal under the full life cycle as shown in Fig. 8.



**Fig. 8.** Magnetic signal distribution under full life cycle

The distribution state of the magnetic signal can be clearly seen from the figure, using this method, the distribution of the magnetic memory signal at a higher number of fatigue times can be predicted, thereby realizing the evaluation of the remaining life of the specimen.

### 5 Conclusion

Magnetic signal eigenvalue analysis can provide a multi-angle evaluation method for the fatigue performance evaluation of specimens. Therefore, the Stein estimation is used to interpolate and expand the magnetic signal, and the one-dimensional signal eigenvalue analysis is introduced into the three-dimensional space domain for more comprehensive and accurate regular conclusions.

- 1) When the fatigue time increases from 3 million to 10 million, the average value of tangential component of magnetic signal increases to 273.76 A/m, and the maximum increase is 80.22%. The signal concentration increased, and the stress concentration was near X = 5 mm and Y = 23 mm.
- 2) The partial derivative of the normal component of the magnetic memory signal after spatial fitting is calculated in the Y direction, and the corresponding gradient distribution map of the two-dimensional magnetic memory signal is made. Compared with the one-dimensional signal, the two-dimensional gradient distribution of the normal component of the magnetic memory signal can reflect the violent change of the magnetic signal more clearly.
- 3) After the spatial location of the magnetic signal is divided, the distribution map of the magnetic memory signal under the full life cycle of the specimen can be obtained, which can be used to detect the position where the specimen may develop into a crack and the remaining life of the specimen to evaluate.

**Acknowledgment.** This research work received financial support from the National Natural Science Foundation of China (51405303), the Special Fund for the Selection and Training of Excellent Young Teachers in Shanghai Universities (ZZyy15110), and the Development of Scientific and Technological Talents for Young and Middle-aged Teachers of Shanghai Institute of Technology (ZQ2019-21).

## References

1. Zhenfa, B., Le, K.: Research on early fault detection of high-speed train wheels based on metal magnetic memory. *J. Hebei Univ. Sci. Technol.* **39**(04), 306–313 (2018)
2. Cui, C., Dong, S., Wang, Y., et al.: Magnetic memory phenomenon of metal in stress concentrated parts under fatigue load. *Vibr. Mech. Eng. Mater.* **32**(12), 51–54 (2008)
3. Ren, J., Liu, H., Song, K.: The rise and development of metal magnetic memory detection technology. *Non-Destr. Test.* **38**(11), 7–15, 20 (2016)
4. Zhang, J., Zhu, S., Bi, Z., et al.: Early fault detection of high-speed wheelset based on metal magnetic memory. *J. Instrum.* **39**(1), 162–170 (2018)
5. Yin, D., Xu, B., Dong, S., et al.: Magnetic memory test of medium carbon steel fatigue test. *J. Mech. Eng.* **43**(3), 60–65 (2007)
6. Jian, W.: Stein estimation of linear regression model parameters and small sample estimation. *Jilin University* (2014)
7. Stein, C.: Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1**, 197–206 (1956)
8. Zhang, J., Wu, X., Liu, M.: Improvement of Stein estimation of linear regression model coefficients. *J. Naval Univ. Eng.* **16**(04), 22–25 (2004)



# Intelligent Frequency Selection of the Sky-Wave Radar Based on Numerical Ray Tracing

Runze Li<sup>(✉)</sup>, Jiangyun Wang, and Guanghong Gong

Beihang University, Beijing 100191, China  
zerunli1996@163.com

**Abstract.** The sky-wave radar uses the reflection effect of the ionosphere to achieve trans-horizon transmission, so its performance is directly affected by the ionospheric environment. In this paper, the latest ionospheric data model IRI2016 is used to generate electron density data, and a three-dimensional short-wave ray tracing algorithm based on Haselgrove equations is implemented. On this basis, an intelligent frequency selection model for fixed-point detection of the sky-wave radar under different ionospheric environments is established and the multipath phenomenon at different operating frequencies is also analyzed. MUF(Maximum Usable Frequency) calculated by the model is compared with the MUF obtained from the measured ionogram. The comparison shows the results calculated by the model are reasonable, which proves that the work done in this article can help the frequency selection of the sky-wave radar.

**Keywords:** Sky-wave radar · Ray tracing · MUF · Multipath phenomenon

## 1 Introduction

The ionospheric environment changes from time to time, and factors such as different regions, years, seasons, and hours greatly affect the state of the ionosphere, mainly manifested by its electron density distribution, which is a key parameter that affects the transmission of short-wave signals. If the sky-wave radar's operating frequency is too high, ionospheric penetration will occur, which means the ray penetrate the ionosphere and can't complete the detection mission on the ground. And if the work frequency is unsuitable, there will be multipath phenomenon, which also affect detection performance of the radar. So it is necessary to achieve intelligent frequency selection of the sky-wave radar.

Many ionospheric data models have been proposed. The ionosphere models which can be used in the real-time processing are the quasi-parabolic model described by Croft [1] and the multi-quasi-parabolic model described by Baker [2]. The two model describe the electron density distribution in the form of

analytical expression. But the accurate models depend on the analysis of the ionogram [3]. IRI model proposed by COSPAR (Committee on Space Research) and USRI (International Union of Radio Science) is the most widely used ionospheric data model. With the replacement of several versions, the accuracy of the model has gradually improved, and been trusted and used by more and more researchers [4]. The new version has many improvements such as building the new hmF2 model, using updated satellite measurements and improvement of the composition model in the topside ionosphere. Zhao X [5] compared the measured ionospheric data from Mohe, Wuhan, Beijing and Sanya ionosondes with the IRI calculation results. All comparison results indicate that IRI has a high accuracy.

Tracing the propagation trajectory of the signal can indicate the detection performance of the radar. The analytic ray tracing model can calculate the great circle path, which is the distance between the radar and target, and the group path, which is signal propagation distance. Chengyu H [6] used the multi-quasi-parabolic model to calculate the two parameters and studied the multipath phenomenon and the multimode phenomenon. But in this method the whole propagation trajectory of the radar signal is unable to get. Numerical ray tracing model does well in it. The numerical algorithm can reproduce the signal propagation between two points by solving the Haselgrove equations [10]. Point-to-point ray tracing technology can be used for fixed-point detection of sky wave-radar. After specifying the location of the radar and the target, the short-wave ray tracing algorithm is able to search out all the rays connecting the two points. Reilly [8] used Newton iteration algorithm to search rays, but it is only effective for the low elevation rays and the algorithm costs a lot of time in the calculation of partial derivatives. Strangeways [9] used a homing-in method which can find the high elevation rays, but the efficiency of the method is not good enough. Coleman [11] studied Fermat principle and proposed a variational method to solve the fixed-point boundary value problem, and the algorithm is relatively complicated.

After considering the above, this paper uses an optimized traversal algorithm to find out all the rays connecting the two points by giving the position of the radar and the target point as few operations as possible. Doing this work can analyze the point-to-point ray transmission status at a certain work frequency of the sky-wave radar. The ionospheric data is generated by IRI2016, which is the latest version of the International Reference Ionosphere model. In order to speed up the signal propagation simulation, the ionospheric data has been discretized and the data grid is established.

## 2 Numerical Ray Tracing Algorithm

Assume that the ionospheric electron concentration is evenly distributed in the horizontal direction, while ignoring the impact of the geomagnetic field and electron collision. The expression of the great circle path and the group path can be deduced from Martyn equivalent theorem and sine theorem, as shown in Eq. 1, where  $r_0$  is radius of earth,  $h'$  is reflection height and  $\beta_0$  is launch elevation angle.

$$\begin{aligned} d &= 2r_0 \left[ \arccos \left( \frac{r_0}{r_0 + h'} \cos \beta_0 \right) - \beta_0 \right] \\ p' &= \frac{2(r_0 + h')}{\cos \beta_0} \sin \left( \frac{d}{2r_0} \right) \end{aligned} \quad (1)$$

However, the actual short wave signal transmission trajectory will cross a certain latitude and longitude range, so that the propagation azimuth angle changes and the propagation trajectory is asymmetric. In order to study the propagation of short-wave signals and get azimuth information in the real ionosphere, this paper uses numerical ray tracing algorithm. By solving Haselgrove equations, the signal propagation conditions under different ionospheric conditions can be simulated and we can get the position of the wavefront at a certain time during the propagation process. And after the entire simulation process, we can get propagation parameters, for instance, the great circle path, the group path and the azimuth of transmitter and receiver. The theoretical signal propagation trajectory can be obtained under the accurate ionospheric model [7].

During the ray propagation, the wave front in spherical coordinate system which takes the earth as the origin is expressed as  $P(r, \theta, \phi)$ , where  $r$  is the distance to the center of the earth,  $\theta$  is the angle to the Arctic axis,  $\phi$  is the angle to the prime meridian. The wave vector is expressed as  $\mathbf{k}(k_r, k_\theta, k_\phi)$ . Haselgrove equations are as follows,

$$\begin{aligned} \frac{dr}{dP'} &= -\frac{1}{c} \frac{\partial H / \partial k_r}{\partial H / \partial \omega} \\ \frac{d\theta}{dP'} &= -\frac{1}{rc} \frac{\partial H / \partial k_\theta}{\partial H / \partial \omega} \\ \frac{d\phi}{dP'} &= -\frac{1}{rc \sin \theta} \frac{\partial H / \partial k_\phi}{\partial H / \partial \omega} \\ \frac{dk_r}{dP'} &= \frac{1}{c} \frac{\partial H / \partial r}{\partial H / \partial \omega} + k_\theta \frac{d\theta}{dP'} + k_\phi \sin \theta \frac{d\phi}{dP'} \\ \frac{dk_\theta}{dP'} &= \frac{1}{r} \left( \frac{1}{c} \frac{\partial H / \partial \theta}{\partial H / \partial \omega} - k_\theta \frac{dr}{dP'} + k_\theta \cos \theta \frac{d\phi}{dP'} \right) \\ \frac{dk_\phi}{dP'} &= \frac{1}{r \sin \theta} \left( \frac{1}{c} \frac{\partial H / \partial \phi}{\partial H / \partial \omega} - k_\phi \sin \theta \frac{dr}{dP'} - k_\phi \cos \theta \frac{d\phi}{dP'} \right) \end{aligned} \quad (2)$$

where  $H$  is Hamilton operator, the expression is

$$H = \frac{1}{2} \operatorname{Re} \left[ \frac{c^2}{\omega^2} (k_r^2 + k_\theta^2 + k_\phi^2) - n^2 \right] \quad (3)$$

where  $n$  is the refractive index of the medium, the expression is

$$\begin{aligned} n^2 &= 1 - 2X[1 - iZ - X/2(1 - iZ)(1 - iZ - X)] \\ &\quad - Y_T^2 \pm \sqrt{Y_T^4 + 4Y_L^4(1 - iZ - X)^2} \end{aligned} \quad (4)$$

where the + signs correspond to an ordinary ray and the - sign corresponds to the extraordinary ray.  $X = \frac{f_p^2}{f^2}$  relates to the plasma frequency  $f_p$  and wave frequency  $f$ ,  $Y = f_H/f$ , where  $f_H$  is the gyro frequency, and  $Y_L, Y_T$  are the components along the direction of the geomagnetic field and its normal.  $Z$  is related with plasma collision, this article ignores the impact of collisions which has little effect on the propagation trajectory.

The variables in Haselgrove equations are expressed as spherical coordinate system. We often use geographic coordinate system which is expressed as latitude, longitude and altitude to indicates the position of the point. So it is necessary to make coordinate transformation, as Eq. 5. If it is south latitude, *lat* needs add a minus sign. If it is west longitude, *lon* needs add a minus sign.

$$\begin{aligned} r &= alt + r_e \\ \theta &= \frac{\pi}{2} - lat \\ \varphi &= lon \end{aligned} \quad (5)$$

The Haselgrove equations is a set of partial differential equations, and the fourth-order R-K algorithm can be used to obtain an accurate numerical solution. The general form of the fourth-order R-K algorithm is expressed as Eq. 6.

$$\begin{aligned} y_{i+1} &= y_i + (K_1 + 2K_2 + 2K_3 + K_4) \frac{\Delta t}{6} \\ K_1 &= f(t_i, y_i) \\ K_2 &= f\left(t_i + \frac{\Delta t}{2}, y_i + \frac{\Delta t}{2} K_1\right) \\ K_3 &= f\left(t_i + \frac{\Delta t}{2}, y_i + \frac{\Delta t}{2} K_2\right) \\ K_4 &= f(t_i + \Delta t, y_i + \Delta t K_3) \end{aligned} \quad (6)$$

The partial derivative in Haselgrove equations is solved by Richard extrapolation. Take parameter *X* as an example, its expression shows as Eq. 7.

$$\begin{aligned} \frac{\partial X}{\partial r} &\approx \frac{4}{3} \frac{X(r+h/2, \theta, \varphi) - X(r-h/2, \theta, \varphi)}{h} - \frac{1}{6} \frac{X(r+h, \theta, \varphi) - X(r-h, \theta, \varphi)}{h} \\ \frac{\partial X}{\partial \theta} &\approx \frac{4}{3} \frac{X(r, \theta+h/2, \varphi) - X(r, \theta-h/2, \varphi)}{h} - \frac{1}{6} \frac{X(r, \theta+h, \varphi) - X(r, \theta-h, \varphi)}{h} \\ \frac{\partial X}{\partial \varphi} &\approx \frac{4}{3} \frac{X(r, \theta, \varphi+h/2) - X(r, \theta, \varphi-h/2)}{h} - \frac{1}{6} \frac{X(r, \theta, \varphi+h) - X(r, \theta, \varphi-h)}{h} \end{aligned} \quad (7)$$

And the initial value of **k** is calculated by Eq. 8, where  $\lambda$  is the wave length of the radar signal,  $\beta$  is the launch elevation angle,  $\alpha$  is the launch azimuth angle.

$$\begin{aligned} |k| &= \frac{2\pi}{\lambda} \\ k_r &= k \sin \beta \\ k_\theta &= k \cos \beta \cos \alpha \\ k_\varphi &= k \cos \beta \sin \alpha \end{aligned} \quad (8)$$

### 3 Electronic Density Data from IRI2016

The input of the numerical ray tracing algorithm contains electronic density of the wave front during simulation. The electronic density data is generated by IRI2016. Enter the date, time, latitude and longitude, IRI2016 can generate a certain altitude range of electron density distribution.

In order to ensure that the data generated by IRI2016 has certain credibility, the generated data is compared with the measured data. The measured data is from National Earth System Science Data Center. The comparison parameter is foF2 in Beijing (39°N, 116°E), which means the characteristic frequency of F2 layer in the ionosphere. The results are as Table 1 (The date format is “month-day-year”).

In the process of numerical integration calculation, if IRI2016 model is called continuously, it will inevitably increase the calculation time. And the difference in electron density between adjacent steps is small, there is no need for each step to run IRI2016 model again. Therefore, the electron density data is discretized with the form of data grid. Every grid has the same density. According to the experiment, it is found that the grid size is set to  $2^\circ \times 2^\circ \times 1\text{ km}$  (longitude, latitude and altitude) to meet the accuracy of ray tracing. If the resolution is further increased, the adjacent data grid has almost no error, but the calculation time increases a lot.

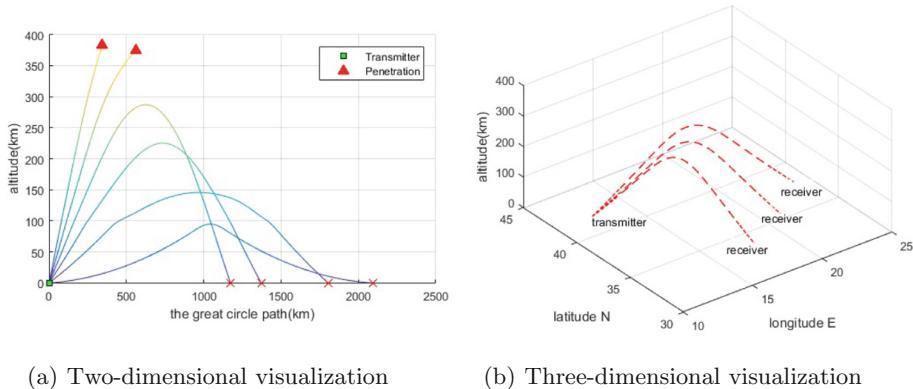
## 4 Analysis of the Sky-Wave Radar Working Frequency

The numerical ray tracing algorithm is implemented through programming. The schematic diagram of the results of the program is Fig. 1. Figure 1(a) is two-dimensional results, showing several signal propagation trajectory of different launch elevation angles. In the figure, there are two higher frequency signals penetrate out of the ionosphere. Figure 1(b) is three-dimensional results, showing several signal propagation trajectory of different launch azimuth angles.

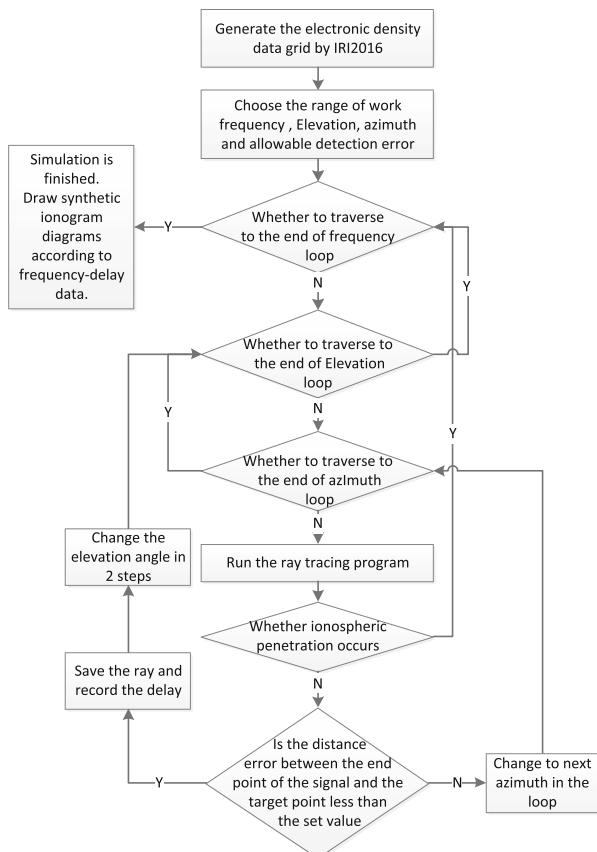
**Table 1.** Comparison of foF2 between generated data by IRI and measured data(MHz)

Time (UT)	Generated data	Measured data	Error
01-01-2011T0:00	4.51	4.90	0.39
01-01-2011T6:00	6.76	6.83	0.07
01-01-2011T12:00	2.72	2.80	0.08
01-01-2011T18:00	3.22	2.92	0.30
07-01-2011T0:00	6.70	6.30	0.40
07-01-2011T6:00	6.70	7.78	1.08
07-01-2011T12:00	6.94	7.40	0.46
07-01-2011T18:00	5.57	5.80	0.23

The flow chart for calculating MUF is as Fig. 2. Compared to the simple three-layer loop traversal algorithm (loop for frequency, elevation and azimuth), two main improvements are proposed: one is abandoning the ray search of different azimuths at the same frequency and elevation angle. After getting a ray that meets the requirement, it exits the azimuth cycle, which can reduce the number of operations. Because at the same frequency and elevation angle different azimuth rays arriving at the same destination have almost the same delay. And there is a certain error in the judgment of the end point. So it is not meaningful to analyze rays of different azimuth launch angle separately. The other is that if the ray that meets the requirement is calculated, the next iteration is a step size of 2 times in the cycle of traversing the elevation angle. This also avoids



**Fig. 1.** Visualization of signal propagation trajectory



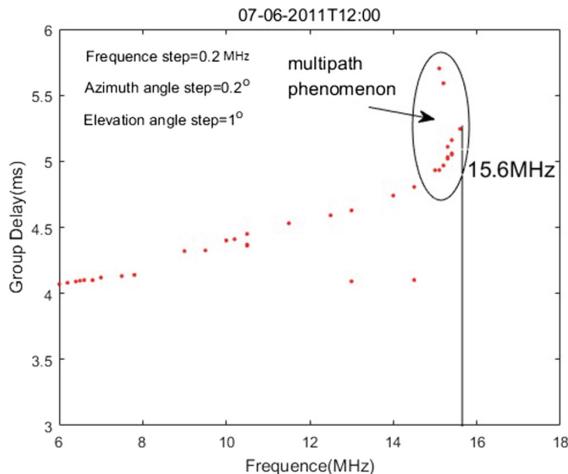
**Fig. 2.** The flow chart for calculating MUF

analyzing and recording multiple sets of rays with a very small delay difference. Through experimental analysis, it is reasonable that the step in frequency is 0.2 MHz, in elevation is  $1^\circ$  and in azimuth is  $0.2^\circ$ . Allowable detection error is 0.1 latitude and 0.1 longitude.

## 5 Results

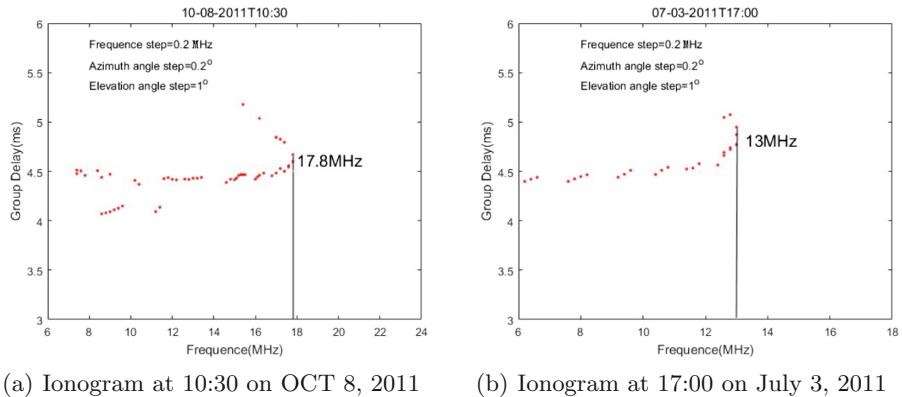
The ionospheric electron density grid used in the experiment is generated by IRI2016. Taking the ionosphere at 12:00 on July 6, 2011 as an example, the maximum usable frequency of sky-wave radar and multipath phenomenon in this ionosphere are analyzed. The synthetic ionogram of it shows as Fig. 3. And the two other experiment results show as Fig. 4. The synthetic ionogram shows the relationship between signal frequency and echo delay.

Multipath propagation refers to the phenomenon that a target's backscattered electromagnetic wave reaches the receiver through different paths under different time delay conditions, and multiple targets appear at the receiving end of the radar. The multipath propagation of radio waves will make target identification and positioning difficult. For the frequency band where multipath transmission occurs, the frequency resolution in the loop can be increased. As shown in the Fig. 3, the points in the multipath area are denser. And Fig. 5 is the schematic diagram of multipath phenomenon.

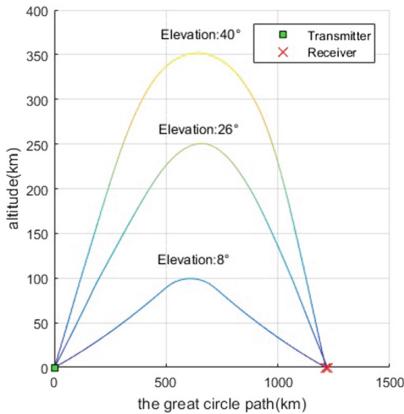


**Fig. 3.** The synthetic ionogram at 12:00 on July 6, 2011

To test the effectiveness of the algorithm, this article compares the MUF value obtained by the algorithm with the measured MUF value on a radio link which is between Rome( $41.8^\circ\text{N}$ ,  $12.5^\circ\text{E}$ ) in Italy and Chania ( $35.7^\circ\text{N}$ ,  $24.0^\circ\text{E}$ ) in Greece [12]. The transmitting system is based on a VOS-1 chirp ionosonde,



(a) Ionogram at 10:30 on OCT 8, 2011      (b) Ionogram at 17:00 on July 3, 2011

**Fig. 4.** Two other synthetic ionogram**Fig. 5.** The multipath phenomenon**Table 2.** Comparison of MUF between model calculation and measured data (MHz)

Time (UT)	Calculated by model	Measured data	Error
07-03-2011T17:00	13.0	15.1	2.1
07-06-2011T12:00	15.6	13.7	1.9
10-08-2011T10:00	17.8	20.5	2.7

sweeping from 2 to 30 MHz at 100 kHz/s with an average power of less than 10 W. The comparison results are shown in the Table 2. The results show that the model of the sky-wave radar intelligent frequency selection has a certain degree of credibility.

## 6 Conclusions

From the comparison between IRI2016 model and measured data, IRI2016 has a certain degree of credibility. The ionosphere data generated by IRI2016 provides environmental factors to choose the suitable frequency of the sky wave radar. Under the premise that the IRI model itself has certain errors, MUF predicted in this paper based on numerical ray tracing is not much different from the actual measured data. It can help the frequency selection of the sky wave radar and reduce the detection error caused by multipath transmission.

## References

1. Croft, T.A., Hoogasian, H.: Exact ray calculations in a quasi-parabolic ionosphere with no magnetic field. *Radio Sci.* **3**(1), 69–74 (1968). <https://doi.org/10.1002/rds19683169>
2. Baker, D.C., Lambert, S.: Range estimation for SSL HFDF systems by means of a multiquasiparabolic ionospheric model. *Microwaves Antennas Propag. IET Proc. H* **136**(2), 120–125 (1989). <https://doi.org/10.1049/ip-h-2.1989.0022>
3. Chengyu, H., Yongzhen, L., Wenjing, H.: Analysis of the ionospheric time-varying effects on the radar echoes based on ionogram inversion. *Neurocomputing* **174**, 966–973 (2015). <https://doi.org/10.1016/j.neucom.2015.09.069>. S0925231215014022
4. Bilitza, D., Altadill, D., Truhlik, V., et al.: International reference ionosphere 2016: from ionospheric climate to real-time weather predictions. *Space Weather* **15**(2), 418–429 (2017). <https://doi.org/10.1002/2016SW001593>
5. Zhao, X., Ning, B., Zhang, M., et al.: Comparison of the ionospheric F2 peak height between ionosonde measurements and IRI2016 predictions over China. *Adv. Space Res.* **60**(7), 1524–1531 (2017). <https://doi.org/10.1016/j.asr.2017.06.056>
6. Chengyu, H., Guo, K., Yili, F.: The sky-wave radar detection performance computing based on the dynamic ionospheric model. *Neurocomputing* **151**(mar.3pt.3), 1305–1315 (2015). <https://doi.org/10.1016/j.neucom.2014.10.073>
7. Kashcheyev, A., Nava, B., Radicella, S.M.: Estimation of higher-order ionospheric errors in GNSS positioning using a realistic 3-D electron density model. *Radio Sci.* **47**(4), 1–7 (2012). <https://doi.org/10.1029/2011RS004976>
8. Reilly, M.H.: Upgrades for efficient three-dimensional ionospheric ray tracing: investigation of HF near vertical incidence sky wave effects. *Radio Sci.* **26**(4), 971–980 (1991). <https://doi.org/10.1029/91RS00582>
9. Strangeways, H.J.: Effect of horizontal gradients on ionospherically reflected or transionospheric paths using a precise homing-in method. *J. Atmos. Solar Terr. Phys.* **62**(15), 1361–1376 (2000). [https://doi.org/10.1016/s1364-6826\(00\)00150-4](https://doi.org/10.1016/s1364-6826(00)00150-4)
10. Haselgrove, J.: Ray theory and new method for ray-tracing. In: Report of the Physical Society Conference. Physics in the Ionosphere, pp. 355–360. The Physical Society, London (1955)
11. Coleman, C.J.: Point-to-point ionospheric ray tracing by a direct variational method. *Radio Sci.* **46** (2011). <https://doi.org/10.1029/2011RS004748>
12. Settimi, A., Pezzopane, M., Pietrella, M., et al.: Testing the IONORT-ISP system: a comparison between synthesized and measured oblique ionograms. *Radio Sci.* **48**(2), 167–179 (2013). <https://doi.org/10.1002/rds.20018>



# Design of Multi-port Energy Conversion System of Electric Vehicle Based on Bridge-Type Buck-Boost Topology

Yunhao Zhang, Xiaonan Xia, Xiaoxing Ge, Wei Tang, and Yu Fang<sup>(✉)</sup>

Institute of Information Engineering, Yangzhou University, Yangzhou 225100, China  
yzfangyu@126.com

**Abstract.** Aiming at the improvement of electric vehicle endurance and its role in the future smart grid construction, this paper proposes a multi-port energy conversion system of electric vehicle based on bridge-type Buck-Boost topology. The energy conversion system consists of an on-board photovoltaic charging circuit, an energy conversion circuit implementing energy exchange between electric vehicles, and an energy conversion circuit implementing energy exchange between grid and electric vehicle. Bridge-type Buck-Boost circuit is adopted between each port, so as to realize the energy conversion of electric vehicle. The feasibility and superiority of the system are verified by experiments.

**Keywords:** Energy conversion system · On-board photovoltaic charging circuit · Electric vehicle

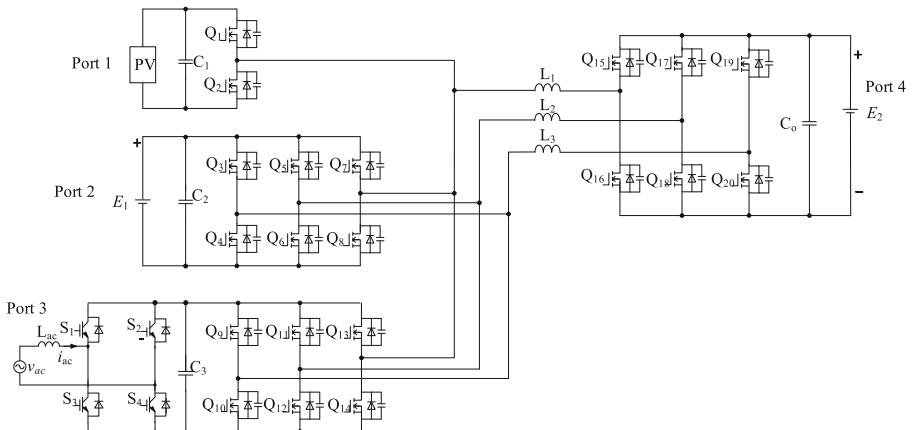
## 1 Introduction

In response to increasingly serious energy shortage and environmental problems, the fuel automotive industry in all countries around the world is accelerating the transition to a new energy vehicle strategy [1,2]. As a clean vehicle, photovoltaic electric vehicles rely on the advantages of low energy consumption, zero pollution, low noise and so on, and the market demand is becoming more and more extensive [3,4]. The use of solar energy in on-board photovoltaic power generation system can improve the economic benefits of electric vehicles. There are currently three major circuit topologies for bidirectional chargers: isolated dual-active bridge circuit [5], isolated bidirectional half-bridge circuit [6], and non-isolated bridge-type Buck-Boost circuit [7]. The non-isolated bridge-type Buck-Boost circuit has a relatively small number of switch devices, simple control and high efficiency. So the non-isolated bridge-type Buck-Boost circuit is the bidirectional charger topology selected in this paper. The bidirectional AC/DC inverter has two major functions [8]. One is to realize the DC bus voltage inversion to the grid, and the other is to rectify the grid to the DC bus side to store energy for the battery. This paper uses a single-phase bidirectional full-bridge AC/DC circuit as the main circuit of a bidirectional inverter. In order to realize

the fast, convenient and reliable charging for power batteries of electric vehicle, this paper proposes a multi-port energy conversion system of electric vehicle based on bridge-type Buck-Boost topology.

## 2 Multi-port Energy Conversion System Based on Bridge-Type Buck-Boost Topology

The multi-port electric vehicle conversion circuit mainly consists of three subsystems, as shown in Fig. 1. The energy generated by the photovoltaic modules is charged by the H4-bridge Buck-Boost circuit for the local battery  $E_2$  of the electric vehicle, and its energy flow is one-way. Among them, Port 1 and Port 4 are respectively connected to the photovoltaic module and the local battery of the electric vehicle. The main function of the double half-bridge circuit is to track the maximum power point of the electric vehicle during charging, so as to realize the fast and high-efficiency charging of the electric vehicle; the energy conversion system between external vehicle battery and local vehicle battery is composed of three bridge-type circuits. Among them, Port 2 and Port 4 are respectively connected to external vehicle battery and local vehicle battery. Three parallel bridge-type Buck-Boost circuits act as a bidirectional flow channel of energy, realizing bidirectional energy conversion between batteries, and achieving efficient use of battery power; the energy conversion system between grid and the local battery of the electric vehicle is composed of a bidirectional AC/DC converter and three parallel bridge-type Buck-Boost circuits. Among them, Port 3 and Port 4 are respectively connected to the grid and the local battery. The AC/DC converter and three parallel bridge-type Buck-Boost circuits act as a bidirectional energy flow channel to realize the mutual flow of energy between grid and local battery.

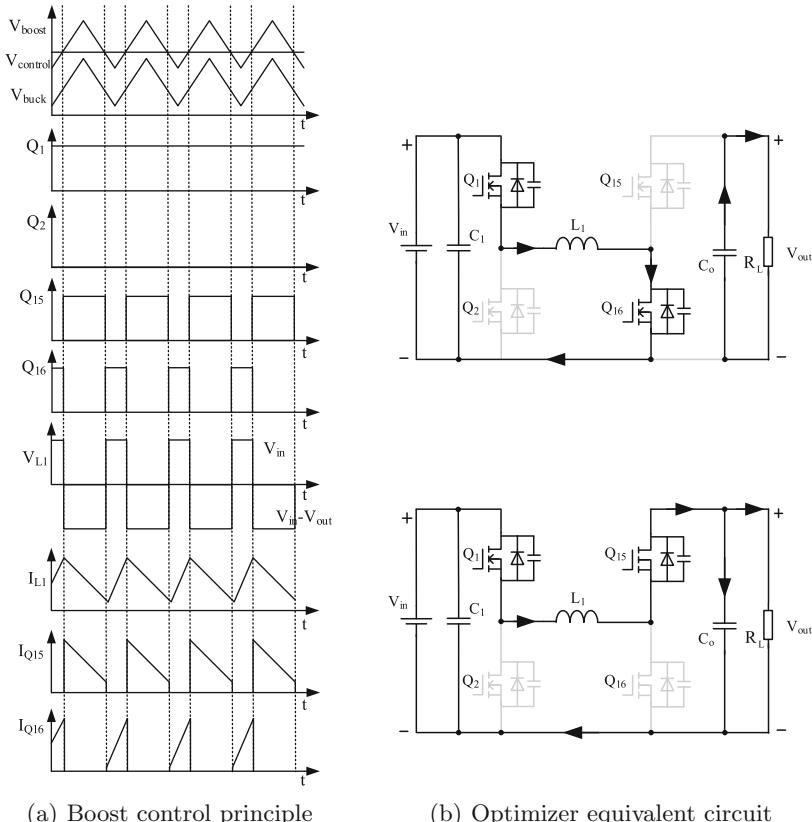


**Fig. 1.** Multi-port main circuit implementing electric vehicle energy conversion

## 2.1 On-board Photovoltaic Charging Circuit

The circuit mainly realizes the energy generated by the photovoltaic module to charge local battery of electric vehicle. Its topology is made of a H4-bridge Buck-Boost circuit, which enables it to operate in Buck-Boost operating mode to convert direct-current voltage. Controllable switch tube  $Q_1, Q_{16}$  are respectively buck, boost major switch tube, and their duty ratio are  $D_1, D_2$ .  $Q_2$  and  $Q_{15}$  are synchronous rectification tube.  $Q_1$  and  $Q_2$  are a set of switch tubes with complementary conduction of control signals, forming buck unit;  $Q_{15}$  and  $Q_{16}$  are a set of switch tubes with complementary conduction of control signals, forming boost unit. Input and output voltage satisfy relationship:  $V_{out}/V_{in} = D_1/(1 - D_2)$ .

In this paper,  $V_{in}$  and  $V_{out}$  are numerically sampled and analyzed by the main chip, and the operating mode of circuit is judged. There are two triangular carrier boost-class triangular carrier  $V_{Boost}$  and buck-class triangular carrier  $V_{Buck}$ . The control voltage  $V_{control}$  is cross-compared with the triangular carrier  $V_{Boost}$  and



**Fig. 2.** Boost operating mode for photovoltaic optimizer

$V_{Buck}$ , and then gets the duty ratio for the boost circuit and the buck circuit. The specific operating principle is as follows:

As shown in Fig. 2(a), when  $V_{in}$  is within  $[V_{in-min}, V_o - \Delta V]$ , the  $V_{control}$  intersects  $V_{boost}$ . The circuit enters boost mode and the duty ratio of the boost circuit is  $D_2$ . The equivalent circuit shown in Fig. 2(b), at which point  $Q_1$  is closed,  $Q_2$  is normally on, and  $Q_{15}$  and  $Q_{16}$  are alternately on.

If  $V_{in}$  rises to the point within  $[V_o - \Delta V, V_o + \Delta V]$ ,  $V_{control}$  is in the overlapping area of  $V_{boost}$  and  $V_{buck}$ . The circuit is in a straight-through state. At this time,  $Q_1$  and  $Q_{15}$  are closed,  $Q_2$  and  $Q_{16}$  are normally on. In this mode,  $V_{out} = V_{in}$ .

If  $V_{in}$  rises to the point within  $[V_o + \Delta V, V_{in-max}]$ ,  $V_{control}$  intersects  $V_{buck}$ . The circuit is in a buck state and has a duty ratio of  $D_1$ . At this time,  $Q_{15}$  is closed,  $Q_{16}$  is normally on, and  $Q_1$  and  $Q_2$  are alternately on.

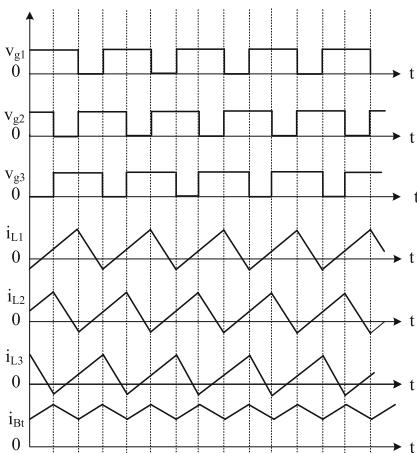
For the large range of output voltages generated by photovoltaic modules in different operating environments, the wide range of voltage access can be achieved by bridge-type Buck-Boost topology to the local battery charging circuit of electric vehicles. In this way, the electric vehicle battery can be placed in a relatively stable charging environment, protecting the battery and extending the battery life.

## 2.2 Energy Conversion Subsystem Between External Vehicle Battery and Local Vehicle Battery

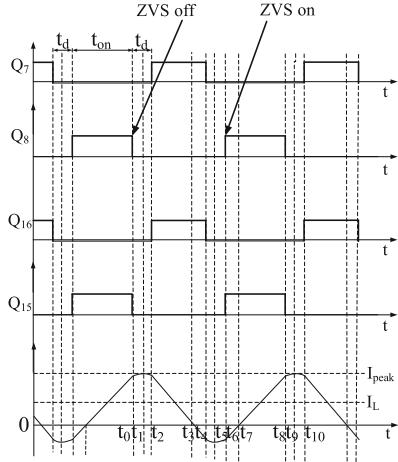
The energy conversion subsystem between external vehicle battery and local vehicle battery adopts three bridge-type Buck-Boost circuits with parallel connection structure, which can realize Buck-Boost operating mode. The three bridge-type Buck-Boost circuits have the same principle. For convenience, one of the bridge Buck-Boost circuits can be analyzed separately.

Figure 3 shows the waveform of the bidirectional DC/DC converter operating in boost mode as an example. The waveform  $v_{g1}$ ,  $v_{g2}$ ,  $v_{g3}$  are the driving signals of  $Q_8$  and  $Q_{15}$ ,  $Q_6$  and  $Q_{17}$ ,  $Q_4$  and  $Q_{19}$  respectively, and the driving signals are  $120^\circ$  phased to each other.  $i_{L1}$ ,  $i_{L2}$ ,  $i_{L3}$  are the current of the energy storage inductor  $L_1$ ,  $L_2$ ,  $L_3$ , and  $i_{Bt}$  is the input current of the battery. Superposition current value of three inductors is always positive, which can eliminate the problem of zero inductance current in each bridge-type Buck-Boost circuit, and reduce the ripple of inductance current. The current at the input of the battery has very little ripple. It can be seen that this circuit can not only realize soft switch, but also solve the problem of large current ripple.

When configuring the drive signals of the switch tubes, the driving signals in the same bridge arms such as  $Q_7$  and  $Q_8$ ,  $Q_{15}$  and  $Q_{16}$  are complementary, but a certain dead time will be left to prevent the switch tubes of the same bridge arm from passing through. The driving signals of the switch tubes in different bridge arms such as  $Q_7$  and  $Q_{16}$ ,  $Q_8$  and  $Q_{15}$  are the same. Figure 4 shows the waveforms of the main components of the circuit in boost mode. Figure 5 shows the eight operating states of the switch at different times during the same cycle in boost mode. The specific situation is analyzed as follows:



**Fig. 3.** Current waveform for three-phase interlaced parallel circuits



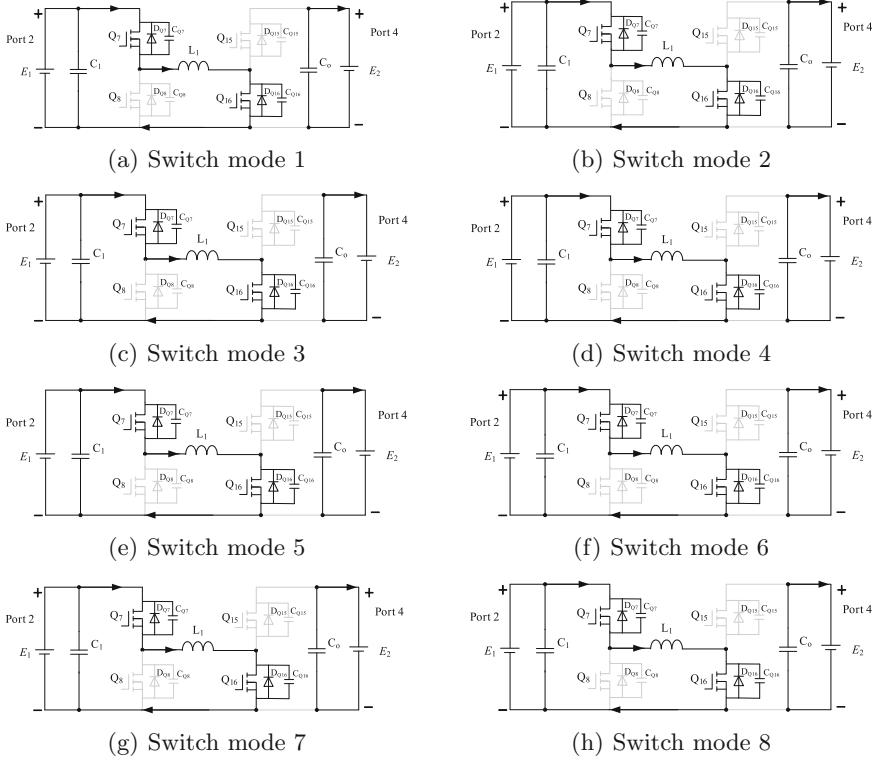
**Fig. 4.** Waveforms in boost mode

Figure 5(a) shows before the moment  $t_0$ , the energy storage inductance  $L_1$  is stored. Figure 5(b) shows, within  $[t_0, t_1]$ , capacitors  $C_{Q8}$  and  $C_{Q15}$  charge,  $C_{Q7}$  and  $C_{Q16}$  discharge. Figure 5(c) shows, within  $[t_1, t_2]$ ,  $C_{Q8}$ ,  $C_{Q15}$  are fully charged,  $C_{Q7}$ ,  $C_{Q16}$  are fully discharged, and  $i_{L1}$  begins to decrease. Figure 5(d) shows, within  $[t_2, t_3]$ , the current value  $i_{L1}$  continues to decrease to zero. Figure 5(e) shows, within  $[t_3, t_4]$ , when the current is reversed, diodes  $D_{Q8}$  and  $D_{Q15}$  cut-off, energy storage inductance  $L_1$  begins to store energy. Figure 5(f) shows, within  $[t_4, t_5]$ , the reverse current of the inductance charges  $C_{Q7}$  and  $C_{Q16}$ .  $C_{Q8}$  and  $C_{Q15}$  discharge. Figure 5(g) shows, within  $[t_5, t_6]$ ,  $D_{Q8}$  and  $D_{Q15}$  are on under the action of inductance current. The reverse current in the inductance continues to decrease. Figure 5(h) shows, within  $[t_6, t_7]$ , the voltage sitting at both ends of capacitors  $C_{Q8}$ ,  $C_{Q15}$  is zero, and the switching tubes  $Q_8$  and  $Q_{15}$  achieve ZVS. The reverse current in the inductance continuously decreases to zero, and  $D_{Q7}$  and  $D_{Q16}$  are naturally shut off. From the above analysis, it can be seen that each diode in the circuit can be naturally turned on and off, so there will be no problems of circuit oscillation and circuit loss due to reverse recovery of the diode. Each switch in the circuit can achieve soft switching operation.

## 2.3 Energy Conversion Subsystem Between Grid and Local Vehicle Battery

Considering the high efficiency and high power density of PCS, this paper selects the single-polarity SPWM modulation mode [9].

Whether operating in rectifier mode or inverter mode, there is a pair of IGBT tubes operating in high-frequency complementary state, a pair operating in the frequency state, effectively reducing the switching loss of the switching tube,



**Fig. 5.** Operating mode of the bidirectional DC/DC converter circuit

thereby improving the conversion efficiency of AC/DC. The DC/DC converter between grid and local vehicle battery is similar to the principle of energy conversion subsystem between external vehicle battery and local vehicle battery, which is not described here.

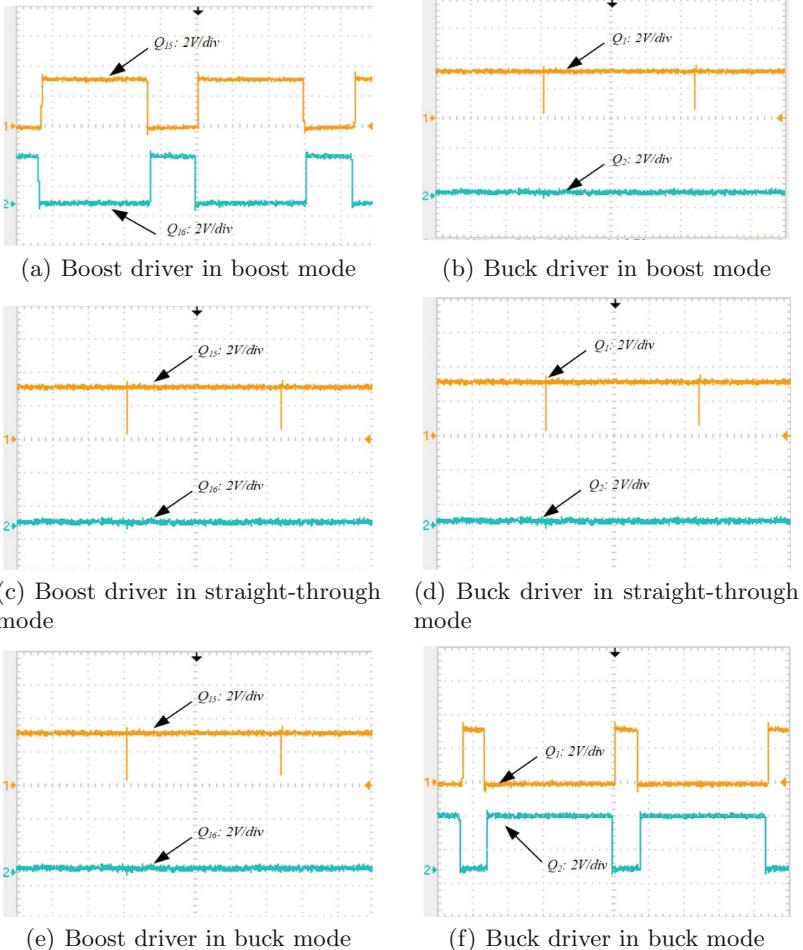
### 3 Experimental Verification

The paper builds a 1.1 kW photovoltaic module converter experimental prototype for experimental verification to test photovoltaic modules to switch circuit operating mode under the change of the external environment, and builds a 3.3 kW bidirectional AC/DC and DC/DC charger to test the driving waveform of the switch tube during the power conversion process of electric vehicles.

Figure 6(a) and (b) show the driving waveforms in boost operating mode. The switch tube  $Q_1$  remains on,  $Q_2$  is always off, and  $Q_{15}$  and  $Q_{16}$  are complementary on. Figure 6(c) and (d) show the driving waveforms when the photovoltaic module enters the straight-through mode. The switch tubes  $Q_1$  and  $Q_{15}$  are always on, and  $Q_2$  and  $Q_{16}$  are always off. Figure 6(e) and (f) show the driving

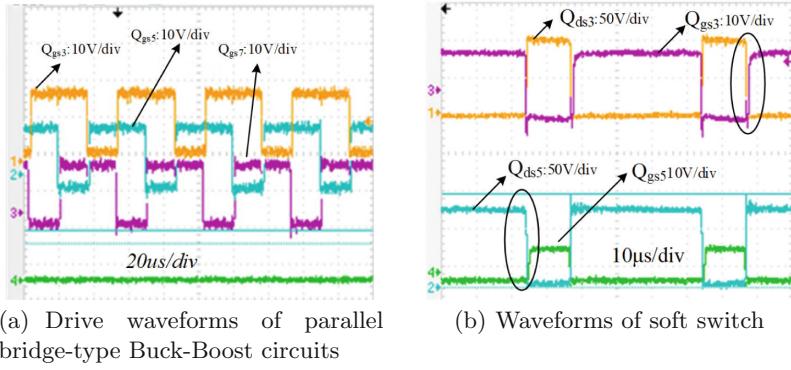
waveforms when the photovoltaic module enters Buck mode. The switch tube  $Q_{15}$  remains on,  $Q_{16}$  is always off, and  $Q_1$  and  $Q_2$  are complementary on.

The bidirectional DC/DC charger used in this paper is three parallel bridge-type Buck-Boost circuits. The phase difference of each phase circuit is  $120^\circ$ . Figure 7(a) shows the driving waveform of upper tube in each phase. The phase of switch tubes differ by  $120^\circ$ . Figure 7(b) shows the soft-switching waveform. Switch tube on and off occur when the drain-source voltage of the switch tube is 0, so the circuit implements soft switching.



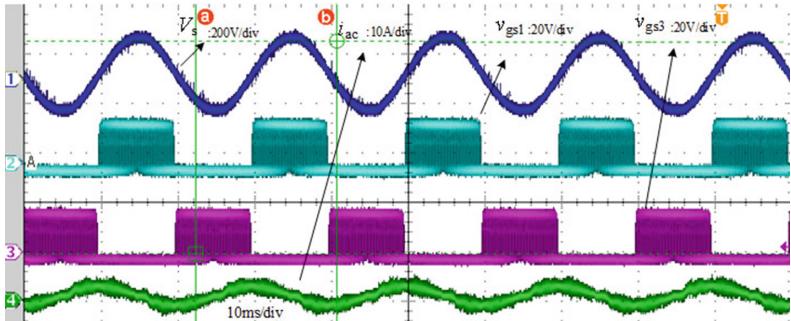
**Fig. 6.** Switch tube waveforms in different circuit operating modes

In this paper, the single-phase bidirectional full-bridge AC/DC circuit uses unipolar SPWM modulation. As shown in Fig. 8, CH1 is the grid voltage  $V_s$ ,



(a) Drive waveforms of parallel bridge-type Buck-Boost circuits

(b) Waveforms of soft switch

**Fig. 7.** Waveforms of switch tubes in DC/DC circuit**Fig. 8.** Waveforms of SPWM unipolar modulation

CH2 is the drive waveform of  $S_1$ , CH3 is the drive waveform of  $S_3$ , and CH4 is the waveform of grid-side current. It can be seen from the drive waveform that the SPWM unipolar modulation is realized.

From the above analysis, it can be seen that the photovoltaic module can switch the operating mode by controlling the on and off time of the switch tube according to the external environment changes. This paper realizes the bidirectional flow by interleaving parallel method between external vehicle battery and local vehicle battery and between grid and local vehicle battery.

## 4 Conclusions

In this paper, multi-port energy conversion system of electric vehicle based on bridge-type Buck-Boost topology is proposed for the application requirements for the energy conversion of electric vehicle. The paper studies on-board photovoltaic charging circuit to realize the rapid and efficient charging of electric vehicles; gives the control circuit for external vehicle battery and local vehicle battery, and uses the interleaving parallel method to realize the bidirectional flow of

energy; and gives the control method of AC/DC unipolar SPWM modulation to realize the energy conversion between grid and local vehicle battery. Through the study on the operating principle of the key circuit topology in energy conversion system and the technical analysis of soft switching, the energy conversion system with high efficiency, high power density and high reliability is realized.

**Acknowledgments.** This paper was supported in part by the Jiangsu Natural Science Foundation under Grant BK20181218, in part by the Equipment pre-research project under Grant 61873346, in part by the Science and Technology Cooperation Fund of Yangzhou City Hall project under Grant YZ2018136, YZ2018212, in part by the Open Project Fund of Yangzhou University Jiangdu Institute of High-end Equipment Engineering Technology under Grant YDJD201902, and in part by the Intelligent Energy Internet Research Institute Joint Fund of State Grid Yangzhou Power Supply Company and Yangzhou University under Grant SGTYHT/17-JS-202.

## References

1. Liu, Z.W., Hao, H., Cheng, X., Zhao, F.Q.: Critical issues of energy efficient and new energy vehicles development in China. *Energy Policy* **115**, 92–97 (2018)
2. Zhang, L., Qin, Q.D.: China's new energy vehicle policies: evolution, comparison and recommendation. *Transp. Res. Part A Policy Pract.* **110**, 57–72 (2018)
3. Araujo, K., Boucher, J.L., Aphale, O.: A clean energy assessment of early adopters in electric vehicle and solar photovoltaic technology: geospatial, political and socio-demographic trends in New York. *J. Clean. Prod.* **216**, 99–116 (2019)
4. Ding, C.W., Li, H.J., Zheng, W.W., Wang, Y.Z., Lin, X.: Reconfigurable photovoltaic systems for electric vehicles. *IEEE Design Test.* **35**, 37–43 (2018)
5. Pan, X.W., Li, H.Q., Liu, Y.T., et al.: An overview and comprehensive comparative evaluation of current-fed-isolated-bidirectional DCDC converter. *IEEE Trans. Power Electron.* **35**(3), 2737–2763 (2020)
6. Elserougi, A., Abdelsalam, I., Massoud, A., Ahmed, S.: A non-isolated hybrid-modular DC-DC converter for DC grids: small-signal modeling and control. *IEEE Access* **7**, 132459–132471 (2019)
7. Chakraborty, S., Chattopadhyay, S.: A dual-active-bridge-based fully ZVS HF-Isolated inverter with low decoupling capacitance. *IEEE Trans. Power Electron.* **35**(3), 2615–2628 (2019)
8. Shan, Y.H., Hu, J.F., Chan, K.W., et al.: Model predictive control of bidirectional DC-DC converters and AC/DC interlinking converters-a new control method for PV-wind-battery microgrids. *IEEE Trans. Sustain Energ.* **10**(4), 1823–1833 (2019)
9. Birbir, Y., Yurtbasi, K., Kanburoglu, V.: Design of a single-phase SPWM inverter application with PIC micro controller. *Eng. Sci. Technol. Int. J. Jestech* **22**(2), 592–599 (2019)



# Realization of Automatic Zero Calibration of Inductive Proximity Sensor Based on Inductive Increment Detection

Yuyin Zhao, Yu Fang<sup>(✉)</sup>, Jiajun Yang, Miao Weng, and Xiaonan Xia

School of Information Engineering, Yangzhou University, Yangzhou 225127, China  
yzfangyu@126.com

**Abstract.** Inductive proximity sensor has the advantages of high sensitivity, good stability and high reliability. According to the principle of electromagnetic induction, it converts the distance signal of the target metal into electrical signal so as to control the mechanical system. In order to obtain high precision of distance measurement, a closed-loop control circuit is designed to realize the bias inductance. The mathematical model of the bridge differential inductance detection circuit is established and the transfer function from the control to the bias inductance is derived. On this basis, an I-PI controller is proposed, the method of automatic zero calibration of inductive proximity sensor is realized and the control parameters of the controller are given, which is helpful to obtain the consistent high reliability of the product. Finally, simulation results verify the correctness of the mathematical model and control strategy.

**Keywords:** Inductive proximity sensor · Control strategy · Small signal model

## 1 Introduction

Because of its small volume, light weight, high precision, long life and no mechanical contact, inductive sensors can be used in industry and aviation. The inductive proximity sensor in military application can realize long-distance measurement and meet the high performance indexes such as sensitivity, stability and reliability. Its military applications range from star, two bombs, aircraft, ships, tanks, artillery and other equipment systems, to individual combat equipment; from communication technology detection system to logistics support system; from military scientific test to military equipment engineering; from battlefield operation to strategic and tactical command. Inductive proximity sensors are widely used in every link of war preparation and war implementation. They expand the time, space and frequency domain of future high-tech war, affect the way and efficiency of war, and improve the power of weapons and the efficiency of command.

When the inductive proximity sensor senses the proximity distance of the metal object, the inductance of the detection coil of the inductive head will

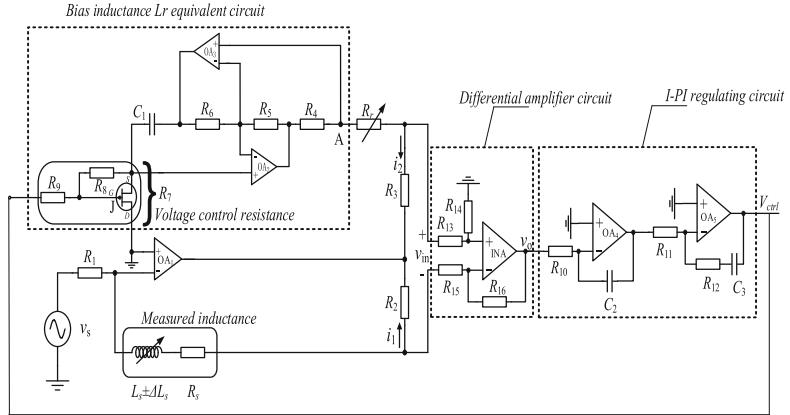
change with the change of the proximity distance of the metal object. Therefore, the proximity distance of the metal object can be measured by the inductance increment of the detection coil.

Resonance frequency measurement is to measure the resonance frequency of the resonance circuit composed of the detection coil and the capacitor in parallel, so as to get the close distance. However, the change of the inductance and the change of the output frequency are not linear, and the linear area is small, so it is difficult to achieve the long-distance measurement. The method of parallel connection of detecting coil and capacitance is also used in amplitude modulation measuring circuit. Different from frequency modulation method, the circuit is not working in resonance state, but in detuning state. This method has the same disadvantages as frequency modulation, so it is difficult to achieve the goal of long-distance measurement. The AC bridge measuring circuit is a kind of measuring circuit derived from the basic AC bridge. The AC bridge circuit can be used to measure the inductance elements with low Q value or high Q value. Because of the symmetry of the bridge arm, it can well suppress the temperature drift [1]. However, this method needs phase sensitive detection. The introduction of phase sensitive detection chip will bring temperature drift. Therefore, based on the AC bridge circuit, the bridge differential inductance detection circuit is adopted in this paper. The measurement method based on the bridge differential inductance detection circuit can cancel most of the common mode noise signals, and make sure that the detected information is effective as far as possible, so that the system has a strong anti-interference ability without affecting its own performance. For the bridge differential inductance detection circuit, it is necessary to use bias inductance and bias resistance to simulate the initial inductance and its equivalent resistance of the detection coil, so as to build a balanced bridge, so that the output of the differential detection circuit is zero before the incremental inductance measurement, which is called zero adjustment [2]. If the precise zero adjustment cannot be achieved, the increment of the measured inductance will be affected.

In this paper, the self-adaptive adjustment of bias inductance in bridge differential inductance detection circuit is realized, which can accurately measure the inductance increment and improve the production efficiency. Due to the use of differential measurement circuit, it can effectively suppress common mode noise, leave useful differential mode signal to represent the proximity distance of inductive proximity sensor, which is helpful for remote measurement. In the closed-loop control circuit to realize the bias inductance, the impedance matcher is used, and the bias inductance is realized by changing the voltage control resistance in the impedance matcher, which effectively solves the problem that it is difficult to make the bias inductance [3]. In the closed-loop control circuit to realize the bias inductance, the inductive proximity sensor self-adaptive zero calibration method based on the I-PI controller is proposed, which can realize the metal object entering. The automatic zero calibration before entering the measurement range of proximity switch ensures the accurate measurement of the inductance increment when the metal objects enter the measurement range

of proximity sensor, and also improves the batch productivity of products. Simulation results show that the system has good dynamic and static performance.

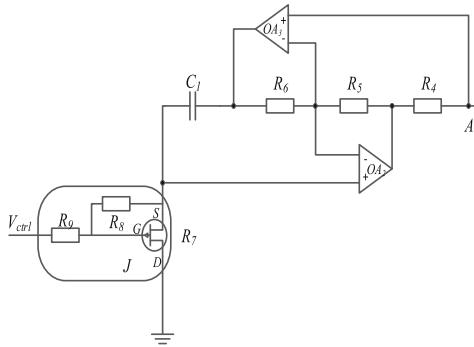
## 2 Working Principle of Measuring Circuit of Inductive Proximity Sensor



**Fig. 1.** Working principle diagram of measuring circuit of inductive proximity sensor

The working principle diagram of measuring circuit of inductive proximity sensor is shown in Fig. 1. When the metal body is close to the inductive surface of inductive proximity sensor, the inductance of the detection coil in the proximity sensor will change, but the inductance change is small, while the self inductance of the detection coil is large. In order to make the inductance change have a corresponding large electric quantity signal to improve or meet the measurement accuracy requirements, the The bridge type differential detection circuit is used, and then the electronic measurement circuit is realized by the instrument amplifier [4]. Figure 1 shows the schematic diagram of this electronic measurement circuit. In the figure,  $L_s$  represents the equivalent inductance of the detection coil,  $L_s$  is the self induction value of the detection coil, and  $R_s$  is the resistance of the detection coil; in the feedback channel of the operational amplifier, variable inductance  $L_r$  and adjustable resistance  $R_r$  series branch are used to correspond to the inductance and resistance of the detection coil. Before metal objects enter the measurement range of inductive proximity sensor, it is necessary to adjust the adjustable resistance  $R_r$  to be equal to the equivalent resistance  $R_s$  of the measured inductance, adjust the voltage control resistance  $R_7$  to make the equivalent bias inductance  $L_r$  equal to the measured inductance  $L_s$ , and select the resistance values of  $R_8$  and  $R_9$  to be equal [5], then the output  $V_{in}$  of the bridge differential inductance detection circuit is zero, and the output  $V_o$  of the amplifier is zero, which means that It's called zeroing. After zero adjustment, when

the metal object enters the measurement range of the inductance close to the sensor, the inductance of the detection coil in the induction head will change, that is, when the measured inductance changes and the increment  $\Delta L_s$  is generated, the output  $V_{in}$  of the bridge differential inductance detection circuit and the output  $V_o$  of the amplifier will not be zero, so the output signal  $V_o$  of the instrument amplifier can be used to represent the increment of the measured inductance, thus realizing The measurement of inductance increment can realize the measurement of proximity distance by inductive proximity sensor [6].



**Fig. 2.** The realization circuit diagram of bias inductance Lr

As shown in Fig. 2, resistance  $R_7$  is realized by the on resistance of the JFET. The model of the JFET in this paper is J2N4091. In this way, the specific value of  $R_7$  can be realized by controlling the gate voltage. Here it may be called voltage control resistance [7], we can get:

$$R_7 = -6.06V_{ctrl} + 40.16 \quad (1)$$

$$L_r = \frac{R_4 R_6 R_7 C_1}{R_5} \quad (2)$$

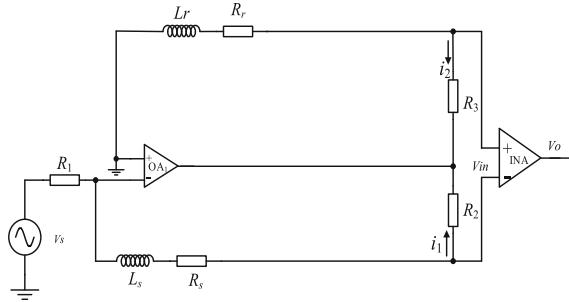
From this we can get:

$$L_r = \frac{R_4 R_6 (-6.06V_{ctrl} + 40.16) C_1}{R_5} \quad (3)$$

The resistance  $R_8$  and  $R_9$  near the JFET are taken as equal values to ensure that the JFET works in the linear region, and finally  $L_r = L_s$  [8].

### 3 Automatic Zero Calibration Method of Inductive Proximity Sensor Based on I-PI Controller

The main circuit schematic diagram of inductive proximity sensor is shown in Fig. 3:



**Fig. 3.** Measuring circuit of inductive proximity sensor

$V_s$  ( $V_s = V_m \sin(\omega T)$ ), equivalent oscillating excitation voltage source, connected to OA1 through  $R_1$ , according to the characteristics of OA1, the following formula is obtained:

$$I_1 = \frac{V_s}{R_1} \quad (4)$$

The current  $I_2$  flowing through  $L_r$ ,  $R_r$  and  $R_3$  is calculated as follows:

$$I_2 = \frac{I_1[R_s + R_2 + j\omega(L_s \pm \Delta L_s)]}{R_3 + R_r + j\omega L_r} \quad (5)$$

Error voltage signal applied to instrument amplifier  $V_{in}$ :

$$V_{in} = I_2 R_3 - I_1 R_2 \quad (6)$$

Substituting formula (4) and formula (5) into formula (6) to obtain:

$$V_{in} = I_1 \left[ \frac{R_s + R_2 + j\omega L_s}{R_3 + R_r + j\omega L_r} R_3 - R_2 \right] \quad (7)$$

Considering the symmetry of bridge structure, select parameters:  $R_3 = R_2 = 300 \Omega$ ,  $R_r = R_s = 36 \Omega$ ,  $L_r = L_s = 0.0005 \text{ H}$ , the amplification gain of instrument amplifier is set as  $A = 2$ , then its output  $V_o$  is calculated as follows (8):

$$V_{out} = 2V_{in} \quad (8)$$

Obviously, when  $R_3 = R_2$ ,  $R_r = R_s$ ,  $L_r = L_s$ , the output  $V_o$  of the circuit is equal to 0, and the circuit reaches a stable state.

### 3.1 Small Signal Modeling of Inductive Proximity Sensor

In the Inductive proximity sensor system, the integral-proportional integral regulator is used as the controller of the Inductive proximity sensor. The structure of I-PI controller is simple. It is composed of an integral regulator and a proportional integral regulator in series. The transfer function is shown in formula (9) [9]:

$$G_i(s) = \frac{K_{i1}}{s} \left( K_p + \frac{K_{i2}}{s} \right) \quad (9)$$

When the system is running, the control voltage  $V_{ctrl}$  to  $L_r$  is nonlinear, which is not convenient for the analysis and design of the control loop. The small signal model reflects the dynamic behavior of the circuit near a steady-state operating point, which is the basis of feedback control. In this paper, the disturbance method is used to build the small signal model [10].

According to Fig. 3, the circuit can be divided into equations:

$$V_{in} + \Delta V_{in} = I_1 \left[ \frac{Rs + R_2 + j\omega Ls}{R_3 + Rr + j\omega Lr} R_3 - R_2 \right] \quad (10)$$

Thus, formula (11) can be obtained:

$$V_{in} + \Delta V_{in} + I_1 R_2 = I_1 \left[ \frac{Rs + R_2 + j\omega Ls}{R_3 + Rr + j\omega Lr} R_3 \right] \quad (11)$$

$V_{in}$  is the steady-state voltage, and it is further simplified as follows:

$$V_{in} j\omega \Delta Lr + I_1 R_2 j\omega \Delta Lr = - \Delta V_{in} (R_3 + Rr + j\omega Lr) \quad (12)$$

$$\frac{\Delta V_{in}}{\Delta Lr} = - \frac{V_{in} j\omega + I_1 R_2 j\omega}{R_3 + Rr + j\omega Lr} \quad (13)$$

Both sides of the equation take Laplace transformation at the same time to get:

$$\frac{\Delta V_{in}(s)}{\Delta Lr(s)} = - \frac{V_{in}s + I_1 R_2 s}{R_3 + Rr + sLr} \quad (14)$$

Then the transfer function from inductance to error voltage can be obtained, as shown in formula (15):

$$G_i(s) = - \frac{V_{in}s + I_1 R_2 s}{R_3 + Rr + sLr} \quad (15)$$

### 3.2 Control Parameter Design of Inductive Proximity Sensor

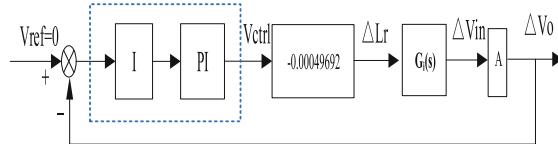
The block diagram of I-PI control loop is shown in Fig. 4,  $R_3 = R_2 = 300 \Omega$ ,  $Rr = Rs = 36 \Omega$ ,  $Lr = Ls = 0.0005 \text{ H}$ ,  $V_{in} = 0 \text{ V}$ ,  $V_s = 8.2 \text{ V}$ ,  $R_1 = 82 \Omega$ . If there is no regulator control, the transfer function of open-loop system is:

$$G_i(s) = \frac{0.0298s}{336 + 0.005s} \quad (16)$$

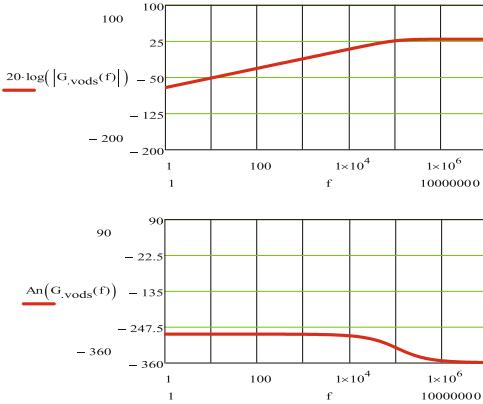
The Bode diagram is shown in Fig. 5. It can be seen from Fig. 5 that before the regulator is added, the cut-off frequency is 3.589 kHz. At this time, the phase angle is  $-271.9^\circ$  (below  $-180^\circ$ ), so the system is unstable and needs to be corrected.

The expression of I-PI controller is:

$$G_i(s) = \frac{K_{i1}}{s} \left( K_p + \frac{K_{i2}}{s} \right) \quad (17)$$



**Fig. 4.** Closed loop control block diagram of realizing bias inductance



**Fig. 5.** Bode plot of current open-loop without regulator

Then the transfer function added to the I-PI controller is:

$$G_{iopen}(s) = \frac{0.0298s}{336 + 0.005s} \cdot \frac{K_{i1}}{s} (K_p + \frac{K_{i2}}{s}) \quad (18)$$

The phase angle margin of the corrected open-loop transfer function is set to  $45^\circ$ , which can be obtained from the control theory:

$$|G_{iopen}(s)| = 1 \quad (19)$$

$$\gamma = 180 + \angle G_{iopen}(s) = 45^\circ \quad (20)$$

The solution is:

$$K_{i1} = 3.086 \times 10^4 \quad (21)$$

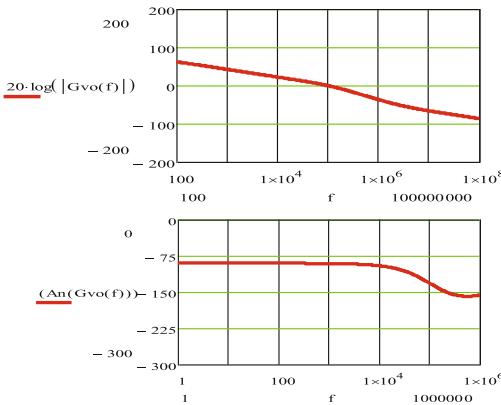
$$K_p = 0.034 \quad (22)$$

$$K_{i2} = 6.279 \times 10^5 \quad (23)$$

Substituting Eq. (17) into Eq. (16), the corrected open-loop transfer function of current loop can be obtained as follows:

$$G_{iopen}(s) = \frac{0.0298}{336 + 0.0005s} \times 30860 \times (0.034 + \frac{6.279 \times 10^5}{s}) \quad (24)$$

Its Bode diagram is shown in Fig. 6, It can be seen from Fig. 6 that the expected correction effect is achieved.

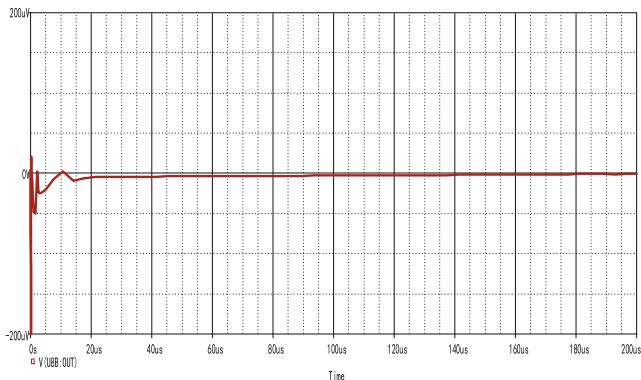


**Fig. 6.** Bode plot with regulator

## 4 Simulation Verification

In order to verify the feasibility of the control scheme, simulation is carried out. Figure 7 shows the simulation waveform.

The waveform is the output waveform of  $V_o$  adjusted by the controller. It can be seen that the input voltage of the upper half of the bridge arm of the bridge differential circuit is adjusted by the controller, so as to change the value of  $L_r$ , make it continuously close to  $L_s$ , and finally reach a balance. The simulation result shows that the circuit has good stability and dynamic performance, which verifies the feasibility of the control strategy.



**Fig. 7.** Simulation waveform of Inductive proximity sensor regulator

## 5 Conclusion

In this paper, the small signal model of Inductive proximity sensor is established, and the transfer function from control voltage to inductance of Inductive proximity sensor is obtained through calculation and derivation, and then the design method of I-PI control regulator parameters is given. The control model proposed in this paper is helpful to realize precise automatic zero adjustment and mass production.

**Acknowledgement.** This paper was supported in part by the Equipment pre-research project under Grant 61873346, in part by the Science and Technology Cooperation Fund of Yangzhou City Hall project under Grant YZ2018136, in part by the Open Project Fund of Yangzhou University Jiangdu Institute of High-end Equipment Engineering Technology under Grant YDJD201902, and in part by the Intelligent Energy Internet Research Institute Joint Fund of State Grid Yangzhou Power Supply Company and Yangzhou University under Grant SGTYHT/17-JS-202.

## References

1. Nie, M., Chen, J.Q., Xu, F.: Design of temperature drift compensation structure for flexible pressure sensor. *J. Sensing Technol.* **10**, 1443–1446 (2019)
2. Li, M.: A kind of sensor auto zero circuit based on Microcomputer. *Internet of things technology* **2**, 49–51 (2011)
3. Yang, L., Dai, J.F., Zhao, H.C.: Research on frequency tracking method of ultrasonic scalpel based on impedance matching. *Appl. Electron. Technol.* **10**, 105–111 (2019)
4. Dai, D.L.: Exploration and application of proximity sensor in a certain aircraft. *Modern Man. Technol. Equ.* **4**, 42–43 (2017)
5. Lan, T.Q., Lai, H.B.: A linear sensitive active half bridge circuit for resistance sensor and its compensation. *J. Electron. Meas. Inst.* **5**, 134–141 (2019)
6. Yan, L.H.: Design and analysis of door sensor of civil aircraft. *Shandong Ind. Technol.* **15**, 104 (2017)
7. Hang, H.Y., Zhao, C., Yu, B.Y.: Design and application of JFET high precision variable resistor. *Instrument* **8**, 1884–1891 (2015)
8. Chang, K., Li, S.J., Li, Z.S.: Application and key technology of inductive proximity sensor in aircraft. *Aviation manufacturing technology* **20**, 76–79 (2015)
9. Xie, C.H., Lu, S.J.: Event driven output feedback control in wireless sensor networks. *Computer Res. Devel.* **11**, 2639–2645 (2017)
10. Han, X.J.: Problems and solutions of two wire proximity switch. *Cement* **1**, 61–62 (2018)



# Sliding Mode Control Method of Powered Parafoil Based on Extended State Observer

Li Yu, Qinglin Sun<sup>(✉)</sup>, and Panlong Tan

Nankai University, Tianjin 300350, China  
sunql@nankai.edu.cn

**Abstract.** Aiming at the problem of altitude control of powered parafoil system with nonlinear and strong coupling characteristic, a powered parafoil flight control algorithm based on extended state observer (ESO) and sliding mode control (SMC) was studied. Considering relative pitch and yaw motion, an eight degree of freedom model of parafoil dynamics model was built. The composite disturbance composed of internal and external disturbances of the system was accurately estimated by designing an extended state observer and compensated by sliding mode controller in real-time, which solved the control problem of the complex second-order system. The stability of the system was proved by the Lyapunov function. Simulation results show that the method can effectively overcome the influence of compound disturbances and implement precise altitude control. Compared with the standard LADRC controller, the sliding mode control based on the extended state observer has higher control accuracy and better robustness.

**Keywords:** Extended state observer · Sliding mode control · Altitude control · Powered parafoil

## 1 Research Background, Significance and Progress

The powered parafoil is a type of aircraft that originated from the air sports of Western countries. It is easy to operate and has a high safety factor. Therefore, it is widely used in airdrop hardware, military reconnaissance, and civilian fields. Different from traditional parafoil, powered parafoil have a power system, thus, they have stronger stagnation capability and better controllability, so they are widely studied. As for the control algorithm, Xie Yarong [1] proposed a nonlinear predictive control method based on the analysis of the homing trajectory, using a fuzzy controller. Zhang Hao [2] proposed a variable gain adaptive fuzzy back-stepping control strategy for the fixed-wing flight of power parachute with uncertain model. Chen Qi [3] Studied the multi-wing parachute cluster control problem, collected the attitude information of each wing parachute, and realized the wing parachute assembly and collision avoidance by the field potential method.

Sliding Mode Control (SMC) is a kind of special nonlinear control method. The sliding mode can be designed with few adjustable parameters, quick response and strong anti-disturbance ability. However, in practice, most of the models need to be accurately modeled. When some state variables cannot be measured, the control method will fail. Moreover, due to the discontinuous switching characteristics of sliding mode variable structure control, chattering phenomenon of the system will occur, which greatly restricts the development of sliding mode control. Li kunyang [4] established the train delay control model and proposed a target speed curve tracking control algorithm based on sliding mode active disturbance rejection, which has good robustness to internal and external disturbances of the system. Jinyue [5] proposed a back-stepping sliding mode control method for ship dynamic positioning control system based on an extended observer and used Lyapunov method to prove the stability of the system. Wang bingyuan [6] studied the attitude control of highly nonlinear flapping wing aircraft, and proposed the adaptive weighted approach law terminal sliding mode control method based on the time-scale separation principle.

Professor Han jing qing proposed Active Disturbance Rejection Control (ADRC), which is a project oriented Control algorithm [7]. It is mainly composed of Extended State Observer (ESO), differential tracker and control law of state error feedback. As the core part of the control algorithm, the Extended State Observer can observe and compensate the system error and nonlinear uncertainty disturbance. Therefore, it is very suitable for flexible aircraft with strong coupling characteristics such as dynamic parafoil [8]. Wang yiyi [9] proposed a method of nonlinear ADRC for attitude control of 2-DOF unmanned helicopter, and developed a nonlinear extended state observer and a nonlinear feedback control law to improve control quality. Pan zhen [10] studied the nonlinear attitude control of tilt-rotor aircraft. Combining the ADRC theory, sliding mode control theory and dynamic surface control theory, the new ADRC sliding mode control algorithm has the advantages of strong disturbance adaptability and clear parameters. Zhang yong [11] designed a dynamic surface controller based on ADRC to solve the control problems of nonlinear, strong coupling and disturbance sensitivity in attitude control of quadrotors.

Based on the 8-DOF parafoil dynamics model established by considering the relative pitch and relative yaw motion between the carrier and the paraglider, the Extended State observer-sliding Mode Control (ESO-SMC) was designed in this paper. The perturbation and uncertainty of the parafoil system are estimated by ESO, and the exponential approach sliding mode controller is designed, which is compared with the traditional active disturbance rejection control method, and the disturbance rejection performance of the control method is verified under gust disturbance.

## 2 Mathematical Model

In the modeling of the parafoil dynamic model, the relative pitch and relative yaw motion between the carrier and the canopy need to be considered, and an

8-DOF parafoil dynamic model based on the Kirchhoff equation of motion is established.

$$\frac{\partial P_w}{\partial t} + W_w \times P_w = F_w^{aero} + F_w^G + F_w^t + F_w^{th} \quad (1)$$

$$\frac{\partial H_w}{\partial t} + W_w \times H_w = M_w^{aero} + M_w^t + M_w^f \quad (2)$$

$$\frac{\partial P_s}{\partial t} + W_s \times P_s = F_s^{aero} + F_s^G + F_s^t \quad (3)$$

$$\frac{\partial H_s}{\partial t} + W_s \times H_s = M_s^{aero} + M_s^t + M_s^f + M_s^G \quad (4)$$

In the above formula,  $V = [u \ v \ w]$  and  $W = [p \ q \ r]$  represent speed and angular velocity respectively;  $P$  and  $H$  represent momentum and momentum moment;  $F$  and  $M$  represent force and torque respectively. The subscripts  $w, s$  represent the coordinate system of the object and the parafoil respectively; the superscripts  $aero, G, t, th$  and  $f$  represent the aerodynamic force, gravity, parafoil rope tension and friction, respectively. The calculation of the aerodynamics of the canopy introduces the idea of piecewise calculation [12], which divides the canopy into eight parts in the direction of the average.

The calculation method of the momentum and momentum moment of an object is as follows:

$$\begin{bmatrix} P_w \\ H_w \end{bmatrix} = \begin{bmatrix} m_w & 0 \\ 0 & J_w \end{bmatrix} \begin{bmatrix} V_w \\ W_w \end{bmatrix} \quad (5)$$

For the calculation of the momentum and momentum moment of the parafoil, the influence of the additional mass must be considered. With reference to the calculation method proposed by Lissaman and Brown, the additional mass and the moment of inertia are expressed in the form of a matrix:

$$\begin{bmatrix} P_s \\ H_s \end{bmatrix} = [A_a + A_r] \begin{bmatrix} V_s \\ W_s \end{bmatrix} \quad (6)$$

Among them,  $A_a, A_r$  respectively represent the true mass and parasitic mass matrix of the parafoil. The parafoil and the object are connected by parafoil ropes. There is relative movement between the two, and the constraints of speed and angular velocity need to be satisfied. At the midpoint of the two connection points, there are

$$V_w + W_w \times L_{w-c} = V_s + W_s \times L_{s-c} \quad (7)$$

$$W_w = W_s + \tau_s + \kappa_s \quad (8)$$

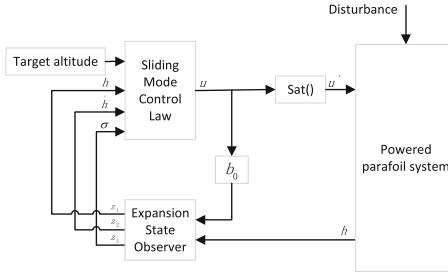
The simultaneous formulas (1)–(8) can be used to establish the 8-DOF model of the parafoil system.  $x = [V_w^T \ W_w^T \ V_s^T \ W_s^T \ \Psi_r \ \theta_r]^T$  Taking  $x$  as the system state variable, the kinematics equation can be obtained:

$$x = \left( [D_1^T \ D_2^T \ D_3^T \ D_4^T]^T \right)^{-1} [E_1^T \ E_2^T \ E_3^T \ E_4^T]^T \quad (9)$$

The detailed derivation process and the specific form of equations can be referred to [13].

### 3 Design of Parafoil Altitude Controller

The method proposed in this paper is mainly composed of an extended state observer and a sliding mode control law. The control structure diagram of the power parafoil system is as follows Fig. (1):



**Fig. 1.** Structure diagram of sliding mode controller based on ESO

#### 3.1 Design of the Extended State Observer

According to the characteristics of the parafoil system model, the current height can be expressed as a second-order form [14]:

$$\ddot{h} = f + b_0 u \quad (10)$$

Among them,  $f = f(h, \dot{h}) + d$  is regarded as the total disturbance of the system,  $f(h, \dot{h})$  is the expression of the system state variable,  $d$  is the external disturbance,  $b_0$  is the input gain. It is assumed that the derivative  $\sigma$  of  $f$  is existential and bounded, i.e.  $|\sigma| < L$ . Rewriting the above Eq. (10) as the extended state equation is:

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = x_3 + b_0 u \\ \dot{x}_3 = \sigma \\ y = x_1 \end{cases} \quad (11)$$

In the formula,  $x_1$  is the current altitude  $h$ ;  $x_2$  is the rate of change of altitude  $\dot{h}$ ;  $\sigma$  is the observable measurement. A third-order linear extended state observer is established:

$$\begin{cases} \dot{z}_1 = z_2 + L_1(y - z_1) \\ \dot{z}_2 = z_3 + b_0 u + L_2(y - z_1) \\ \dot{z}_3 = L_3(y - z_1) \end{cases} \quad (12)$$

The observer can realize that  $t \rightarrow \infty$ ,  $z_i(t) \rightarrow x_i(t)$ , and the observation gain is  $L = [3\omega_0 \ 3\omega_0^2 \ \omega_0^3]^T$ ,  $\omega_0$  is the bandwidth of the observer.

### 3.2 Analyze the Estimation Ability of the Observer

Define the observation error  $\eta = [\eta_1 \ \eta_2 \ \eta_3]$ , here,  $\eta_i = x_i - z_i$ , estimation errors equation of state can be obtained:

$$\dot{\eta} = A\eta + B\sigma \quad (13)$$

$$\text{where, } A = \begin{bmatrix} -3\omega_0 & 1 & 0 \\ -3\omega_0^2 & 0 & 1 \\ -\omega_0^3 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Obviously, when  $\omega_0 > 0$ ,  $A$  is Hurwitz, the characteristic roots are in the negative half-plane, so the system is stable. For any given positive definite matrix  $Q$ , there exists a symmetric positive definite matrix  $P$ , satisfying  $A^T P + PA = -Q$ . Select  $V_o = \eta^T P \eta$  as Lyapunov function, then

$$\dot{V}_o = \dot{\eta}^T P \eta + \eta^T P \dot{\eta} \leq -\lambda_{min}(Q) \|\eta\|^2 + 2\|\eta\| \|PB\| \sigma \quad (14)$$

Because of  $\frac{V_o}{\lambda_{max}(P)} \leq \|\eta\|^2 \leq \frac{V_o}{\lambda_{min}(P)}$ , let  $W = \sqrt{V_o}$ , substitute formula (14),

$$\dot{W} \leq -\frac{\lambda_{min}(Q)}{2\lambda_{max}(P)} W + \frac{\|PB\| \sigma}{\sqrt{\lambda_{min}(P)}} \quad (15)$$

Solving the inequality,

$$W \leq \frac{2\lambda_{max}(P) \|PB\| \sigma}{\lambda_{min}(Q) \sqrt{\lambda_{min}(P)}} + \left( -\frac{2\lambda_{max}(P) \|PB\| \sigma}{\lambda_{min}(Q) \sqrt{\lambda_{min}(P)}} + W(t_0) \right) e^{-\frac{\lambda_{min}(Q)}{2\lambda_{max}(P)}(t-t_0)} \quad (16)$$

$$\lim_{t \rightarrow \infty} W = \frac{2\lambda_{max}(P) \|PB\| \sigma}{\lambda_{min}(Q) \sqrt{\lambda_{min}(P)}} \quad (17)$$

By further analysis, the upper bound of the error of the extended state observer is obtained:

$$\lim_{t \rightarrow \infty} \|\eta\| \leq \frac{\sqrt{V_o}}{\sqrt{\lambda_{min}(P)}} = \frac{W}{\sqrt{\lambda_{min}(P)}} = \frac{2\lambda_{max}(P) \|PB\| \sigma}{\lambda_{min}(Q)} \quad (18)$$

Therefore, for a general second-order system, in the case where the disturbance  $\sigma$  is derivable and bounded, there is always a linear extended observer to make the estimation error bounded.

### 3.3 Design of Sliding Mode Controller

Definition of sliding surface:

$$s = ce + \dot{e} \quad (19)$$

Among,  $e = x_1 - x_d$ ,  $c > 0$ . Define the Lyapunov function:

$$V_s = \frac{1}{2} s^2 \quad (20)$$

Design the sliding mode control law based on the extended state observer:

$$u = \frac{1}{b_0}(-k_1\hat{s} - k_2 sgn(s) - \hat{v} - \hat{f}) \quad (21)$$

Among them,  $\hat{v} = c\dot{\hat{e}} - \ddot{x}_d$ ,  $\hat{e} = z_1 - x_d$ ,  $\dot{\hat{e}} = z_2 - \dot{x}_d$ , then,

$$\begin{aligned} \dot{V}_s &= s\dot{s} = s(c\dot{\hat{e}} + \ddot{e}) \\ &= s(c\dot{\hat{e}} + bu + f - \ddot{x}_d) \\ &= s(-k_1s - k_2 sgn(s) + e_3) \\ &= -k_1s^2 - k_2|s| + se_3 \end{aligned} \quad (22)$$

Among them,  $e_3$  depends on the estimation error of each state of the extended state observer. Set  $k_2 > \|e_3\|$ , therefore,

$$\dot{V}_s(t) < 0 \quad (23)$$

The convergence of the sliding surface can be guaranteed.

When the sliding mode surface converges to  $s = 0$ , it can be known from Eq. (19)

$$\dot{e} = -c_0e \quad (24)$$

Let  $V_e = \frac{1}{2}e^2$ , and take the derivative of it,

$$\dot{V}_e = e\dot{e} = -c_0e^2 < 0 \quad (25)$$

Therefore, the closed-loop system converges gradually to realize the tracking of the target height  $h_d$ .

## 4 Simulation Experiment

In order to better appraisal the height tracking effect of the SMC-ESO, two indicators for evaluating tracking performance are defined.

1. Maximum overshoot of trajectory tracking  $e_{max}^h$

$$e_{max}^h = \max(|h(t)|), t = t_0 \sim t_1 \quad (26)$$

It refers to the maximum value of the deviation of the parafoil system when flying at a fixed height.

2. Variance  $\sigma^h$

$$\sigma^h = \frac{1}{N} \sum_{i=1}^N [h(i)]^2 \quad (27)$$

**Table 1.** Main performance parameters of power parafoil

Parameters	Value	Unit
Chord length	1.30	m
Horizontal length	4.50	m
Aspect ratio	3.46	
Length of parafoil rope	3.00	m
Parafoil area	6.50	$m^2$
Angle of installation	10.00	(·)
Parafoil quality	1.70	kg
Load mass	20.00	kg
Thrust force	0~400.00	N

Refers to the variance of the error between the reference height and the actual height in  $N$  samplings after the preset height is reached for the first time.

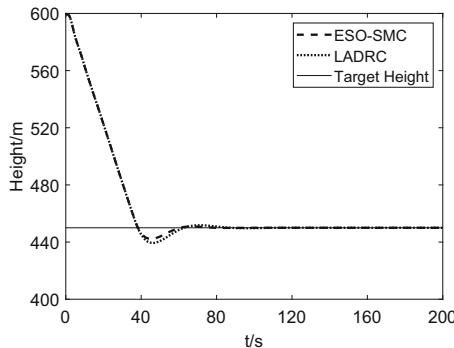
In this paper, the parafoil model is used to perform simulation experiments with reference to the actual airdrop parafoil type. The main performance parameters of the parafoil are as follows (Table 1):

The simulation experiment was carried out by using the parafoil system dynamic model and altitude control algorithm constructed above. The initial speed of the powered parafoil is 0 m/s, the initial height is 600 m, and the target height is 450 m. The ESO-SMC method and the LADRC method are used to implement height control experiments with or without external wind disturbances [15]. The simulation time is 200 s and the step length is 0.025 s. Among them, in the 100 s, a discrete gust with a maximum wind speed of 3 m/s was added to the system, and the action time was 15 s. Controller parameter selection: LADRC controller parameters are set  $\omega_0 = 1$ ,  $k_p = 1$ ,  $k_d = 2$ ,  $b_0 = 0.2$ ; ESO-SMC controller parameter are set  $\omega_0 = 1$ ,  $b_0 = 0.2$ ,  $k_1 = 1$ ,  $k_2 = 1$ . It can be seen that the bandwidth  $\omega_0$  of the observer is the same, which can verify the performance of the height control of the powered parafoil under the same extended state observer. The simulation results are as follows: Note: Due to the obvious chattering caused by the symbolic function  $\text{sgn}(\cdot)$  in sliding mode control, the hyperbolic tangent function  $\tanh(\cdot)$  is used in practice instead.

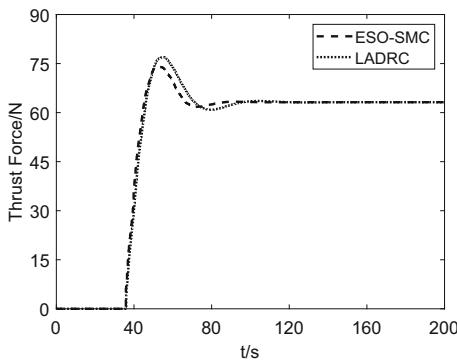
Table 2 shows the performance comparison between the ESO-SMC method and the LADRC method with and without wind disturbance.

**Table 2.** Comparison of height errors

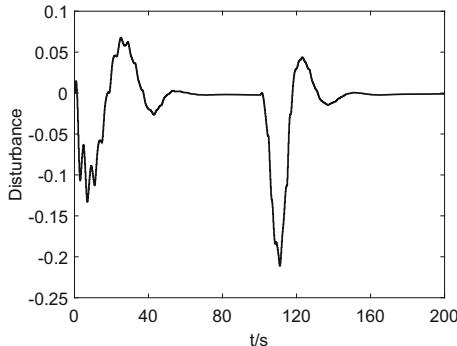
Environment	ESO-SMC		LADRC	
	$e_{max}^h$	$\sigma^h$	$e_{max}^h$	$\sigma^h$
No disturbance	7.52	0.69	10.61	1.14
Gust disturbance	2.54	0.32	3.14	0.52



**Fig. 2.** Height variation curve

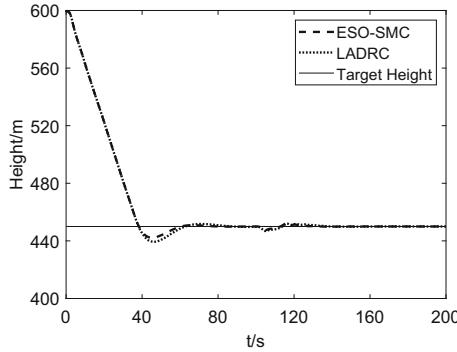


**Fig. 3.** Controlling thrust output

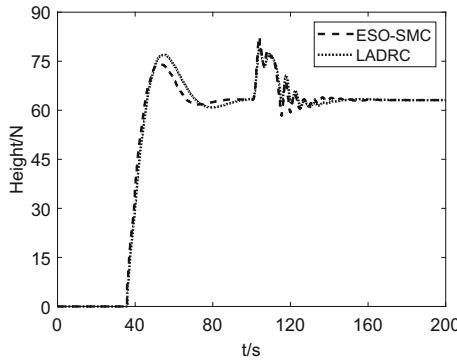


**Fig. 4.** Estimation of disturbance by the extended state observer

It can be seen from Fig. 2 and Fig. 3 that the ESO-SMC method and LADRC method used in this paper both can achieve the tracking of the specified height



**Fig. 5.** Height variation curve under external wind disturbance



**Fig. 6.** Controlling thrust output under external wind disturbance

without external disturbance. However, under the same extended state observer, the controller designed in this paper responds faster, and the output overshoot of the control quantity is small. From the perspective of height error, the overshoot of the height error is also smaller and the convergence speed is faster.

It can be seen from Fig. 4 that after the parafoil system has tracked to the target height steadily, the observed system also tends to be stable. When it reaches 100s, the horizontal gust disturbance is added to the system, and it can be clearly seen that the observer responded quickly and estimate accurately the internal and external disturbances of the system. As a flexible aircraft, the powered parafoil has a strong coupling between the horizontal movement and the vertical movement, so the horizontal gust disturbance will also have a certain effect on the altitude control. It can be seen from Fig. 5 and Fig. 6 that both algorithms show good anti-disturbance performance. In the face of gust disturbances, the height control does not show a large control deviation. However, it can be seen from Table 2 that the method used in this paper has smaller overshoot and quick response; and the variance of height error is smaller than that of LADRC

method during the entire disturbance stage, which indicates that the ESO-SMC method is more sensitive to disturbance. In summary, the sliding mode control method based on the extended state observer proposed in this paper has good control performance, can achieve a stable output of the control quantity, and ensure the accuracy of the powered parafoil fixed-height flight.

## 5 Conclusion

This paper proposes a sliding mode control method based on the extended state observer for the 8-DOF parafoil dynamic model, combined with the advantages of sliding mode variable structure control and extended state observer. This method can effectively use the extended state observer to accurately estimate the composite disturbance composed of the internal disturbance and the external disturbance, and get the compensation in the sliding mode controller, which is suitable for practical engineering applications. The simulation results show that the ESO-SMC method can more effectively improve the control accuracy and anti-disturbance performance of the parafoil system fixed height control than the traditional LADRC method. The focus of the following work is to consider the coupling of the parafoil system and extend the algorithm to horizontal trajectory tracking control. At the same time, the actual flight verification of the control strategy proposed in this paper is also required.

**Acknowledgements.** This work was supported by grant-61973172 and 61973175 of the National Natural Science Foundation of China, the grant-19JCZDJC32800 of the key Technologies Research and Development Program of Tianjin. This work also has been supported by the funding Research on intelligent control theory of large parafoil from Beijing Institute of Space Mechanics Electricity, fifth institute, China Aerospace Science and Technology Corporation.

## References

1. Xie, Y.: Research on Modeling and Flight Control under the Airdrop Mission. Nanjing University of Aeronautics and Astronautics(2011). <https://doi.org/10.7666/d.d166701>
2. Zhang, H., Chen, Z., Qiu, J.: Adaptive fuzzy backstepping for reducing altitude control of parafoil based on variable gain. *Syst. Eng. Electron. Technol.* **38**(440(05)), 156–161(2016). <https://doi.org/10.3969/j.issn.1001-506X.2016.05.24>
3. Chen, Q., Zhao, M., Zhao, Z., Ma, M., Huang, R.: Multiple autonomous parafoils system modeling and rendezvous control. *Acta Aeronautica Sinica* **37**(10), 3121–3130 (2016). <https://doi.org/10.7527/S1000-6893.2016.0047>
4. Li, K., Chen, H., Guo, J., et al.: Research on automatic train operation algorithm for high-speed trains based on sliding mode active disturbance rejection control. *Mod. Comput.* **15**, 25–31 (2019)
5. Jin Yue, Yu., Menghong, Y.W., et al.: Back-stepping sliding mode control of ship dynamic positioning system based on extended state observer. *Ship Sci. Technol.* **2**, 103–107 (2017). <https://doi.org/10.3404/j.issn.1672-7619.2017.02.021>

6. Wang, B., Zhang, S., Zheng, F., Li, X.: Attitude control of flapping-wing aircraft based on adaptive terminal sliding mode. *Control Eng. China* **27**(02), 309–315 (2020). <https://doi.org/10.14107/j.cnki.kzgc.20190134>
7. Han, J.: From PID to active disturbance rejection control. *IEEE Trans. Ind. Electron.* **56**(3), 900–906 (2009). <https://doi.org/10.1109/TIE.2008.2011621>
8. Tan, P., Luo, S., Sun, Q.: Control strategy of power parafoil system based on coupling compensation. *Trans. Beijing Inst. Technol.* **039**(004), 378–383 (2019). <https://doi.org/10.15918/j.tbit1001-0645.2019.04.008>
9. Wang, Y., Zhao, Z.: Nonlinear active disturbance rejection attitude control of two-DOF unmanned helicopter. *Acta Automatica Sinica* **05**(06), 1–14 (2020). <https://doi.org/10.16383/j.aas.c190521>
10. Zhen, P., Chengzhi, C., Jingkai, Z., et al.: Nonlinear attitude control of tilt rotor aircraft based on active disturbance rejection sliding mode theory. *Aero Weaponry* **308**(06), 46–51 (2018). <https://doi.org/10.19297/j.cnki.41-1228/tj.2018.06.007>
11. Zhang, Y., Chen, Q., Zhang, X., Sun, Q., Sun, M.: Dynamic surface attitude control of quad-rotor UAV based on ADRC . *J. Jilin Univ. (Engineering and Technology Edition)* **49**(02), 562–569 (2019). <https://doi.org/10.13229/j.cnki.jxbgxb20171241>
12. Xiong, J.: Research on the dynamics and homing project of parafoil system. National University of Defense Technology (2005). <https://doi.org/10.3969/j.issn.1009-8518.2003.03.003>
13. Zhu, E., Sun, Q., Tan, P., Chen, Z., et al.: Modeling of powered parafoil based on Kirchhoff motion equation. *Nonlinear Dyn.* **79**(1), 617–629 (2015). <https://doi.org/10.1007/s11071-014-1690-9>
14. Jin, T., Qinglin, S., Qiang, C., et al.: Linear active disturbance rejection altitude control for parawing unmanned aerial vehicle. *J. Natl. Univ. Defense Technol.* **039**(006), 103–110 (2017). <https://doi.org/10.11887/j.cn.201706016>
15. Xiong, J., Qin, Z., Cheng, W.: The characteristics and description of mid-high altitude wind in recovery. *Spacecr. Recovery Remote Sens.* **03**, 13–18 (2003). <https://doi.org/10.3969/j.issn.1009-8518.2003.03.003>



# Analysis of Accelerated Vibration-Magnetic Effect of 25CrMo4

Zhenfa Bi<sup>(✉)</sup> and Zongkai Wang

School of Railway Transportation, Shanghai Institute of Technology, Shanghai, China  
bizhenfa@sit.edu.cn, 1064169568@qq.com

**Abstract.** With the purpose of researching the effect of train vibrancy on the security of high-speed rail wheels, it is depended on Metal Magnetic Memory detection technology, accelerated vibration fatigue test method is adopted, and the ANSYS simulation software is used to analyze the impact of the accelerated vibrancy of 25CrMo4 stuff on the magnetic memory signal. By increasing the vibration amplitude and frequency, the actual vibration time 800d, 1200d, 1600d, 2000d is shortened equivalently. From 800d to 2000d, the tangential average value of magnetic memory values are reduced from 104.5 A/m to -10 A/m, the normal mean value increases from -32 A/m to -20 A/m. In addition, as the vibrancy time augments, the magnetic field strength vector value gradually decreases. According to the analysis of frequency spectrum, the lower frequency signal is located 0 Hz~10 Hz. As the vibrancy time increases, the magnetic memory signal energy concentration phenomenon appears. Furthermore, the effect of high-frequency noise signals on the energy distribution of metal magnetic memory signal gradually decreases. Through acceleration and vibration simulation analysis of 25CrMo4 material, the connection of metal magnetic memory value with vibrancy time is verified. The relevance at simulation and measured signals provide significant guides to wheelset operation security detection.

**Keywords:** High-speed rail wheels · 25CrMo4 · Magnetic memory detection · Accelerated vibration · ANSYS

## 1 Introduction

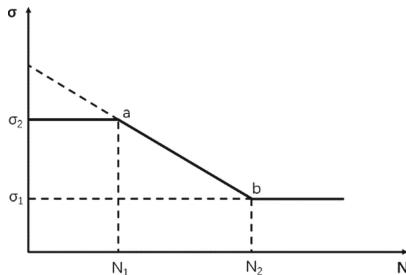
As the velocity of high-speed rail and the load capacity continues to increase, the long-term moving results in the working environment of the high-speed rail wheels worse. During the high-speed operation of the high-speed rail, a series of external factors, such as the collision between wheelset and track and the slight height inequality of track, cause the train wheelset to produce a long time of violent vibration. The fatigue accumulation inside the high-speed rail wheels has seriously affected the security and stability of high-speed train operation [1]. Therefore, the efficient detection and early prediction of wheelset is an momentous research to insure the security of high-speed railway.

The fatigue vibration test of the wheelset is an indispensable part of the vibration fatigue life analysis and rapid fault detection of the high-speed rail wheels. However, the vibration test time under the actual load excitation is usually long. Under general experimental conditions, that is hard to simulate the long-term vibration of the high-speed rail wheels. Therefore, the accelerated vibration test method is used to abbreviate the experiment time, enhance the experiment efficiency, decrease the experiment cost, and quickly obtain the vibration fatigue characteristics of the wheels [2].

At present, in terms of early damage detection of the high-speed rail wheels, metal magnetic memory detection technology can quickly and accurately detect macroscopic and microscopic defects of ferromagnetic workpieces without special magnetization devices. Besides, it can also warn of potential failures of wheels [3]. Combining the principle of accelerated vibration and the metal magnetic memory detection method, through the ANSYS dynamics simulation, the magnetic memory signal of the high-speed rail wheels material under long-term vibration was studied, which verified the feasibility of failure early warning and security inspection of Magnetic Memory detection method on high velocity rail wheels.

## 2 Principle of Accelerated Vibration

According to fatigue damage theory, acceleration vibration usually adopts two methods of changing the frequency characteristics of random vibration and artificially increasing the vibration level within a certain range [4]. The same load acts on the workpiece for a long time, the generated stress  $\sigma$  and the number of stress cycles  $N$  [5] under which the structure is broken, the relationship is described in Fig. 1.



**Fig. 1.** Typical  $\sigma$ - $N$  curve of accelerated vibration

The curve in the figure can provide a basis for the specimen to undergo fatigue failure under vibration conditions. It can be seen from the a-b section of Fig. 1. that when the stress on the vibration test piece is low, the more failure cycles can be experienced. Conversely, when the stress experienced by the vibration test piece is high, the number of failure cycles that can be experienced is less, that is,

the shorter the vibration test time required. Replacing the above  $\sigma$ -N relationship with the A-N relationship can be used in the actual accelerated vibration test [6]. The recommended empirical formula for the relationship between A-N is:

$$\frac{N}{N_{e1}} = \left( \frac{A}{A_{e1}} \right)^{-k} \quad (1)$$

Where N is the number of failure cycles at any point on the curve. A is the corresponding vibration acceleration.  $N_{e1}$  is the number of failure cycles at the turning point of the curve.  $A_{e1}$  is the corresponding vibration acceleration. K is a constant greater than zero.

The above formula shows that if the original vibration acceleration and vibration frequency produce a kind of fatigue damage, fatigue damage will also occur after increasing the vibration acceleration and reducing the vibration frequency according to the A-N curve. This can ensure the equivalence of accelerated and non-accelerated tests.

If the vibration time is shortened to  $\frac{1}{10}$ , that is,  $\frac{N}{N_{e1}} = \frac{1}{10}$ , according to formula (2):

$$\alpha = \frac{A}{A_{e1}} = \text{arc lg} \left( \frac{1}{K} \right) \quad (2)$$

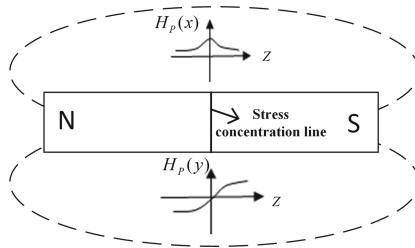
From the above formula, to shorten the vibration time to  $\frac{1}{10}$  of the original, the test value is increased to the original  $\alpha$  times. According to the relationship between sinusoidal vibration acceleration, amplitude and frequency, it can be obtained that the vibration acceleration and amplitude have a positive correlation. When the vibration frequency is constant, the vibrancy rate and vibrancy time are negatively correlated. When the vibrancy amplitude is constant, the vibration frequency and the vibration time have a negative correlation. Therefore, the method of increasing vibration acceleration or vibration frequency is used to shorten the vibration time.

### 3 Metal Magnetic Memory Method

Under the effect of external load, the iron-made components in the geomagnetic environment will undergo orientation and irreversible reorientation of the magnetic domain structure with magnetostrictive properties. The self-magnetization and residual magnetism of ferromagnetic materials are directly related to mechanical stress. This characteristic was called magnetomechanical effect. The magnetomechanical effect makes the magnetic field on the surface of the ferromagnetic metal workpiece stress area strengthen, and the enhanced magnetic field “memorizes” the location of stress concentration of the component, which is the magnetic memory effect.

The theory of metal magnetic memory detection is shown in Fig. 2. In the pressure concentration area, tangent direction weight of leakage magnetic field  $H_p^X$  has maximum value, and the normal direction weight  $H_p^Y$  transforms sign and has the zero value [7]. Through the external load is removed, the irreversible

change of the magnetic domain state will continue. It can determine the stress concentration area and realize the early diagnosis of micro cracks by metering the normal direction component of leakage magnetic field.



**Fig. 2.** Schematic diagram of magnetic memory detection

In the environmental geomagnetic field where the vibrating specimen is located, comprehensively considering the active magnetic field on the vibrating specimen and the external stress on the vibrating specimen, the metal magnetic memory technology is used to detect the change of the magnetic memory signal on the surface of the vibrating specimen, and combined. Regarding the theory of magnetic memory detection technology, through the combination of experiment and theory, the following correspondence formula can be obtained:

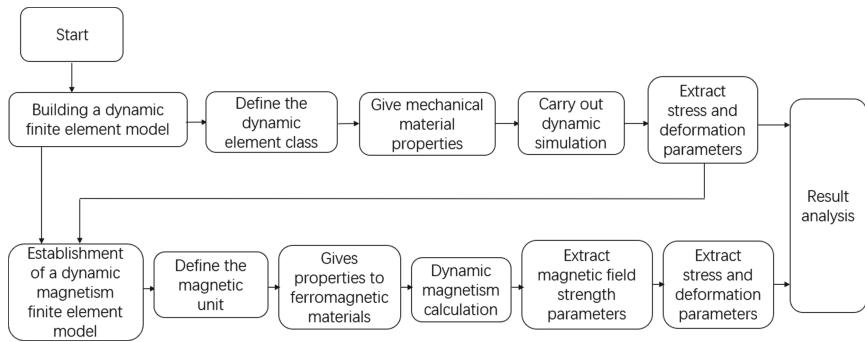
$$H_\sigma = \lambda^H \Delta\sigma / \mu_0 + H + M_\alpha \quad (3)$$

In this formula,  $H_\sigma$  is magnetic field on component.  $\lambda^H$  is magnetoelastic constant component.  $\Delta\sigma$  is partial stress of the workpiece.  $\mu_0$  is vacuum permeability.  $H$  is ambient magnetic field where the component is located,  $M_\alpha$  represents the exchange between magnetic domains [8].

#### 4 ANSYS Model Simulation Analysis

The vibration of 25CrMo4 test plate is simulated in the magnetic field environment by using the finite element analysis software ANSYS, and the magnetic field intensity corresponding to each condition is obtained. The simulation results are analyzed and processed by different numerical calculation methods, and the corresponding relation between different numerical calculation amount and different vibration time is obtained. The finite element simulation under the combined action of vibration and magnetic field is divided into two analysis processes: structural dynamics analysis and vibration magnetic analysis. The process is described in Fig. 3.

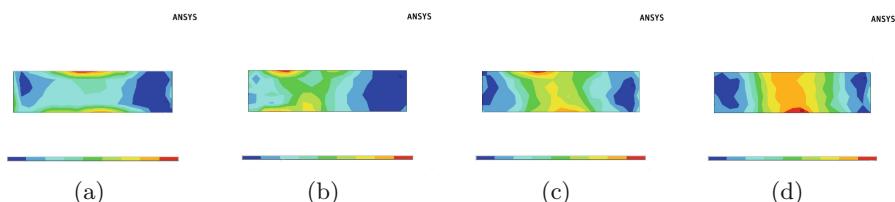
The 25CrMo4 alloy structural steel plate model was established by ANSYS finite element simulation software, and its parameters were as follows: length 195 mm, width 50 mm, height 1.5 mm. This material is the high-speed rail



**Fig. 3.** Finite element simulation flow

wheels material of CRH3 EMU, that has fine machinability, weldability and hardenability. That mostly used to manufacture wheels and axle parts.

In dynamic simulation, including solid185 and 8-node hexahedral element, and sinusoidal load with frequency 45 Hz was first applied to the lower surface of the vibrancy with metal experimental sheet. According to principle of accelerated vibration, vibration of 800d, 1200d, 1600d, 2000d time is equivalently shortened to one thousandth of the original. Set the vibrancy time history to 19.2 h, 28.8 h, 38.4 h, and 48 h respectively. It performs transient dynamics calculation with time history. Extract the stress and strain values of all nodes on the lower surface of the plate, read the node data after the plate structure changes and obtain the modified plate entity, and select solid97, 8-node hexahedral element. Vertical steady magnetic field boundary condition applying to the surface of whole workpiece. The magnetic field strength is 39.8 A/m, which places the simulated entity in a geomagnetic environment. The intensity distribution of the magnetic field combined vibrancy time of 800d, 1200d, 1600d and 2000d, the magnetic field after acceleration vibration is displayed in Fig. 4.



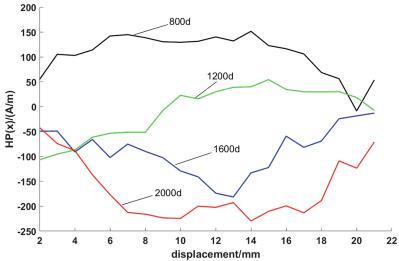
**Fig. 4.** Vibration simulation results

It can be seen from magnetic flux density distribution of four different vibrance times as shown above figure. With combined action of vibrance and terrestrial magnetic field, a significant change in the magnetic field strength signal appears in the middle of the vibration test plate. This change trend is in

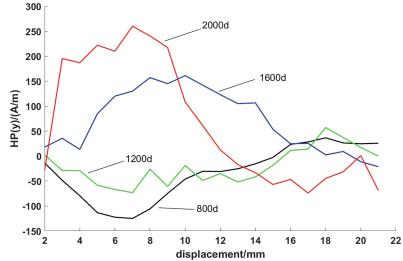
line with change rule that the pressure at middle position of vibration test plate first increases. With the vibration time from 800d to 2000d, the area where the stress increases in the middle position of the vibration test board gradually increases and the color gradually deepens. So as to more clearly represent the change rule of this metal magnetic memory semaphore, the magnetic memory signal are collected the alignment of unit panel point. And characteristic values are extracted strength signal are drawn in Table 1. The trends of magnetic field power tangential and normal values are displayed in Fig. 5 and Fig. 6.

**Table 1.** Characteristic values of metal magnetic memory simulation signals in different vibrancy times

Time (d)	Tangential maximum (A/m)	Tangential minimum (A/m)	Tangential mean (A/m)	Normal maximum (A/m)	Normal minimum (A/m)	Normal average value (A/m)
800	151	-8.5	104.5	36	-124	-32
1200	54	-106	-10	57	-73	-20
1600	-12	-181	-85	330	-74	40
2000	-36	-212	-128	420	-24	128



**Fig. 5.** Tangential component of magnetic field strength



**Fig. 6.** Normal component of magnetic field strength

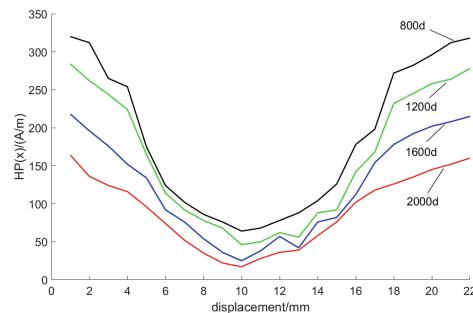
Under the condition of maintaining a fixed vibrancy frequency, with vibrancy time of specimen increases, the tangent direction and normal direction components of metal magnetic memory value of the specimen transform greatly. The tendency is more evident. While vibrancy time increases between 800d through 1200d, that tangent direction weight of magnetic memory value changes between  $-8.5 \text{ A/m} \sim 151 \text{ A/m}$  through  $-106 \text{ A/m} \sim 54 \text{ A/m}$ . Furthermore, the average value reduces from  $104.5 \text{ A/m}$  through  $-10 \text{ A/m}$ . The normal direction component in magnetic memory value changes between  $-124 \text{ A/m} \sim 36 \text{ A/m}$  to  $-73 \text{ A/m} \sim -20 \text{ A/m}$ . The average value increases from  $-32 \text{ A/m}$  to  $-20 \text{ A/m}$ . With vibrancy time raises from 1200d through 1600d, range the tangent direction weight of magnetic memory value transforms from  $-106 \text{ A/m} \sim 54 \text{ A/m}$  to  $-181 \text{ A/m} \sim -12 \text{ A/m}$ .

$A/m \sim -12 A/m$ . Besides, the average value descends from  $-10 A/m$  through  $-85 A/m$ . The normal direction weight of magnetic memory value changes between  $-73 A/m \sim 57 A/m$  to  $-74 A/m \sim -330 A/m$ , the average value raises from  $-20 A/m$  to  $40 A/m$ . When vibrancy time rises by 2000d, the range of the tangent direction component of metal magnetic memory vaule changes from  $-181 A/m \sim -12 A/m$  to  $-212 A/m \sim -36 A/m$ , and the average value decreases from  $-85 A/m$  to  $-128 A/m$ . The normal component of the magnetic memory signal changes from  $-74 A/m \sim -330 A/m$  to  $-24 A/m \sim -420 A/m$ . The average value raises from  $-40 A/m$  through  $128 A/m$ . The above three different time periods have the same change trend, while maintaining a fixed vibrancy frequency, vibrations of different time periods have an effect on the tangent direction and normal direction components of metal magnetic memory signal on workpiece surface, with vibrancy time grows, normal component and tangential component show opposite trends.

To analysing the influence of metal magnetic memory value specifically, calculated the modulus value on strength vector of magnetic field. Experimental results are shown in the following table and figure.

**Table 2.** Leakage magnetic field vector modulus

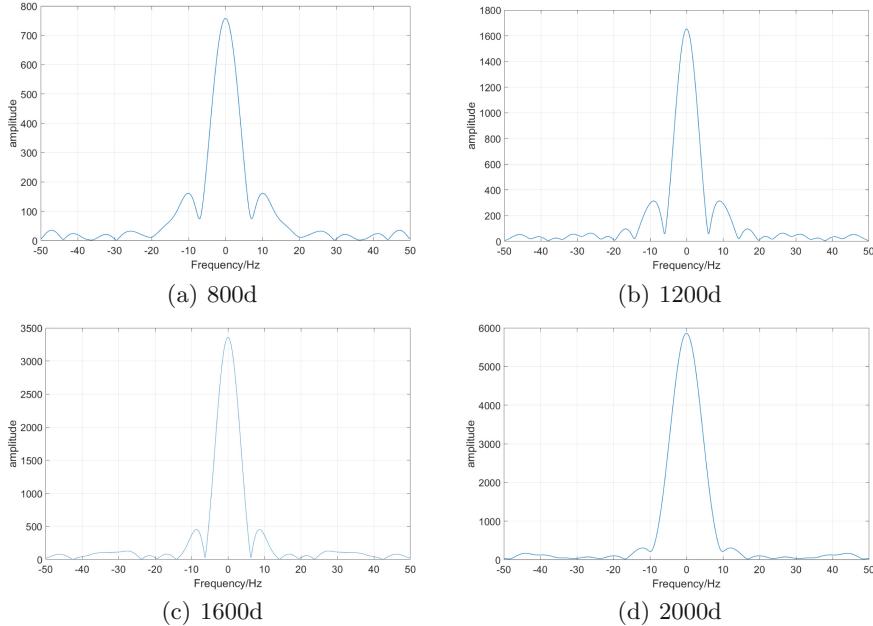
Time (d)	Maximum (A/m)	Minimum (A/m)	Mean (A/m)
800	320	64	214
1200	284	46	153
1600	218	25	112
2000	164	17	76



**Fig. 7.** Leakage magnetic field vector modulus

When the vibration time changes from 800d to 2000d, the magnetic field strength vector value gradually decreases from  $214 A/m$  to  $76 A/m$ , reflecting the phenomenon that the magnetic field strength of vibrating frequency spectrum gradually reduces as the raise of the vibrating time minimal theory.

To analysing the impact with four different vibration times on the energy distribution of metal magnetic memory value specificly, the spectrum acquired from simulation data is transformed to obtain four distinct vibrating time magnetic memory value spectrum diagrams, as display in Fig. 8.



**Fig. 8.** Spectrum analysis

The value of low-frequency signal are 0 Hz to 15 Hz. With the increasing of vibration time, the primary frequency weights of low frequency domain are reduced in bandwidth. The frequency range reduced between 0 Hz ~15 Hz through 0 Hz~10 Hz, the peak value of metal magnetic memory value keeps increasing. Furthermore, the phenomenon of energy that magnetic memory value centralized with vibrating time increased are more evident. The influence of high frequency noise values on energy distribution of metal magnetic memory value are gradually diminished. And four different vibration times have little effect in frequency spectrogram components of metal magnetic memory value, while have remarkable impact in its power distribution. Through the accelerated vibration simulation analysis of the 25CrMo4 material, the obtained magnetic memory values changes trend as vibrating time result regular conclusions.

## 5 Conclusion

Through the accelerated vibration simulation analysis of 25CrMo4 material, it is known that when the vibration time is 800d~2000d, the change trend of the obtained magnetic memory signal as follows.

- 1) The vibrancy frequency is constant, and the maximum change range of the tangential component of the magnetic memory signal is -212~151A/m. The range of average value is -128~-104.5 A/m. The maximum value of the normal magnetic memory signal varies within -124~420A/m. The average change range is -32~128 A/m, showing a trend that the tangential magnetic memory signal gradually decreases and the normal direction gradually increases.
- 2) After frequency spectrum transformation, the low-frequency target signals are all in the range 0 Hz~15Hz. As the vibration time increases, the influence of high-frequency noise signals on the energy distribution of the magnetic memory signal gradually decreases.
- 3) The vector value of the magnetic field intensity gradually decreased from 214A/m to 76A/m, reflecting that the magnetic field intensity of the test piece gradually decreased with the increase of vibrancy time.

According accelerated vibration simulation analysis of 25CrMo4 material, the change trend of the magnetic memory value with the longer vibrancy time is obtained. It is proved that application of magnetic memory detection technology in the actual vibration state detection of high-speed rail wheels has certain research value.

**Acknowledgment.** This research work was supported by the National Natural Science Foundation of China under Grant No.51405303, the Special Fund for the Selection and Training of Excellent Young Teachers in Shanghai Universities (ZZyy15110), and the Fund for the Development of Science and Technology Talents for Young and Middle-aged Teachers of Shanghai Institute of Technology (ZQ2019-21).

## References

1. Chao, C.: Analysis of the influence of wheel profile wear on wheel-rail contact characteristics and vibration characteristics of tracks and bridges. East China Jiaotong University (2018). 2.1018.818669
2. Zhang, F., Zhou, L., Jiang, J.: Design of random vibration fatigue acceleration test based on frequency domain method. Vibration, Testing and Diagnosis **39**(4), 306–313 (2016)
3. Wang, W., Yi, S., Su, S.: Research status and key issues of metal magnetic memory nondestructive testing. China Journal of Highway and Transport (2019)
4. Lu, Z., Zhu, X., Wei, G.: Accelerated test research of PCB board vibration based on fatigue cumulative damage equivalent theory. Equ. Env. Eng. **015**(003), 53–56 (2018)
5. Wang, Q., Guan, D.: Accelerated Fatigue Test Method for Auto Parts. Autom. Technol. **000**(011), 14–17 (1997)

6. Jiang, Y., Tao, J., Chen, X.: Research on Super Gaussian Random Vibration Accelerated Test Model. *Vibration and Shock* (2017). <https://doi.org/10.13465/j.cnki.jvs.2017.09.038>
7. Dubov, A., Kolokolnikov, S.: The metal magnetic memory method application for online monitoring of damage development in steel pipes and welded joints specimens. *Weld World* **57**(1), 123–136 (2013)
8. Wen, W., Sa, S.: The mechanism and realization of metal magnetic memory detection. *J. Northern Jiaotong University* (2002). <https://doi.org/10.3969/j.issn.1673-0291.2002.04.017>



# Automated Prediction of Cervical Precancer Based on Deep Learning

Bing Zhang<sup>1,2</sup>, Qingyuan Zhang<sup>3</sup>, Hao Zhou<sup>1,2</sup>, Chengyi Xia<sup>1,2(✉)</sup>,  
and Juan Wang<sup>3</sup>

<sup>1</sup> Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300384, People's Republic of China  
[cxyia@email.tjut.edu.cn](mailto:cxyia@email.tjut.edu.cn), [xialooking@163.com](mailto:xialooking@163.com)

<sup>2</sup> Key Laboratory of Computer Vision and System (Ministry of Education), Tianjin, University of Technology, Tianjin 300384, People's Republic of China

<sup>3</sup> School of Electrical and Electronic Engineering, Tianjin University of Technology, Tianjin 300384, People's Republic of China

**Abstract.** Cervical cancer is one of the most common malignant ones among women worldwide. Visual examination of the cervix using acetic acid or Lugol's iodine has been widely adopted in cervical cancer screening. However, there also exist problems of high misdiagnosis and low diagnostic efficiency during the manual screening of cervical images. Here, we present a new method for automatic cervical intraepithelial neoplasia (CIN) classification combining acetic acid and Lugol's iodine cervigrams. We take use of convolutional neural network (CNN) and bi-directional long short-term memory (BiLSTM) to extract the maximum acetic acid images' features and temporal characteristics, and then collect spatial features for Lugol's iodine cervigrams. Meanwhile, we design a feature fusion module to combine features of different dimensions in cervigrams to further enhance the predictive performance. The experimental results indicated that the current method achieves an accuracy of 96.04%, which is superior to previous related works and those observed by clinicians. Our work will help to enhance the early detection of cervical cancers.

**Keywords:** Cervical cancers · Early detection · Computer-aided diagnosis · Deep learning

## 1 Introduction

According to the statistics in 2018, there were almost 570,000 cases and 311,000 deaths of cervical cancer [1], and the cervical cancer has become one of four types of most common cancer and main lethal reasons of cancer in women. In particular, the cervical cancer continues to rank the top two deadly causes of cancer for women from 20 to 39 years old, leading to 9 deaths per week in this group [2].

Vaccination and screening have played an important role in reducing the burden of cervical cancer, especially in countries with fewer resources [3]. In the

clinical practice, colposcopy is a widely used early screening method because of its low cost and feasibility, which mainly checks the characteristics of cervical epithelium after continuously applying the normal saline, 3-5% dilute acetic acid and Lugol's iodine solution [4]. Adding 3-5% acetic acid to cervical epithelial cells causes the white appearance of epithelial cells in CIN and early stage cancer, and this reaction is termed as acetowhitening and may last for 2-3 min. The iodine solution will make CIN and invasive cancers present a thick mustard yellow area. Based on the data from the World Health Organization (WHO) [5], CIN is usually classified into three-different levels or grades: CIN1 (mild), CIN2 (moderate) and CIN3 (severe). However, due to the complexity of cancer and individual differences, manual screening is very dependent on the diagnostic experience of clinical experts, which leads to the low efficiency of colposcopy and high misdiagnosis rate.

To this end, many researchers devote to solving the aforementioned problems. At present, researches on the classification of cervical cancer can be briefly classified into two main types: the first one is the traditional learning-based method (e.g., k-Nearest Neighbors, decision tree, random forest), which utilizes some manually extracted features, such as color histogram, gradient and texture information [6-8]; the second one is the method based on deep learning (e.g., CNN) [9,10]. Compared with some manually extracted features, CNN features can be automatically picked up by a certain number of images, some of which may be neglected by the manual extraction. Current researches on the deep learning methods for cervix imaging predominantly focused on visual inspection with acetic acid (VIA) images [11], but these methods will be limited in terms of sensitivity and specificity [12]. Clinically, it is necessary to take use of VIA and visual inspection with Lugol's iodine (VILI) to detect cervical abnormalities, and some researchers have also proposed algorithms for categorizing these two kinds of images. As an example, in Ref. [10], the authors classified cervical images as CIN+ with an accuracy of 80.0% and their method of manually extracting the features of VIA image may lose some information of image grading judgment, and thus the classification results are not enough since only two grades can be obtained. In addition, Ref. [13] proposed the CRCNN, which only extracted the information of forwarding codes for VIA and ignored the feature of bidirectional codes for sequential images.

In the current work, based on the deep learning method, we develop a novel framework to further improve the computer aided diagnosis for cervical cancers. In this method, the image features of VIA and VILI can be automatically learned, and the cervigram can be categorized into five different levels: normal, CIN1, CIN2/3 and cancer, which is also the most frequently used in clinical practice. Furthermore, we have not only verified the temporal features of VIA, but also further experimentally demonstrated that it has bidirectional encoding information. Finally, we design a specific software module that combines the features extracted from VIA and VILI to get the final prediction grades, evaluates the performance of the proposed framework and analyzes experimental results.

The remaining parts of the paper are structured as follows. At first, Sect. 2 describes the dataset and overall framework of proposed method. Next, Sect. 3

presents the results of related experiments based on the CNN. Finally, in Sect. 4, we summarize the paper with some concluding remarks and future works.

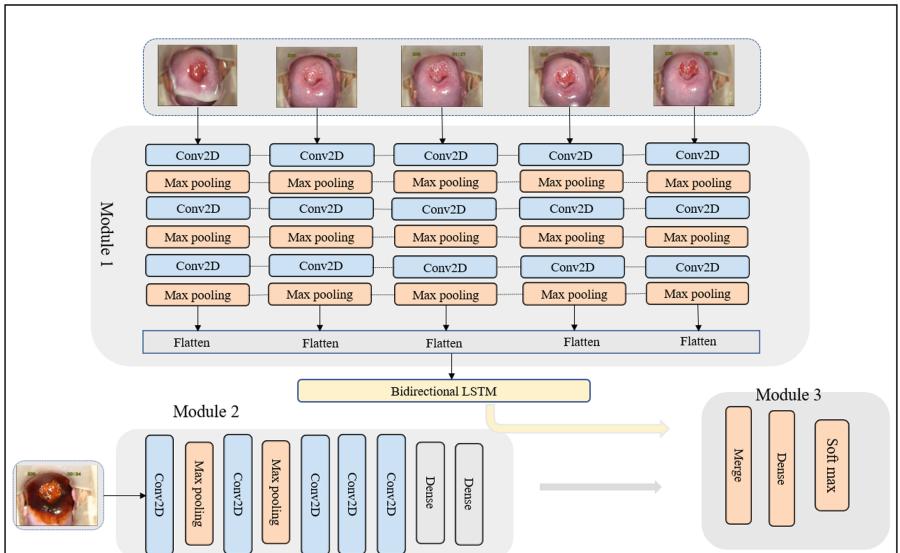
## 2 Methods and Materials

### 2.1 Dataset

All data used in the experiments were provided by the First Affiliated Hospital of University of Science and Technology at Hefei, China, which include 262 normal cases, 133 CIN1 cases, 183 CIN2/3 cases and 60 cancer cases. In this dataset, there are a total of 3828 images, all of which were marked by clinically experienced experts. Each case consisted of 6 cervigrams, five of which were consecutive images obtained every 30s by colposcopy in acetic acid white test and one of which was VILI image.

### 2.2 Overall Framework

Figure 1 depicts the overall structure of our proposed framework, which is composed of three main modules: a) CNN-BiLSTM module, which is adapted to acquire the VIA images' feature like bidirectional sequence feature at utmost; b) Alexnet adaptation module [13], which is used to draw the VILI spatial features of images; c) The main function, which is applied to fuse the features extracted from VIA and VILI and obtain the diagnosis results, namely, normal, CIN1, CIN2/3 or cancer.



**Fig. 1.** The structure of our proposed framework.

### 2.3 Via Feature Extraction

**Spatial features:** in order to extract spatial features effectively, we adopt the network structure of three-layered convolution layer and three-layered pooling layer, as shown in Table 1. The pooling layer is the maximum pooling, and the activation function is the ReLU function. The data input into CNN have the 6404803 pixels, where 640480 is the size of VIA and 3 denotes the number of the color channels of cervigram. Finally, the last flattened layer of the CNN network flattens the feature map into various vectors  $y_i$ .

**Table 1.** The parameters of CNN.

Layer	Kernel Size	Stride	Channel
Conv1	9*9	4	64
Max pooling1	3*3	2	64
Conv2	5*5	1	128
Max pooling2	3*3	2	128
Conv3	3*3	2	256
Max pooling3	3*3	2	256
Flatten	-	-	-

**Bidirectional sequence feature:** It is found that the five sequential VIA cervigrams obtained by acetowhite test, which is a gradual process, had temporal characteristics. Although the LSTM model can better capture the longer distance dependencies, BiLSTM can capture bidirectional dependencies [14]. We input the vector sequence  $\{y_1, y_5\}$  from the previous step into BiLSTM to obtain the corresponding feature vector  $z_1$ . In this module, we used 512 BiLSTM units.

BiLSTM is composed of the forward LSTM and backward LSTM and its structure at time step  $t$  is pictured in Fig. 2. Among them, the hidden layer of BiLSTM holds two values:  $S$  and  $S'$ , in which  $S$  participates in the forward calculation and  $S'$  in the reverse calculation.

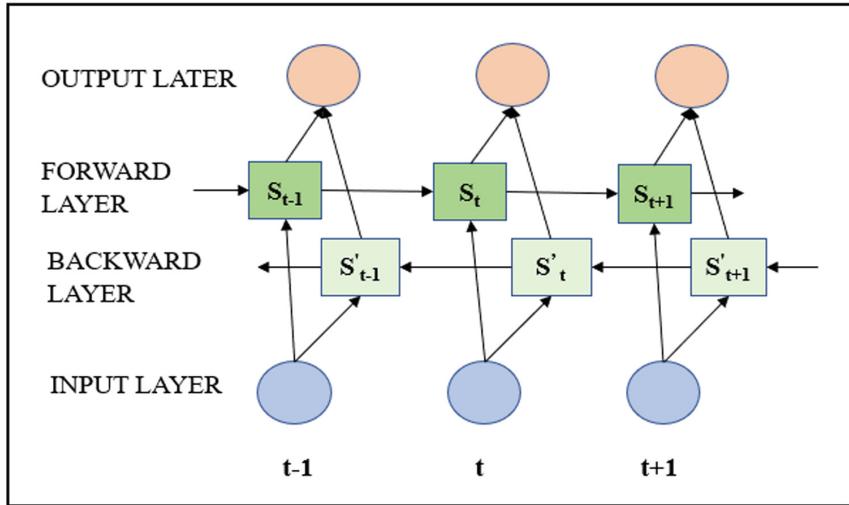
The calculation formula is as follows:

$$o_t = g(VSt + V'S') \quad (1)$$

$$S_t = f(UX_t + WS_{t-1}) \quad (2)$$

$$S'_t = f(U'X_t + W'S'_{t+1}) \quad (3)$$

From Eqs. (1) to (3),  $U$  and  $U'$ ,  $W$  and  $W'$ ,  $V$  and  $V'$  are all different weight matrices. The activation functions are  $g$  and  $f$ .  $X_t$  and  $o_t$  are the inputs and outputs of the network. Furthermore,  $S_t$  represents the hidden layer's value in the forward calculation at time step  $t$ , and  $S'_t$  is the hidden layer's value in the reverse calculation at time step  $t$ .



**Fig. 2.** The structure of Bidirectional LSTM.

## 2.4 VILI Feature Extraction

Unlike the CNN-BiLSTM network in VIA, we take use of a deeper improved version of Alexnet [13] to collect the spatial features of VILI. We input the VILI cervigrams into Alexnet networks to get the space vector  $z_2$  for the next step and optimizer adopted is adam.

## 2.5 Combining Features of via and VILI

In order to fuse the features extracted from VIA and VILI, we design a feature fusion module, where we combine the vector  $z_1, z_2$  obtained in the above two steps and train the predicted grading results from the softmax layer in the three-layered neural network. Meanwhile, the activation function adopts the ReLU form in the current setup.

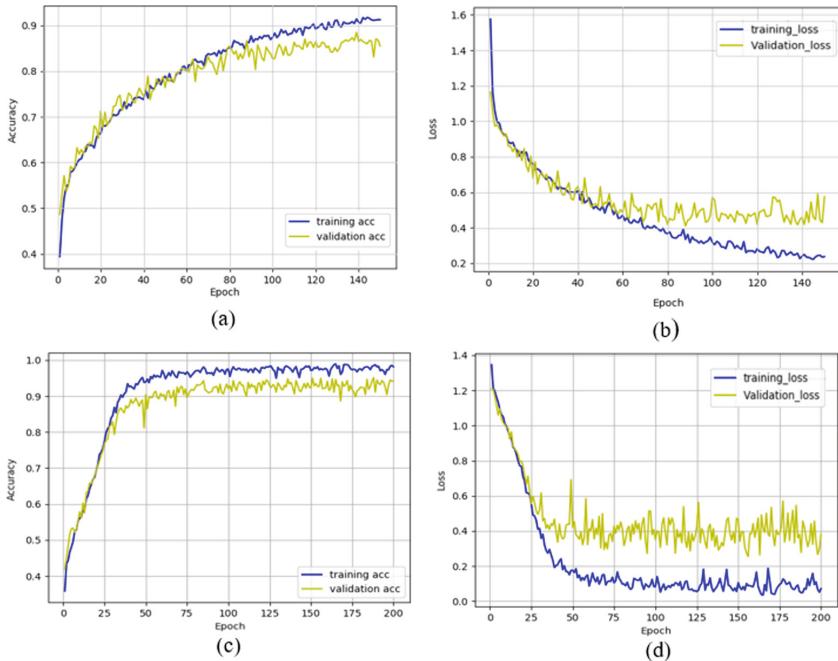
In this paper, all the experiments were run on a workstation installing the Windows 10 operating system, the related algorithms are programmed with Python. The hardware environment of the workstation is Intel Core @3.60 GHz CPU, 4 pieces of NVIDIA GeForce RTX 2080 Ti graphics card and 64 GB RAM.

## 3 Experimental Results

The experimental dataset has been described in Sect. 2, and the size of each image contains 640480 pixels. During the training process, we adopt some common methods of deep learning to train the model, such as data augmentation technology, dropout and so on. The augmented dataset contains 19,120 cervigrams through data augmentation technology like rotation, horizontal flip.

In addition, the loss function used in our experiment is the cross entropy loss function. The accuracy is defined as the ratio of the number of correct samples predicted by the classification model to the total number of samples.

Figure 3 depicts the accuracy and loss rate as a function of epoch for the current setup. Among them, Fig. 3 (a) and (b) show the accuracy and loss obtained from validation and training data set, during which VIA is trained with CNN, while Fig. 3 (c) and (d) present the accuracy and loss curves, where the network training VILI is assumed to be Alexnet. From Fig. 3, we can find that the loss and accuracy curves of the VIA's CNN tend to be stable after around 120 epochs, where the accuracy of prediction is about 85.50%. Meanwhile, from Fig. 3 (c) and Fig. 3 (d), it can be observed that the curve of Alexnet's accuracy and loss function converges after around 75 epochs with an accuracy up to 91.03%. Therefore, it can be found that the CNN-based method can effectively extract the spatial features of VIA and VILI images.

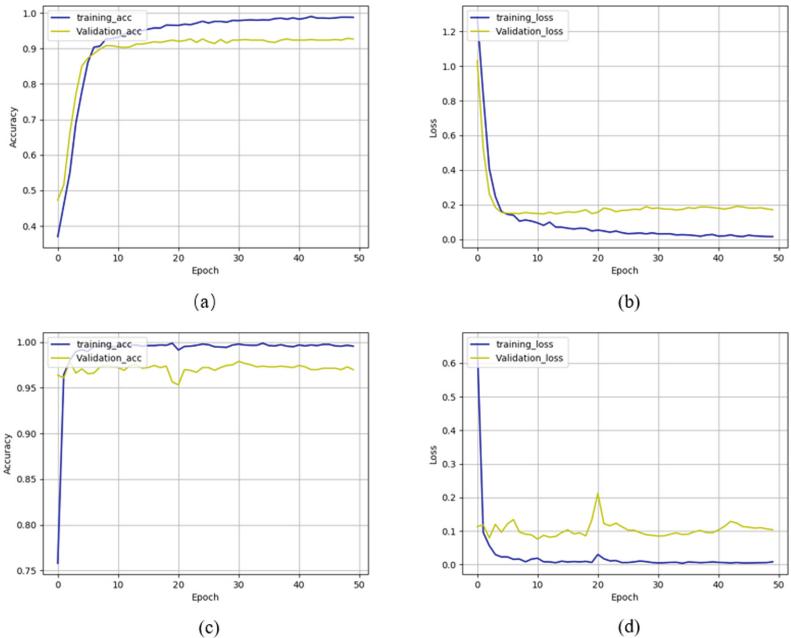


**Fig. 3.** The loss-value and accuracy curves under training set and validation set. (a) The CNN accuracy curves of the date set of VIA images. (b) The CNN loss curves of the date set of VIA images. (c) The Alexnet accuracy curves of the date set of VILI images. (d) The Alexnet loss curves of the date set of VILI images.

As a further step, Fig. 4 provides the accuracy and loss curves of extracting VIA's sequential features using LSTM and BiLSTM under the same conditions. By comparing Fig. 3 (a) and Fig. 4 (a), we can find that the VIA image has

features of forwarding codes. From Fig. 4 (a) and Fig. 4 (c), the accuracy curve finally converges to 91.03% , and at the same time, we can find that the curve of accuracy of CNN-BiLSTM of VIA converges to 95.02%. And furthermore, it is suggested that the VIA images have the bidirectionally encoding information and the features extracted by BiLSTM can better help to determine the CIN grades.

Finally, Table 2 shows the accuracy of each part of our proposed network and other works in the recent years. In Table 2, it can be found that the accuracy rate of the whole network is up to 96.02%, which is higher than that in previous stages, since we designed a feature fusion and merging module in the end to further synthesize the features of extracted VIA and VILI images. Through the large-scale experiments, we can find the necessity of two kinds of images to improve the classification results of cervical cancer and the existence of bidirectional encoding information in the VIA sequence images.



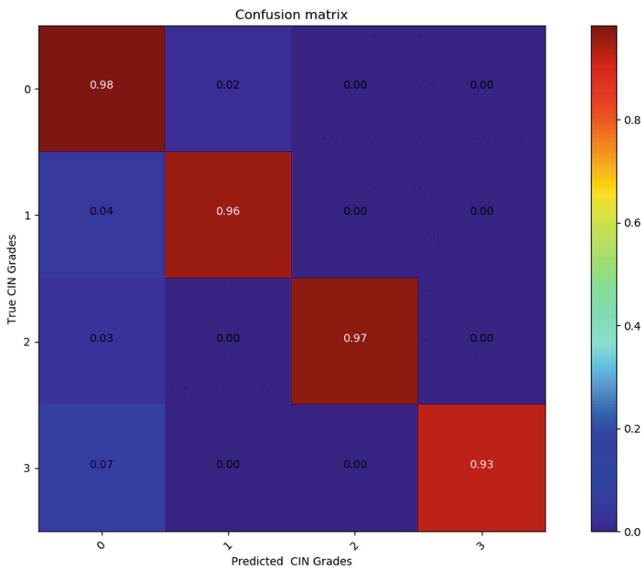
**Fig. 4.** The loss-value and accuracy curves under the training set and validation set. (a) The CNN-LSTM accuracy curves of the date set of VIA images. (b) The CNN-LSTM loss curves of the date set of VIA images. (c) The CNN-BiLSTM accuracy curves of the date set of VIA images. (d) The CNN-BiLSTM loss curves of the date set of VIA images.

In order to further evaluate the performance of predictive classification model, the confusion matrix between the prediction and actual result is provided in Fig. 5, in which each column denotes the predicted classification result,

**Table 2.** The Networks Accuracy of each part and other works.

Network	Accuracy
CNN of VIA in our study	85.50%
Alexnet of VILI in our study	91.03%
CNN-LSTM of VIA in our study	92.51%
CNN-BiLSTM of VIA in our study	95.2%
Asiedu, M et al. [10]	80.0%
CNN-LSTM for acetic acid test [13]	94.21%
This study	96.04%

and each row represents the actual grade on the label. The values of 0, 1, 2, 3 on the coordinate axis represent CIN grades, which include normal, CIN1, CIN2/3, and cancer, respectively. As an example, it can be observed in Fig. 5 that, if the actual sample is a normal one, the accuracy of being predicted as a normal sample is 98% and the actual sample is incorrectly predicted as CIN1 is 2%, and the related prediction accuracy for other grades can be found in the other rows. Finally, we can find that our proposed method has a better classification result for cervigrams of normal grade with an accuracy of 98% and a lower classification accuracy for cancer grade with an accuracy of 93%.

**Fig. 5.** The confusion matrix of our overall network structure.

## 4 Conclusion

In this work, based on the deep learning method, we proposed a new framework for automatically extracting features of cervigrams of VIA and VILI images and classifying CIN grades. Extensive experiments also demonstrate that the VIA sequence image is not only sequential, but also holds the bidirectionally encoding information. Meanwhile, as far as the prediction performance is concerned, our framework is superior than those obtained by the experienced clinicians and several recently proposed methods. In general, our work can, to a large extent, assist clinicians to perform the early diagnosis of cervical cancer and further contribute to carry out the computer-assisted diagnosis.

In the future, we will use the image processing methods to further automatically perform the early detection and screening of cervical cancers, such as the removal of reflectors, and integrating the manual feature extraction into the deep learning framework, and even construct the knowledgeable database to help the experts to quickly and efficiently diagnose the cervical cancer.

## References

- Bray, F., Ferlay, J., Soerjomataram, I., et al.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *J. CA: A Cancer Journal for Clinicians*. **68**(6), 394–424 (2018)
- Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics. *J. CA: A Cancer J. Clinicians*. **69**(1), 7–34 (2019)
- Torre, L.A., Siegel, R.L., Ward, E.M., Jemal, A.: Global cancer incidence and mortality rates and trends—an update. *Cancer Epidemiol. Biomark. Prev.* **25**(1), 16–27 (2016)
- Sellors, J.W., Sankaranarayanan, R.: Colposcopy and Treatment of Cervical Intraepithelial Neoplasia: A Beginners' Manual. Diamond Pocket, India (2003)
- WHO, ICO.: Human papillomavirus and related diseases report C WORLD. HPV Information Centre. Barcelona, Albania (2014)
- Liu, J., Peng, Y., Zhang, Y.: A fuzzy reasoning model for cervical intraepithelial neoplasia classification using temporal grayscale change and textures of cervical images during acetic acid tests. *J. IEEE Access*. **7**, 13536–13545 (2019)
- Ji, Q., Engel, J., Craine, E.: Texture analysis for classification of cervix lesions. *J. IEEE Trans. Med. Imag.* **19**(11), 1144–1149 (2000)
- Xue, Z., Antani, S., Long, L.R., Thoma, G.R.: November. An online segmentation tool for cervicographic image analysis. In: Proceedings of the 1st ACM International Health Informatics Symposium pp. 425–429 (2010)
- Xue, Z., Antani, S., Long, L.R., Jeronimo, J., Thoma, G.R.: March. Comparative performance analysis of cervix ROI extraction and specular reflection removal algorithms for uterine cervix image analysis. In: Medical Imaging 2007: Image Processing, International Society for Optics and Photonics Vol. 6512, p. 65124I (2007)
- Asiedu, M.N., Simhal, A., Chaudhary, U., et al.: Development of algorithms for automated detection of cervical pre-cancers with a low-cost, point-of-care, pocket colposcope. *J. IEEE Trans. Biomed. Eng.* **66**, 2306–2318 (2019)

11. Marquez-Grajales, A., Acosta-Mesa, H.G., Mezura-Montes, E., et al.: Cervical image segmentation using active contours and evolutionary programming over temporary acetowhite patterns. In: 2016 IEEE Congress on Evolutionary Computation (CEC), pp. 3863–3870. IEEE (2016)
12. Catarino, R., Schäfer, S., et al.: Accuracy of combinations of visual inspection using acetic acid or lugol iodine to detect cervical precancer: a meta-analysis. *J. Bjoг Int. J. Obstetrics Gynaecol.* **125**, 545–553 (2018)
13. Yue, Z.J., Ding, S., Zhao, W.D., et al.: Automatic CIN grades prediction of sequential cervigram image using LSTM with multistate CNN features. *IEEE J. Biomed. Health Inform. Inf.* **24**(3), 844–854 (2020)
14. Ullah, A., Ahmad, J., Muhammad, K., et al.: Action recognition in video sequences using deep Bi-directional LSTM with CNN features. *IEEE Access*. **99**, 1 (2017)



# Dynamic Economic Dispatch Considering Wind Based on Adaptive Crisscross Optimization

Panpan Mei, Lianghong Wu<sup>(✉)</sup>, Hongqiang Zhang, and Zhenzu Liu

Hunan University of Science and Technology, Xiangtan 411100, Hunan, China  
lhwu@hnu.edu.cn

**Abstract.** Due to the characteristic of randomness and volatility of the wind, broad-scale wind farm connected makes the economic dispatch of power systems more complicated. Since the crisscross optimization algorithm has no mutation operation, it is simple comparatively for the evolution method, to avoid the deficiency of the unitary evolution operation, the horizontal crossover strategy is improved. It is no longer the two paired individuals, but the two paired individuals and the global optimal individual implement horizontal crossover operation. And the adaptive strategy of parameter with learning capability is put forward to better balance the global exploration and local exploitation. At the same time, a heuristic constraint processing method is proposed to effectively dealing with the constraints. A classic 5-machine system is proposed to testify the raised algorithm, the simulation results reveal that the algorithm possesses good convergence performance, moreover, it is an effective strategy for dealing with dynamic economic dispatching considering the wind power integration.

**Keywords:** Wind power · Dynamic economic dispatch · Crisscross optimization algorithm · Parameter adaptive strategy

## 1 Introduction

In general, the Economic Dispatch (ED) is to reasonably allocate the output value of generators under the condition of known system load demand, in order to make the total power generation cost minimal under the premise of satisfying the operating constraint conditions of power system. However, economic dispatch is a typical static optimization problem, which only considers one hour scheduling period [1, 2], therefore the dynamic economic dispatch (DED) problem is proposed, which considers the whole dispatching period. DED is more in line with the actual situation of power system dispatching, but it is also more difficult to deal with [3]. In order to alleviate the problems of energy shortage and environmental pollution, wind power generation technology has developed speedy in recent years, at the same time, the wind power accounts for an increasing proportion in the power grid. However, wind power has strong intermittency

and random fluctuation, which brings new challenges to power system economic dispatch [4,5]. The ED problem with consideration of valve point effect is a complicated optimization problem which has the characteristics of strictly constraint, nonlinear and multi-peak. Over the years, many scholars have proposed a variety of traditional mathematical methods to solve DED problems, including linear programming (LP) [6], dynamic programming (DP) [7], quadratic programming (QP) [8], nonlinear programming (NLP) [9], Lagrange relaxation (LR) [10], etc. However these traditional solving strategies require that these objective functions must be derivable and defined in the convex feasible region [11]. Therefore, traditional mathematical methods are difficult to solve complex nonlinear economic scheduling problems. In order to get over the limitations of traditional strategies, the intelligent optimization algorithm such as differential evolution algorithm(DE) [12], genetic algorithm (GA) [13], harmony search (HS) [14] and artificial swarm optimization algorithm [15] have been applied to cope with complex DED problems [16]. Zhang et al. [17] put forward an improved PSO algorithm to cope with the DED problem and obtained good optimization results, but only one-hour scheduling time was considered in the economic dispatching problem. Liang et al. [18] proposed an improved hybrid bat algorithm to deal with the DED problem, but the effect of valve point was ignored. Elattar et al. [19] put forward a hybrid optimization algorithm of bacterial foraging algorithm and genetic algorithm to cope with the DED problem. Zhang et al. [20] proposed a Stackelberg game model for economic dispatch and the teaching and learning algorithm (TLBO) is applied to deal with the model. Wu et al. [21] put forward a fast adaptive DE algorithm to cope with the DED problem, which improved the robustness of the algorithm. Meng et al. [22] put forward a crisscross optimization algorithm(CSO) to cope with DED problem, however it lacked effective processing methods for complex constraints. Xie et al. [23] proposed a fuzzy model to solve the DED problem considering wind power, but the membership function proposed was relatively simple, which could hardly reflect the real situation and human factors were relatively large. Zaman et al. [24] put forward an improved enhanced differential evolution algorithm (HDE) to cope with the DED problem, and obtained good optimization results, however, the HDE algorithm had too many control parameters which were not easy to adjust. Guo et al. [25] put forward an improved competitive group optimization algorithm (ICSO) to cope with the economic dispatching problem of power system, but there were too many parameters to be adjusted and the treatment of constraint conditions wasn't mentioned.

In this paper, the adaptive crisscross optimization (A-CSO) improved is proposed to cope with the dynamic economic scheduling problem considering wind power. In order to preserve the diversity and guarantee the feasibility of the solutions, a heuristic repaired strategy is put forward to cope with these constraints. For the purpose to prove the effectiveness and superiority of the strategy put forward in this paper, a simulation calculation is carried out on a 5-machine system and contrasted with other heuristic algorithms. The experimental results

display that this strategy is an effective strategy to cope with the dynamic economic dispatch of power system with the consideration of wind power.

## 2 The Proposed Model Considering Wind Power

### 2.1 Objective Function

The goal of DED problem is to make the fuel cost of power system minimal. Wind farms do not need consume fuel [18]. However, when the intake valve of steam turbine is turned on suddenly, the phenomenon of wire drawing would appear. The research shows that ignoring the effect of valve point will have a big influence on the accuracy of solution [4]. When the objective function considers non-linear factors like effect of the valve point, the objective function of fuel cost [24] can be expressed as:

$$F(P_{i,t}) = \sum_{i=1}^N \sum_{t=1}^T \{ a_i p_{i,t}^2 + b_i p_{i,t} + c_i + |e_i \sin [f_i(p_i^{\min} - p_{i,t})]| \} \quad (1)$$

where,  $F$  is the total fuel cost of conventional unit operation during whole dispatching cycle ( $T$  hours).  $T$  is the dispatching period,  $N$  is conventional units number,  $a_i$ ,  $b_i$ ,  $c_i$ ,  $e_i$  and  $f_i$  are the cost coefficients of  $i$ -th unit,  $P_{i,t}$  is the output power of  $i$ -th unit during  $t$ -th hour, and  $P_i^{\min}$  is the minimum output value of  $i$ -th unit.

### 2.2 Constraints

1) The constraint of power balance.

$$\sum_{i=1}^N P_{i,t} + P_{w,t} - P_{loss,t} = PD_t \quad (2)$$

Where  $P_{w,t}$  is the output value of the wind farm at  $t$ -th hour,  $P_{loss,t}$  is the load of system at  $t$ -th hour, that is shown as:

$$P_{loss,t} = \sum_{i=1}^N \sum_{j=1}^N P_{i,t} \times B_{i,j} \times P_{j,i} \quad (3)$$

Where  $B$  is the loss coefficient.

2) The limits of active power

$$P_i^{\min} < P_{it} < P_i^{\max} \quad (4)$$

3) The limits of ramp rate

$$P_{i,t} - P_{i,t-1} < UR_i \quad (5)$$

$$P_{i,t-1} - P_{i,t} < DR_i \quad (6)$$

where  $UR_i$  and  $DR_i$  respectively represent the ramp rate of  $i$ -th unit.

### 2.3 The Model of Wind Turbine Output

The wind power owns the characteristic of random fluctuations, and lots of measurement data demonstrate that the wind speed mostly follows the *Weibull* distribution of two-parameter, shown as:

$$F(v) = 1 - \exp(-(v/c)^k) \quad (7)$$

where  $c$  and  $k$  represent the scale and shape parameter of the *Weibull* distribution respectively;  $v$  represents the wind speed; The probability density of the wind speed could be shown as:

$$f(v) = (k/c) (v/c)^{k-1} \exp\left[-(v/c)^k\right] \quad (8)$$

According to the probability distribution of  $v$ , the relationship between the active output of wind power and wind speed  $v$  can be obtained, thus the active output of wind power could be shown as [26]:

$$P_{w,t}(v_t) = \begin{cases} 0, & (v_t < v_{ci}) \cup (v_t > v_{co}) \\ \frac{P_{rate}(v_t^3 - v_{ci}^3)}{v_R^3 - v_{ci}^3}, & (v_{ci} < v_t < v_R) \\ P_{rate}, & (v_R < v_t < v_{co}) \end{cases} \quad (9)$$

Where  $P_{rate}$  represents the output value of wind turbine power,  $v_{ci}$ ,  $v_R$  and  $v_{co}$  respectively represent cut-into, rated and cut-out wind speed [27].

## 3 The Improved CSO Algorithm

CSO is a new intelligent optimization algorithm proposed in 2014 [22]. The CSO algorithm is constituted of three operators: the horizontal cross operation, vertical cross operation, and selection operation. The horizontal and vertical cross operations perform different crossing operations in opposite directions respectively, and lots of adjustment solutions would be produced at each iteration. The parent population would be updated by these newly generated adjustment solutions through a greedy selection strategy. Only the better individuals can be retained to the next generation.

### 1) Horizontal cross operation:

Horizontal crossing is an arithmetic crossover of two paired individuals for all dimensions in a population. It is assumed that the parent solutions  $X_i$  and  $X_j$  carry out the horizontal crossing of the  $d$ -th dimension, the offspring solutions are produced as:

$$\begin{cases} MS_{hc}(i, d) = r_1 \times X_{i,d} + (1 - r_1) \times X_{j,d} + c_1 \times (X_{i,d} - X_{j,d}) \\ MS_{hc}(j, d) = r_2 \times X_{j,d} + (1 - r_2) \times X_{i,d} + c_2 \times (X_{j,d} - X_{i,d}) \end{cases} \quad (10)$$

Where  $c_1$  and  $c_2$  represent expansion coefficients in  $[-1, 1]$ ,  $r_1$  and  $r_2$  represent random number uniformly distributed in  $[0, 1]$ .  $MS_{hc}(i, d)$  and  $MS_{hc}(j, d)$  respectively represent the offspring of  $X_{i,d}$  and  $X_{j,d}$ .

2) Vertical cross operation:

The vertical crossing carries out the arithmetic crossing between two different dimensions of all solutions. It is assumed that  $d_1$  and  $d_2$  are different dimensions of solution  $X_i$ ,  $MS_{vc}(i, d_1)$  can be generated by Eq. 11.

$$MS_{vc}(i, d_1) = r \times X_{i,d_1} + (1 - r) \times X_{i,d_2} \quad (11)$$

where  $d_1$  and  $d_2 \in N(1,D)$ ,  $i \in N(1,M)$ ,  $r \in U(0,1)$ .

3) Improved horizontal cross operation:

The CSO algorithm has the characteristics of simple principle, easy implementation, and strong search ability. When solving optimization problems, it shows pretty global search capability and fast convergence performance. But, CSO algorithm has no mutation operations, and it is single comparatively for the evolution strategy. When solving complicated optimization problems, it's easy to trapped in the local optimal solution and is difficult to obtain the optimal solution. Therefore an improvement was made to the original horizontal cross operation in this paper: The paired individuals no longer perform horizontal crossover operations with each other, but perform horizontal crossover operations with each other and the global optimal solution simultaneously with a random probability shown as Eq. 12. And a parameter adaptive strategy with learning ability was introduced, a corresponding control parameter  $F$  was generated for each individual in the population. During the iterative process, the control parameters are updated according to Eq. 13. If the offspring's fitness is better than the parent's, the offspring would replace the parent's individual, and the control parameters corresponding to the offspring will also replace the parent's control parameters.

$$\begin{cases} MS_{hc}(i, d) = c_1 \times X_{i,d} + TF(i) \times (X_{i,d} - X_{j,d}) \\ \quad + (1 - TF(i)) \times (X_{gbest,t} - X_{i,d}) \\ MS_{hc}(j, d) = c_2 \times X_{j,d} + TF(j) \times (X_{j,d} - X_{i,d}) \\ \quad + (1 - TF(j)) \times (X_{gbest,t} - X_{j,d}) \end{cases} \quad (12)$$

Where

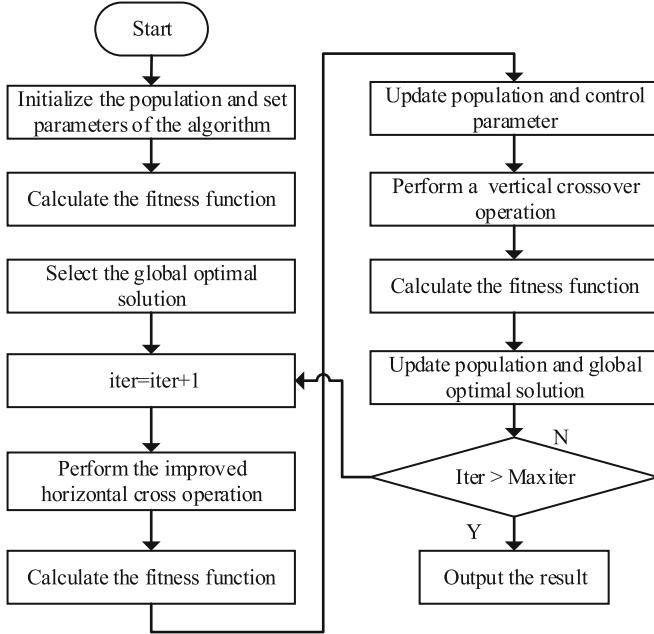
$$TF(i) = F(r_1) + rand \times (F(r_2) - F(r_3)) \quad (13)$$

where,  $c_1$  and  $c_2$  both are random numbers between  $[0, 1]$ , and  $TF$  is the control parameter corresponding to the offspring individual.  $r_1$ ,  $r_2$ ,  $r_3$  are different individual indexes, and are chosen from the cluster  $\{1, 2, \dots, NP\}$  randomly.  $NP$  is the size of population.

The flowchart of the improved adaptive crisscross algorithm is expressed as Fig. 1.

## 4 The Adjustment Mechanism for Infeasible Solutions

The offspring generated by the random optimization algorithm through crossover operations often cannot meet the constraints of equality and inequality. But some



**Fig. 1.** The flow chart of improved adaptive crisscross optimization algorithm

Infeasible solutions may contain significant message to find the optimal solution. In order to make sure the diversity of offspring, a repair technique is put forward, which does not require the selection of penalty factors and any other parameters.

Step 1: Arrange decision variables into matrix form

$$P = \begin{bmatrix} P_1^1 & P_2^1 & \cdots & P_n^1 \\ P_1^2 & P_2^2 & \cdots & P_n^2 \\ \cdots & \cdots & \cdots & \cdots \\ P_1^T & P_2^T & \cdots & P_n^T \end{bmatrix} \quad (14)$$

Where, the matrix  $P$  represents the output of all units in the entire period.  
 Step 2: Initialize the output of each generator.

$$P_{i,t} = P_i^{\min} + (P_i^{\max} - P_i^{\min}) \times r_0 \quad (15)$$

Where  $r_0$  is a random number between [0,1].

Step 3: In order to satisfy the constraints of the limits of ramp rate and active power simultaneously, the upper and lower limits of each generator in the  $t$ -th time period are updated respectively shown as Eq. (16–17).

$$P_{i,t}^{\max} = \begin{cases} P_i^{\max} & \text{if } t == 1 \\ \min [P_i^{\max}, (P_{i,t-1} + UR_i)] & \text{otherwise} \end{cases} \quad (16)$$

$$P_{i,t}^{\min} = \begin{cases} P_i^{\min} & \text{if } t == 1 \\ \max [P_i^{\min}, (P_{i,t-1} - DR_i)] & \text{otherwise} \end{cases} \quad (17)$$

Determine whether the elements in the matrix  $P$  meet the requirements according to the updated  $P_{i,t+1}^{\min}$  and  $P_{i,t+1}^{\max}$ . If it is not satisfied, the boundary absorption is performed to make them equal to the boundary value.

Step 4: Determine the feasibility of the updated candidate solution, as shown in Eq. 18:

$$\left| \delta = \sum_{i=1}^N P_{i,t} - P_{loss,t} - PD_t \right| \leq \varepsilon \quad (18)$$

Where,  $\varepsilon$  is the allowable error of the equality constraint. If the candidate solution meets the constraint conditions, end Step 4; otherwise, enter Step 4.1.

Step 4.1: Set up the number of cycles  $k_{max}$ .

Step 4.2: Randomly select a unit at a certain time and adjust its output as shown in Eq. 19.

$$P(t, Rg(q)) = P(t, Rg(q)) + \delta \quad (19)$$

Where,  $Rg(q)$  is the unit number randomly selected after scrambling all the unit numbers. And judge whether the new solution meets the unit's output constraints.

If the newly generated solution surpasses the lower limit, it is adjusted according to Eq. 20:

$$\begin{aligned} \delta &= P(t, Rg(q)) - P_t^{\min}(Rg(q)) \\ P(t, Rg(q)) &= P_t^{\min}(Rg(q)) \end{aligned} \quad (20)$$

If the newly generated solution surpasses the upper limit, the output is adjusted according to Eq. 21:

$$\begin{aligned} \delta &= P(t, Rg(q)) - P_t^{\max}(Rg(q)) \\ P(t, Rg(q)) &= P_t^{\max}(Rg(q)) \end{aligned} \quad (21)$$

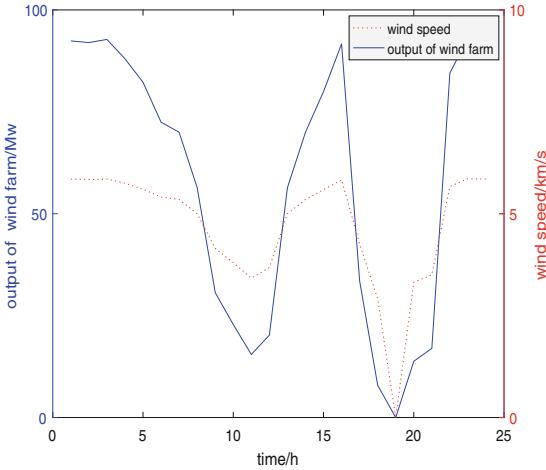
Step 4.3: Recheck whether the solution meets the all constraints. Stop if the solution is feasible, otherwise check if the number of cycles has reached  $k_{max}$ . If the iteration attains  $k_{max}$ , the equation repair process is terminated, otherwise return to step 4.2.

## 5 Simulation Analysis

For the purpose of proving the effectiveness and superiority of the put forward adaptive crisscross optimization algorithm (A-CSO) in the DED problem with consideration of wind farms, a 5-machine test system is used to perform simulation calculations. The load curve and the physical characteristics of each unit are obtained through [24]. The wind farm with a capacity of 100MW is joined to the power system. The wind speed and the corresponding output value in 24 h [28] are shown in Fig. 2. The characteristic parameters of the five units are listed in Table 1. And the key parameters of A-CSO algorithm are set up, shown as:

**Table 1.** System unit characteristic parameters

Unit number	Cost coefficients					The unit output range		Ramp rate limits	
	a	b	c	e	f	$P^{\min}/\text{Mw}$	$P^{\max}/\text{Mw}$	DR/Mw	UR/Mw
1	0.008	2.0	25	100	0.042	10	75	30	30
2	0.003	1.8	60	140	0.040	20	125	30	30
3	0.0012	2.1	100	160	0.038	30	175	40	40
4	0.001	2.0	120	180	0.037	40	250	50	50
5	0.0015	1.8	40	200	0.035	50	300	50	50

**Fig. 2.** Speed of the wind and the output of wind farm

$NP = 100$ , the maximum evolution algebra  $G_{\max}$  set as 8000, the initial control parameter  $F$  is randomly generated in  $[0, 1]$ ; vertical cross competition operator [22]  $P_v$  is set as 0.8.

The scheduling results solved by the put forward A-CSO algorithm are contrasted with three other intelligent algorithms: Crisscross optimization algorithm [22], improved enhanced differential evolution algorithm [24] and improved competitive group optimization algorithm [25]. For the purpose to avoid randomness, all optimization algorithm run 30 times respectively. At the same time, aiming at verifying the effectiveness of the put forward control parameter adaptive method with self-learning ability, the improved algorithm that does not use parameter strategy is applied to solve economic dispatching problem of power systems. All the examples in the experiment consider the valve point effect and the system network loss.

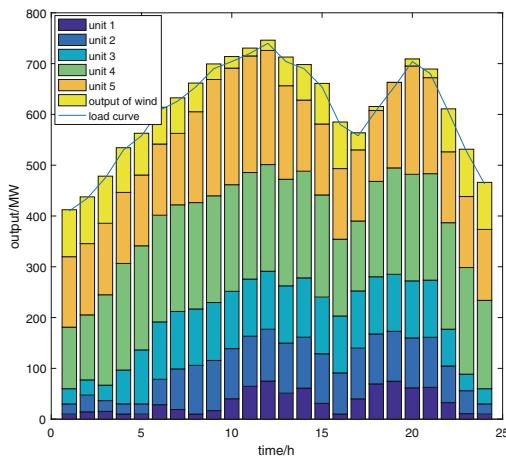
Table 2 shows a group of typical optimal solutions gotten from the A-CSO algorithm. The detailed output of each unit in the scheduling period is shown in Fig. 3. The statistics of the fuel cost gotten from the four algorithms were given in Table 3. The average optimal fitness curve of the four algorithms has been displayed in Fig. 4. Figure 5 shows the optimal fitness curve for dealing with the

**Table 2.** The output of each unit in the dispatch period

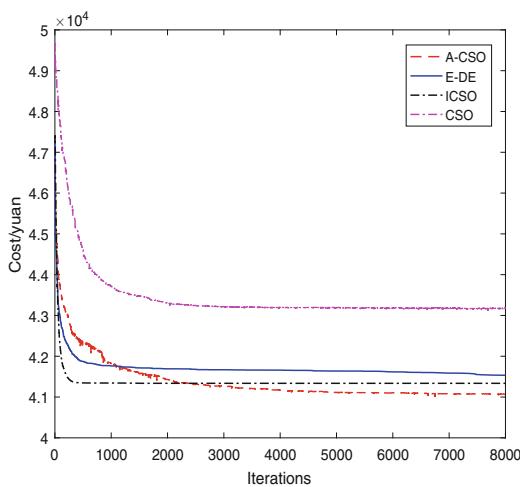
Time/h	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5
1	10.000	20.000	30.000	121.110	138.774
2	14.373	32.969	30.010	127.989	140.317
3	15.274	21.260	30.418	177.989	140.709
4	10.170	20.035	66.555	209.801	139.853
5	10.000	20.000	106.364	204.981	139.307
6	28.488	49.988	113.020	210.025	140.128
7	18.948	79.968	113.097	210.045	140.594
8	10.016	95.953	110.991	209.422	178.951
9	16.861	98.738	114.050	210.133	228.951
10	40.127	98.724	112.803	209.959	229.511
11	64.729	98.544	112.500	209.814	229.519
12	75.000	102.000	114.210	210.027	229.569
13	51.170	98.660	112.672	209.806	184.217
14	61.096	100.277	116.826	209.973	139.996
15	31.097	97.701	111.560	201.110	139.608
16	10.004	81.089	112.073	151.110	139.081
17	40.004	99.816	112.639	137.708	140.048
18	69.427	98.303	112.665	187.708	139.559
19	74.695	98.282	112.191	209.548	168.343
20	61.645	98.231	112.420	209.739	218.302
21	62.602	98.545	112.535	209.675	188.999
22	32.604	71.970	72.545	209.711	139.605
23	10.895	45.146	32.549	209.975	139.766
24	10.000	20.000	30.000	173.882	139.695

**Table 3.** The output of each unit in the dispatch period

Algorithms	Minimum cost/yuan	Maximum cost/yuan	Average cost/yuan
ACSO	41069	41523	41083
E-DE	41154	42277	41827
ICSO	41258	42725	41897
CSO	43521	44623	44123



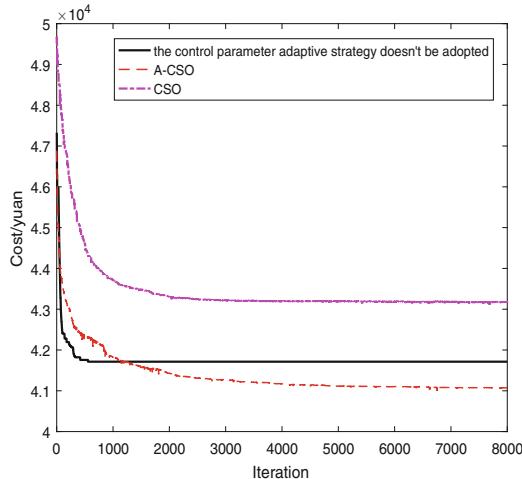
**Fig. 3.** The output of each unit and the load curve



**Fig. 4.** Average optimal fitness evolution curve for different algorithms

**Table 4.** The output of each unit in the dispatch period

algorithms	Minimum cost/yuan	Maximum cost/yuan	Average cost/yuan
ACSO	41069	41523	41083
No adaptive strategy	41786	42011	41846
CSO	43521	44623	44123



**Fig. 5.** Optimal fitness curve with and without control parameter adaptive strategy

DED problem using with and without self-learning control parameter adaptive strategy. Table 4 shows the statistics of the objective function value with and without using control parameter adaptive strategy for solving economic dispatch problem.

As can be seen from Table 3 and Fig. 4, the proposed algorithm has the lowest economic cost, compared with the other three algorithms, indicating that the algorithm has superior optimization performance and good convergence performance. It is an effective way for DED problem. It can be drawn from Fig. 5 and Table 4 that in solving the DED problem, the introduction of global optimal individuals can effectively improve the effectiveness of the algorithm, and the introduction of self-learning control parameter adaptive strategy makes the algorithm have better convergence performance. The lower cost demonstrates the rationality and effectiveness of the proposed strategy.

## 6 Conclusion

In this paper, a DED model with the consideration of wind power, the valve-point effect and system network loss is established. In view of the shortcomings of the crisscross optimization algorithm, the horizontal cross operation of the CSO algorithm is improved, and the self-learning control parameter adaptive strategy is introduced. Aiming at a series of constraints, a heuristic constraint processing method is proposed to maintain the diversity of individuals. From the simulation, it can be drawn that the adaptive crisscross optimization algorithm put forward in this paper has the lowest cost in solving the economic dispatching of power system, and can meet the requirements of economic dispatching. Although E-DE algorithm and ICSO algorithm converged faster in the early stage, they

were surpassed later by the adaptive crisscross optimization algorithm, which indicated that the adaptive crisscross optimization algorithm could effectively shook off the local optimal value, had better global search capability, and could well balance local search and global search. Finally, the simulation results show that the put forward method can obtain better results, has good optimization performance and global convergence, and is an effective strategy for dealing with DED problem.

**Acknowledgment.** The work is supported by Hunan Graduate Research and Innovation Project (CX20190807), National Natural Science Foundation of China (Grant Nos. 61603132, An Improved Competitive Swarm Optimizer for Large Scale Optimization 11 61672226), Hunan Provincial Natural Science Foundation of China (Grant No. 2018JJ2137, 2018JJ3188), Science and Technology Plan of China (2017XK2302), and Doctoral Scientific Research Initiation Funds of Hunan University of Science and Technology (E56126).

## References

1. Wenchuan, M., Jiaju, Q.: Chaotic particle swarm optimization algorithm for economic load dispatch of power system. Proc. CSU-EPSA **9**(2), 114–119 (2007). <https://doi.org/10.1002/jrs.1570>
2. Yinggan, T., Yuhong, C., Leijie, Q.: Application of simplex search method and particle swarm optimization in economic dispatch. Proc. CSU-EPSA **21**(1), 20–26 (2009). [https://doi.org/10.1002/smр.397](https://doi.org/10.1002/smr.397)
3. Qu, B.Y., Zhu, Y.S., Jiao, Y.C., Uw, M.Y., Suganthan, P.N., Liang, J.J.: A survey on multi-objective evolutionary algorithms for the solution of the environmental/economic dispatch Problems. Swarm Evol. Comput. **38**(1), 1–11 (2018). <https://doi.org/10.1016/j.swevo.2017.06.002>
4. Wang Bao, X., Jian, S.Y.: Stochastic dynamic economic dispatch of power systems considering wind power based on versatile probability distribution. Autom. Electric Power Syst. **40**(16), 17–23 (2016). <https://doi.org/10.7500/AEPS20150807004>
5. Yue, W., Yan, Z., Dong, W.: Optimization and scheduling of power system stochastic model predictive control based optimization and scheduling for power system with large scale wind integrated. Control Decis. **34**(08), 1616–1625 (2019)
6. Jabr, R.A., Coonick, A.H., Cory, B.J.: A homogeneous linear programming algorithm for the security constrained economic dispatch problem. IEEE Trans. Power Syst. **15**(3), 930–936 (2000). <https://doi.org/10.1109/59.871715>
7. Travers, D.L., Kaye, R.J.: Dynamic dispatch by constructive dynamic programming. IEEE Trans. Power Syst. **13**(1), 72–78 (1998). <https://doi.org/10.1109/59.651616>
8. Papageorgiou, L.G., Fraga, E.S.: A mixed integer quadratic programming formulation for the economic dispatch of generators with prohibited operating zones. Electr. Power Syst. Res. **77**(10), 1292–1296 (2007). <https://doi.org/10.1016/j.epsr.2006.09.020>
9. Chen, C.L.: Non-convex economic dispatch: a direct search approach. Energy Convers. Manag. **48**(1), 219–225 (2007). <https://doi.org/10.1016/j.enconman.2006.04.010>

10. Hindi, K., Ab, G.M.: Dynamic economic dispatch for large scale power systems: a Lagrangian relaxation approach. *Int. J. Electr. Power Energy Syst.* **13**, 51–56 (1991)
11. Lin, W., Cheng, F., Tsay, M.: Nonconvex economic dispatch by integrated artificial intelligence. *IEEE Trans. Power Syst.* **16**(2), 307–311 (2001). <https://doi.org/10.1109/59.918303>
12. Zuo Lixia, Y., Yuan, S.H.: Research on dynamic economic emission dispatch model of power system. *J. East China Jiaotong Univ.* **161**(03), 138–146 (2018)
13. Chiang, C.L.: Improved genetic algorithm for power economic dispatch of units with valve-point effects and multiple fuels. *IEEE Trans. Power Syst.* **20**(4), 1690–1699 (2005). <https://doi.org/10.1109/tpwrs.2005.857924>
14. Niu, Q., Zhang, H., Li, K., et al.: An efficient harmony search with new pitch adjustment for dynamic economic dispatch. *Energy* **65**, 25–43 (2014). <https://doi.org/10.1016/j.energy.2013.10.085>
15. Basu, M.: Artificial bee colony optimization for multi-area economic dispatch. *Int. J. Electr. Power Energy Syst.* **49**(49), 181–187 (2013). <https://doi.org/10.1016/j.ijepes.2013.01.004>
16. Sun, Q., Yang, L., Zhang, H.: Smart energy - applications and prospects of artificial intelligence technology in power system. *Control Decis.* **5**, 938–949 (2018)
17. Zhang, J., Long, J., Yue, C.: An modified particle swarm optimizer for short-term hydrothermal scheduling with cascaded reservoirs. *Control Decis.* **26**(3), 407–412 (2011)
18. Liang, H., Liu, Y., Shen, Y., et al.: A hybrid bat algorithm for economic dispatch with random wind power. *IEEE Trans. Power Syst.*, 1 (2018). <https://doi.org/10.1109/TPWRS.2018.2812711>
19. Elattar, E.E.: A hybrid genetic algorithm and bacterial foraging approach for dynamic economic dispatch problem. *Int. J. Electr. Power Energy Syst.* **69**, 18–26 (2015). <https://doi.org/10.1016/j.ijepes.2014.12.091>
20. Menglin, Z., Xiaomeng, A., Jinyu, W.: Economic dispatch for power system integrated with wind power using Stackelberg game. *Control Theory Appl.* **35**(05), 80–88 (2018). <https://doi.org/10.7641/CTA.2017.70676>
21. Lianghong, W., Yaonan, W., Xiaofang, Y.: Fast self-adaptive differential evolution algorithm for power economic load dispatch. *Control Decis.* **28**(4), 557–562 (2013)
22. Meng, A., Hu, H., Hao, Y., et al.: Crisscross optimization algorithm for large-scale dynamic economic dispatch problem with valve-point effects. *Energy* **93**, 2175–2190 (2015). <https://doi.org/10.1016/j.energy.2015.10.112>
23. Xie, L., Ilic, M.D.: Model predictive economic/environmental dispatch of power systems with intermittent resources. In: Power and Energy Society General Meeting (PES 2009), pp. 1–6. IEEE (2009). <https://doi.org/10.1109/PES.2009.5275940>
24. Zaman, M.F., Elsayed, S.M., Ray, T.: Evolutionary algorithms for dynamic economic dispatch problems. *IEEE Trans. Power Syst.* **31**(2), 1486–1495 (2016). <https://doi.org/10.1109/TPWRS.2015.2428714>
25. Yanyan, G., Guojian, X.: Large scale power system economic dispatch based on an improved competitive swarm optimizer. *Power Syst. Protect. Control* **45**(15), 97–103 (2017)
26. Jie, C., Shen, Y., Lu, X.: An intelligent multi-objective optimized method for wind power prediction intervals. *Power Syst. Technol.* **40**(8), 2281–2283 (2016)
27. Shuo, Z., Gengyin, L., Ming, Z.: Reliability modeling of large-scale wind farms. *Power Syst. Technol.* **33**(13), 37–41 (2009)
28. Na, Z.: Economic comparison for different generation schedulings with large scale wind power. In: Power System. North China Electric Power University (2012)



# Consensus for Heterogeneous Networked Systems Based on Second-Order Neighbors' Information

Lei Wang, Huanyu Zhao<sup>(✉)</sup>, Dongsheng Du, and Hongbiao Zhou

Faculty of Automation, Huaiyin Institute of Technology, Huai'an 223003, China  
hyzhao@163.com

**Abstract.** In this paper, we focus on the consensus for a class of heterogeneous networked systems by applying the second-order neighbors information. According to the communication with their neighbors, two kinds of linear consensus control algorithms are put forward for agents. Through using graph theory and matrix theory, it provides sufficient conditions for the consensus of heterogeneous networked systems with undirected connected topology.

**Keywords:** Networked system · Heterogeneous system · Consensus

## 1 Introduction

In recent years, as the basic and critical issue in cooperation control of systems, information consensus attracts considerable attention from artificial intelligence, automatic control and other fields. The key issue of consensus is that try to construct appropriate control algorithms, then agent systems can reach an agreement by communicating with their neighbors with limited information. By employing graph theory, matrix theory, Lyapunov theorem and others, researchers studied the problem that first-order, second-order and higher-order homogeneous multi-agent systems how to get the consensus [1–9]. However, in practice, dynamic systems differ from each other, so most of them are heterogeneous systems. Zheng et al. [10] studied the condition of heterogeneous multi-agent systems to reach an agreement under linear consensus algorithm and saturated consensus algorithm when the velocity information of second-order agents was not measurable. Sun et al. [11] studied the expression of the final convergence value of the consensus protocol for heterogeneous multi-agent systems, then gave the convergence interval. Literature [12] proposed a composite controller, which designed a disturbance observer with limited time and used the technology based on the power integrator to guarantee the leader-following rendezvous in finite time. Through the action of the leader, literature [13] made the multi-agent system flock based on the robust control of high gain feedback.

In most of the above studies, first-order neighbors' information was considered. However, with the increasing complexity of the networked systems, the

information exchange mode of first-order neighbor protocol can't meet our needs any longer. Compared with first-order neighbor protocol, except the information of agent neighbors, second-order neighbor protocol still uses the information of second-order neighbors, which can increase efficiency, save cost and optimize performance. At the same time, the communication topology of the system has not changed. Pan et al. [14] and Wang et al. [15] concluded that the agent convergence rate under the second-order protocol is faster in comparison with the convergence rate under the general protocol. Xia et al. [16] adopted an event-based control method to study the problem of using second-order neighbors information to speed up the multi-agent systems to converge under fixed and switched topology network respectively.

This study investigates the consensus problem for networked systems composed of first-order and second-order agents. According to the information of the agent itself and its neighbors, two types of linear consensus control algorithms are put forward. Through employing graph theory and matrix theory, the convergence values of agents are given when the time approaches infinity, and the sufficient conditions for heterogeneous networked systems with fixed topology to reach consensus are obtained.

## 2 Preliminaries and Problem Statement

### 2.1 Preliminaries

In this paper, let  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$  be an undirected graph, where  $\mathcal{V} = \{v_1, \dots, v_n\}$ ,  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$  and  $A = [a_{ij}]_{n \times n}$  denote node set, edge set and weighted adjacency matrix with non-negative elements  $a_{ij}$ , respectively. If  $(v_i, v_j) \in \mathcal{E}$ , node  $v_j$  can receive information from node  $v_i$ . If  $(v_i, v_j) \in \mathcal{E}$ ,  $a_{ij} > 0$ , otherwise  $a_{ij} = 0$ . Suppose that  $a_{ii} = 0$  for  $i = 1, 2, \dots, n$ .  $a_{ij} = a_{ji}$  if the graph is undirected.  $N^i = \{v_j \in \mathcal{V} : (v_i, v_j) \in \mathcal{E}\}$  denotes the set of neighbors of node  $v_i$ . The Laplacian matrix  $L = [l_{ij}]_{n \times n}$  of graph  $\mathcal{G}$  is defined as

$$l_{ij} = \begin{cases} \sum_{j=1, j \neq i}^n a_{ij}, & i = j \\ -a_{ij}, & i \neq j \end{cases}, \quad i, j = 1, \dots, n.$$

### 2.2 Problem Statement

In this study, the heterogeneous networked system contains  $n$  agents, the number of agents with second-order dynamics is  $m$  ( $m < n$ ) and the rest are agents with first-order dynamics. These agents can get information from second-order neighbors.

The dynamics of the system are described by

$$\begin{cases} \dot{x}_i(t) = v_i(t) \\ \dot{v}_i(t) = u_i(t) \end{cases}, \quad i = 1, \dots, m$$

$$\dot{x}_i(t) = u_i(t), \quad i = m + 1, \dots, n \tag{1}$$

where  $x_i \in R$ ,  $v_i \in R$ , and  $u_i \in R$  respectively, denote position state, velocity state and control input of agent  $i$ .

If agents communicate with each other based on the information from first-order neighbors, the adjacency matrix  $A$  is written as

$$A = \begin{bmatrix} A_s & A_{sf} \\ A_f & A_{fs} \end{bmatrix}$$

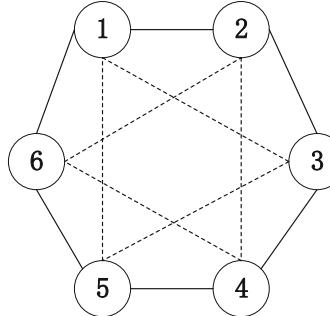
where  $A_s \in R^{m \times m}$  and  $A_f \in R^{(n-m) \times (n-m)}$ .

Then, the Laplacian matrix  $L$  can be expressed as

$$L = \begin{bmatrix} L_s + D_{sf} & -A_{sf} \\ -A_{fs} & L_f + D_{fs} \end{bmatrix}$$

where  $L_f$  and  $L_s$  represent the Laplacian matrix of the first-order agents and the second-order agents, respectively, with  $D_{sf} = \text{diag}(\sum_{j \in N^{sf}} a_{1j}, \dots, \sum_{j \in N^{sf}} a_{mj})$  and  $D_{fs} = \text{diag}(\sum_{j \in N^{fs}} a_{m+1j}, \dots, \sum_{j \in N^{fs}} a_{nj})$ .

This paper investigates the networked systems based on second-order neighbors' information. That is to say, for agent  $i$ , the information from first-order neighbors and second-order neighbors are considered at the same time, which corresponds to add virtual second-order paths to the original graph. Define  $\hat{\mathcal{G}} = (\mathcal{V}, \hat{\mathcal{E}}, \hat{A})$  as second-order neighbor communication topology, the set of second-order neighbors is denoted by  $N^2$ , as shown in Fig. 1, original graph is composed of nodes  $\{v_1, v_2, \dots, v_6\}$  with the edges and the second-order communication topology is composed of the same nodes with dotted lines.



**Fig. 1.** Communication graph with second-order neighbors.

Therefore, the adjacency matrix  $\tilde{A} = A + \hat{A}$  in this paper can be described as

$$\tilde{A} = \begin{bmatrix} A_s + \hat{A}_s & A_{sf} + \hat{A}_{sf} \\ A_f + \hat{A}_f & A_{fs} + \hat{A}_{fs} \end{bmatrix} = \begin{bmatrix} \tilde{A}_s & \tilde{A}_{sf} \\ \tilde{A}_f & \tilde{A}_{fs} \end{bmatrix}$$

The Laplacian matrix  $\tilde{L} = L + \hat{L}$  is written as

$$\tilde{L} = \begin{bmatrix} L_s + D_{sf} + \hat{L}_s + \hat{D}_{sf} & -A_{sf} - \hat{A}_{sf} \\ -A_{fs} - \hat{A}_{fs} & L_f + D_{fs} + \hat{L}_f + \hat{D}_{fs} \end{bmatrix} = \begin{bmatrix} \tilde{L}_s + \tilde{D}_{sf} & -\tilde{A}_{sf} \\ -\tilde{A}_{fs} & \tilde{L}_f + \tilde{D}_{fs} \end{bmatrix}$$

Let  $x_s = [x_1, x_2, \dots, x_m]^T$ ,  $v_s = [v_1, v_2, \dots, v_m]^T$ ,  $x_f = [x_{m+1}, x_{m+2}, \dots, x_n]^T$  and  $y = [x_s^T, v_s^T, x_f^T]^T$ . The initial condition of the system can be expressed as  $y(0) = [x_s^T(0), v_s^T(0), x_f^T(0)]^T$ .

**Definition 1.** For any initial conditions  $x_i(0)$  ( $i = 1, \dots, n$ ) and  $v_i(0)$  ( $i = 1, \dots, m$ ), the heterogeneous networked system (1) is said to reach consensus asymptotically if we have

$$\begin{aligned} \lim_{t \rightarrow \infty} \|x_i(t) - x_j(t)\| &= 0, (i, j = 1, 2, \dots, n), \\ \lim_{t \rightarrow \infty} \|v_i(t) - v_j(t)\| &= 0, (i, j = 1, 2, \dots, m). \end{aligned}$$

### 3 Main Results

#### 3.1 Consensus Analysis for Case I

In order to solve the consensus problem of system (1), we consider the information from relative position and the absolute velocity, the following consensus control algorithm is designed for each agent:

$$u_i(t) = \begin{cases} \sum_{j \in N} a_{ij} \beta_1 (x_j - x_i) + \sum_{k \in N^2} a_{ik} \beta_1 (x_k - x_i) - \beta_2 v_i & i = 1, 2, \dots, m \\ \sum_{j \in N} a_{ij} \beta_3 (x_j - x_i) + \sum_{k \in N^2} a_{ik} \beta_3 (x_k - x_i) & i = m + 1, \dots, n \end{cases} \quad (2)$$

where  $\beta_1 > 0$ ,  $\beta_2 > 0$  and  $\beta_3 > 0$  are parameters to be determined later. Using (2), the dynamic of the heterogeneous system (1) can be written in a matrix form as

$$\dot{y} = \Omega y$$

$$\text{where } \Omega = \begin{bmatrix} 0 & I_m & 0 \\ -\beta_1 \bar{L}_s & -\beta_2 I_m & \beta_1 \tilde{A}_{sf} \\ \beta_3 \tilde{A}_{fs} & 0 & -\beta_3 \bar{L}_f \end{bmatrix} \text{ with } \bar{L}_s = \tilde{L}_s + \tilde{D}_{sf} \text{ and } \bar{L}_f = \tilde{L}_f + \tilde{D}_{fs}.$$

**Lemma 1** [17]. Suppose that  $L_n$  is the Laplacian matrix associated with graph  $\mathcal{G}_n$ . Then, the following three conditions are equivalent:

1.  $L_n$  has only one zero eigenvalue with eigenvector  $1_n$ , and all non-zero eigenvalues have positive real parts;
2.  $\text{rank}(L_n) = n - 1$ ;
3.  $\mathcal{G}_n$  has a spanning tree.

**Lemma 2.** The matrix  $\Omega$  has only one zero eigenvalue, and all non-zero eigenvalues have negative real parts, if  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$  satisfy the following conditions:

- a)  $0 < \beta_1 < \frac{\beta_2 - 1}{\max d_i}$ ,  $i = 1, 2, \dots, m$ ;
- b)  $\beta_2 > 1$ ;
- c)  $\beta_3 > 0$ .

*Proof.* The nonsingular transformation of matrix  $\Omega$  is as follows:

$$\bar{\Omega} = M\Omega M^{-1} = \begin{bmatrix} -I_m & I_m & 0 \\ -\beta_1 \bar{L}_s + (\beta_2 - 1) I_m & I_m (1 - \beta_2) & \beta_1 \tilde{A}_{sf} \\ \beta_3 \tilde{A}_{fs} & 0 & -\beta_3 \bar{L}_f \end{bmatrix}$$

where  $M = \begin{bmatrix} I_m & 0 & 0 \\ I_m & I_m & 0 \\ 0 & 0 & I_{n-m} \end{bmatrix}$  with  $I_m$  and  $I_{n-m}$  are the unit matrices.

Due to the transformation of matrix  $\Omega$  is nonsingular, the eigenvalue properties of  $\Omega$  and  $\bar{\Omega}$  are the same. When all the conditions in Lemma 2 are satisfied, the matrix  $-\bar{\Omega}$  denotes the Laplacian matrix of  $(n + m)$  agents. Then, it can be inferred from Lemma 1 that the matrix  $\bar{\Omega}$  only contains the zero eigenvalue and all non-zero eigenvalue with negative real part.

Next, several elementary rank transformations will be employed on  $\bar{\Omega}$ . Exchange the position of the second column and the first column. Then the first column is added to the second column after the exchange. Then the third line is divided by  $-\beta_3$ , the first line is multiplied by  $(\beta_2 - 1)$  and added to the second line. Then the second line is divided by  $-\beta_1$ . We obtain that

$$\bar{\Omega} \rightarrow \begin{bmatrix} I_m & 0 & 0 \\ 0 & \bar{L}_s & -\tilde{A}_{sf} \\ 0 & -\tilde{A}_{fs} & \bar{L}_f \end{bmatrix} \rightarrow \begin{bmatrix} I_m & 0 \\ 0 & \tilde{L} \end{bmatrix}$$

The communication topology graph is undirected connected. Hence, we know that  $\text{rank}(\tilde{L}) = n - 1$  from Lemma 1. Because the elementary transformation does not change the rank of the matrix,  $\text{rank}(\bar{\Omega}) = m + \text{rank}(\tilde{L}) = m + n - 1$ . We can draw a conclusion that the matrix  $\bar{\Omega}$  has only one zero eigenvalue.

Through above analysis, there is only one zero eigenvalue in matrix  $\Omega$ , and all non-zero eigenvalues have negative real parts. Therefore, the proof is completed.

**Theorem 1.** Under the condition that parameters  $\beta_1, \beta_2, \beta_3$  satisfy the conditions in Lemma 2. The system (1) with algorithm (2) is said to achieve consensus asymptotically if the communication graph contains a spanning tree.

*Proof.* From Lemma 2, there is only one zero eigenvalue in matrix  $\Omega$ , and all non-zero eigenvalues have negative real parts. Jordan canonical form of matrix  $\Omega$  can be expressed as

$$\Omega = PJP^{-1} = [w_1, w_2, \dots, w_{n+m}] \begin{bmatrix} 0 & 0_{1 \times (n+m-1)} \\ 0_{(n+m-1) \times 1} & J' \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_{n+m}^T \end{bmatrix},$$

where  $w_j \in R^{n+m}$ ,  $j = 1, 2, \dots, n+m$  is the right eigenvector and the generalized right eigenvector of matrix  $\Omega$ ;  $v_j \in R^{n+m}$ ,  $j = 1, 2, \dots, n+m$  is the left eigenvector and the generalized left eigenvector of matrix  $\Omega$ ;  $J'$  is the Jordan block corresponding to the non-zero eigenvalue of the matrix  $\Omega$ .

Supposing  $w_1 = [1_m^T, 0_m^T, 1_{n-m}^T]^T$ , we have  $\Omega w_1 = 0$ .  $w_1$  is the eigenvector of matrix  $\Omega$  corresponding to the zero eigenvalue. It's easy to figure out that the left eigenvector  $v_1$  corresponding to  $w_1$  is  $v_1^T = (\alpha\beta_2\beta_3 1_m^T, \alpha\beta_3 1_m^T, \alpha\beta_1 1_{n-m}^T)$ .  $w_1$  is the column vector of invertible matrix  $P$ ,  $v_1^T$  is the row vector of invertible matrix  $P^{-1}$ .  $PP^{-1} = I$ , so  $w_1 v_1^T = 1$ . That is  $\alpha\beta_2\beta_3 m + \alpha\beta_1(n-m) = 1$ , so

$$\alpha = \frac{1}{m(\beta_2\beta_3 - \beta_1) + n\beta_1}.$$

Due to  $e^{\Omega t} = Pe^{Jt}P^{-1} = P \begin{bmatrix} 1 & 0 \\ 0 & e^{J't} \end{bmatrix} P^{-1}$ , we obtain that

$$\lim_{t \rightarrow \infty} e^{\Omega t} = [w_1 v_1^T] = \begin{bmatrix} \alpha\beta_2\beta_3 1_m 1_m^T & \alpha\beta_3 1_m 1_m^T & \alpha\beta_1 1_m 1_{n-m}^T \\ 0_{m \times m} & 0_{m \times m} & 0_{m \times (n-m)} \\ \alpha\beta_2\beta_3 1_{n-m} 1_m^T & \alpha\beta_3 1_{n-m} 1_m^T & \alpha\beta_1 1_{n-m} 1_{n-m}^T \end{bmatrix}.$$

Therefore,

$$\begin{aligned} \lim_{t \rightarrow \infty} \begin{bmatrix} x_s \\ v_s \\ x_f \end{bmatrix} &= \lim_{t \rightarrow \infty} e^{\Omega t} \begin{bmatrix} x_s(0) \\ v_s(0) \\ x_f(0) \end{bmatrix} \\ &= \begin{bmatrix} \alpha\beta_2\beta_3 1_m 1_m^T & \alpha\beta_3 1_m 1_m^T & \alpha\beta_1 1_m 1_{n-m}^T \\ 0_{m \times m} & 0_{m \times m} & 0_{m \times (n-m)} \\ \alpha\beta_2\beta_3 1_{n-m} 1_m^T & \alpha\beta_3 1_{n-m} 1_m^T & \alpha\beta_1 1_{n-m} 1_{n-m}^T \end{bmatrix} \begin{bmatrix} x_s(0) \\ v_s(0) \\ x_f(0) \end{bmatrix} \end{aligned}$$

Because  $\alpha$  is a constant,  $x_s$ ,  $v_s$  and  $x_f$  converge to a fixed value when  $t \rightarrow \infty$ . Hence,  $\lim_{t \rightarrow \infty} \|x_i(t) - x_j(t)\| = 0$  ( $i, j = 1, 2, \dots, n$ ),  $\lim_{t \rightarrow \infty} \|v_i(t) - v_j(t)\| = 0$  ( $i, j = 1, 2, \dots, m$ ). Therefore, the consensus of the heterogeneous system (1) is achieved with algorithm (2). This completes the proof.

### 3.2 Consensus Analysis for Case II

In this section, the relative position and velocity information of networked systems (1) are considered. The consensus control algorithm is devised as follows:

$$u_i(t) = \begin{cases} \sum_{j \in N} a_{ij} [\gamma_1(x_j - x_i) + \gamma_2(v_j - v_i)] \\ + \sum_{k \in N^2} a_{ik} [\gamma_1(x_k - x_i) + \gamma_2(v_k - v_i)] & i = 1, 2, \dots, m \\ \sum_{j \in N} a_{ij} \gamma_3 (x_j - x_i) \\ + \sum_{k \in N^2} a_{ik} \gamma_3 (x_k - x_i) & i = m + 1, \dots, n \end{cases} \quad (3)$$

where  $\gamma_1 > 0$ ,  $\gamma_2 > 0$  and  $\gamma_3 > 0$  are parameters to be determined later.

With the control algorithm described by (3), the heterogeneous system (1) is rewritten in a matrix form as

$$\dot{y} = \Omega' y$$

where  $\Omega' = \begin{bmatrix} 0 & I_m & 0 \\ -\gamma_1 \bar{L}_s & -\gamma_2 \bar{L}_s & \gamma_1 \tilde{A}_{sf} \\ \gamma_3 \tilde{A}_{fs} & 0 & -\gamma_3 \bar{L}_f \end{bmatrix}$  with  $\bar{L}_s = \tilde{L}_s + \tilde{D}_{sf}$  and  $\bar{L}_f = \tilde{L}_f + \tilde{D}_{fs}$ .

**Lemma 3.** *The matrix  $\Omega'$  has only one zero eigenvalue, and all non-zero eigenvalues have negative real parts, if  $\gamma_1, \gamma_2, \gamma_3$  satisfy the following conditions:*

- a)  $0 < \gamma_1 < \gamma_2 - \frac{1}{\max d_i}, \quad i = 1, 2, \dots, m;$
- b)  $\gamma_2 > \frac{1}{\max d_i};$
- c)  $\gamma_3 > 0.$

*Proof.* By using nonsingular transformation on matrix  $\Omega'$ , we obtain that

$$\bar{\Omega}' = M \Omega' M^{-1} = \begin{bmatrix} -I_m & I_m & 0 \\ (\gamma_2 - \gamma_1) \bar{L}_s - I_m & I_m - \gamma_2 \bar{L}_s & \gamma_1 \tilde{A}_{sf} \\ \gamma_3 \tilde{A}_{fs} & 0 & -\gamma_3 \bar{L}_f \end{bmatrix}$$

where  $M = \begin{bmatrix} I_m & 0 & 0 \\ I_m & I_m & 0 \\ 0 & 0 & I_{n-m} \end{bmatrix}$  with  $I_m$  and  $I_{n-m}$  are the unit matrices.

Because the transformation of matrix  $\Omega'$  is nonsingular, the eigenvalue properties of  $\Omega'$  and  $\bar{\Omega}'$  are the same. When all conditions in Lemma 3 are satisfied, the matrix  $-\bar{\Omega}'$  represents the Laplacian matrix corresponding to the communication topology of  $(n+m)$  agents. According to Lemma 1, the matrix  $\bar{\Omega}'$  simply contains the zero eigenvalue and the non-zero eigenvalue with negative real part.

Some elementary rank transformations are taken on  $\bar{\Omega}'$  as follows. Swap the second column with the first column, Then, add the first column to the second column after the exchange. Next, the third line is divided by  $-\gamma_3$ , the first line is multiplied by  $-\gamma_2 \bar{L}_S$  and added to the second line. Lastly, the second line is divided by  $-\gamma_1$ . We obtain that

$$\bar{\Omega}' \rightarrow \begin{bmatrix} I_m & 0 & 0 \\ 0 & \bar{L}_s & -\tilde{A}_{sf} \\ 0 & -\tilde{A}_{fs} & \bar{L}_f \end{bmatrix} \rightarrow \begin{bmatrix} I_m & 0 \\ 0 & \bar{L} \end{bmatrix}$$

Similar to the proof of Lemma 2, it can be proved that  $\text{rank}(\bar{\Omega}') = m + \text{rank}(\bar{L}) = m + n - 1$ . Therefore, the matrix  $\bar{\Omega}'$  has only one zero eigenvalue.

Through above analysis, there is only one zero eigenvalue in matrix  $\bar{\Omega}'$ , and all non-zero eigenvalues have negative real parts. This completes the proof.

**Theorem 2.** *Under the condition that parameters  $\gamma_1, \gamma_2, \gamma_3$  satisfy the conditions in Lemma 3. Supposing the communication graph contains a spanning tree, the system (1) with algorithm (3) is said to achieve consensus asymptotically.*

*Proof.* There is only one zero eigenvalue in matrix  $\bar{\Omega}'$ , and all non-zero eigenvalues have negative real parts from Lemma 3. Jordan canonical form of matrix  $\bar{\Omega}'$  can be written as

$$\Omega' = PHP^{-1} = [p_1, p_2, \dots, p_{n+m}] \begin{bmatrix} 0 & 0_{1 \times (n+m-1)} \\ 0_{(n+m-1) \times 1} & H' \end{bmatrix} \begin{bmatrix} q_1^T \\ \vdots \\ q_{n+m}^T \end{bmatrix},$$

where  $p_j \in R^{n+m}$ ,  $j = 1, 2, \dots, n+m$  is the right eigenvector and the generalized right eigenvector of matrix  $\Omega'$ ;  $q_j \in R^{n+m}$ ,  $j = 1, 2, \dots, n+m$  is the left eigenvector and the generalized left eigenvector of matrix  $\Omega'$ ;  $J'$  is the Jordan block corresponding to the non-zero eigenvalue of the matrix  $\Omega'$ .

Letting  $p_1 = [1_m^T, 0_m^T, 1_{n-m}^T]^T$ , we obtain  $\Omega' p_1 = 0$ . Zero is the eigenvalue of matrix  $\Omega'$  corresponding to the  $p_1$  eigenvector. It's easy to find out that the left eigenvector  $q_1$  corresponding to  $p_1$  is  $q_1^T = (\alpha\gamma_2\gamma_3 1_m^T, \alpha\gamma_3 1_m^T, \alpha\gamma_1 1_{n-m}^T)$ .  $p_1$  is the column vector of invertible matrix  $P$ ,  $q_1^T$  is the row vector of invertible matrix  $P^{-1}$ .  $PP^{-1} = I$ , so  $p_1 q_1^T = 1$ . That is  $\alpha\gamma_2\gamma_3 m + \alpha\gamma_1(n - m) = 1$ , so  $\alpha = \frac{1}{m(\gamma_2\gamma_3 - \gamma_1) + n\gamma_1}$ .

Because  $e^{\Omega' t} = Pe^{Ht}P^{-1} = P \begin{bmatrix} 1 & 0 \\ 0 & e^{H't} \end{bmatrix} P^{-1}$ , we obtain that

$$\lim_{t \rightarrow \infty} e^{\Omega' t} = [p_1 q_1^T] = \begin{bmatrix} \alpha\gamma_2\gamma_3 1_m 1_m^T & \alpha\gamma_3 1_m 1_m^T & \alpha\gamma_1 1_m 1_{n-m}^T \\ 0_{m \times m} & 0_{m \times m} & 0_{m \times (n-m)} \\ \alpha\gamma_2\gamma_3 1_{n-m} 1_m^T & \alpha\gamma_3 1_{n-m} 1_m^T & \alpha\gamma_1 1_{n-m} 1_{n-m}^T \end{bmatrix}.$$

Therefore,

$$\begin{aligned} \lim_{t \rightarrow \infty} \begin{bmatrix} x_s \\ v_s \\ x_f \end{bmatrix} &= \lim_{t \rightarrow \infty} e^{\Omega' t} \begin{bmatrix} x_s(0) \\ v_s(0) \\ x_f(0) \end{bmatrix} \\ &= \begin{bmatrix} \alpha\gamma_2\gamma_3 1_m 1_m^T & \alpha\gamma_3 1_m 1_m^T & \alpha\gamma_1 1_m 1_{n-m}^T \\ 0_{m \times m} & 0_{m \times m} & 0_{m \times (n-m)} \\ \alpha\gamma_2\gamma_3 1_{n-m} 1_m^T & \alpha\gamma_3 1_{n-m} 1_m^T & \alpha\gamma_1 1_{n-m} 1_{n-m}^T \end{bmatrix} \begin{bmatrix} x_s(0) \\ v_s(0) \\ x_f(0) \end{bmatrix}, \end{aligned}$$

Due to  $\alpha$  is a constant,  $x_s$ ,  $v_s$  and  $x_f$  converge to a fixed value when  $t \rightarrow \infty$ . Hence,  $\lim_{t \rightarrow \infty} \|x_i(t) - x_j(t)\| = 0$  ( $i, j = 1, 2, \dots, n$ ),  $\lim_{t \rightarrow \infty} \|v_i(t) - v_j(t)\| = 0$  ( $i, j = 1, 2, \dots, n$ ). Thus, under algorithm (3), the heterogeneous system (1) achieve consensus. This completes the proof.

## 4 Conclusion

This paper has studied that the heterogeneous networked systems how to reach consensus by applying the second-order neighbors' information. Two types of linear consensus control algorithms have been proposed based on the information

of agent itself and its neighbors and independent of any assumptions. Through using the graph theory and the matrix theory, it has provided sufficient conditions for heterogeneous multi-agent systems with the fixed topology to achieve consensus. This paper does not consider the communication delay. Future work will concentrate on this part.

**Funding.** The work was supported by “333” Project in Jiangsu Province (Grant No. BRA2019285).

## References

1. Gao, Y., Liu, B., Zuo, M.: Consensus of first-order multi-agent systems. *Math. Pract. Theory* **43**(10), 106–110 (2013). (In Chinese)
2. Wang, N., Wang, J., Cao, Z.: Sampled-data consensus of first-order multi-agent systems with weighted average prediction. *J. Air Force Eng. Univ. (Nat. Sci. Edn.)* **18**(1), 105–110 (2017). (In Chinese)
3. Oyedele, M.O., Mahmoud, M.S.: Couple-group consensus conditions for general first-order multiagent systems with communication delays. *Syst. Control Lett.* **117**, 37–44 (2018)
4. Huang, H., Huang, T., Wu, S.: Leader-following consensus of second-order multi-agent systems via event-triggered control. *Control Decis.* **31**(5), 835–841 (2016). (In Chinese)
5. Guo, W.: Leader-following consensus of the second-order multi-agent systems under directed topology. *ISA Trans.* **65**, 116–124 (2016)
6. Yang, T., Meng, Z., Dimarogonas, D.V., et al.: Periodic behaviors for discrete-time second-order multiagent systems with input saturation constraints. *IEEE Trans. Circ. Syst. II Express Briefs* **63**(7), 663–667 (2017)
7. Ge, Y., Chen, Y., Zhang, Y., et al.: State consensus analysis and design for high-order discrete-time linear multiagent systems. *Math. Prob. Eng.* **2013**(7), 1–13 (2013)
8. Wang, J., Chen, K., Zhang, Y.: Consensus of high-order nonlinear multiagent systems with constrained switching topologies. *Complexity* **2017**(11), 1–11 (2017)
9. Ma, Q., Lu, J., Xu, H.: Consensus for nonlinear multi-agent systems with sampled data. *Trans. Inst. Measur. Control* **36**(5), 618–626 (2014)
10. Zheng, Y., Wang, L.: Consensus of heterogeneous multi-agent systems without velocity measurements. *Int. J. Control* **85**(7), 906–914 (2012)
11. Sun, Y., Zhang, G., Zhang, S., et al.: Convergence analysis for consensus protocol of heterogeneous multi-agent systems. *Control Theory Appl.* **31**(11), 1524–1529 (2014)
12. Li, P., Xu, S., Chu, Y., et al.: Finite-time leader-following rendezvous for Euler–Lagrange multi-agent systems with an uncertain leader. *Trans. Inst. Measur. Control* **40**(6), 1766–1775 (2018)
13. Qing, Z., Jie, W., Zhengquan, Y., et al.: High gain feedback robust control for flocking of multi-agents system. *Trans. Inst. Measur. Control* **41**(13), 3769–3776 (2019)
14. Pan, H., Nian, X., Guo, L.: Second-order consensus in multi-agent systems based on second-order neighbors' information. *Int. J. Syst. Sci.* **45**(5), 902–914 (2014)

15. Wang, K., Ji, Z., Chao, Y.: A control strategy for maitaining controllability and observability of a multi-agent system with the second-order neighborhood protocol. *CAAI Trans. Intell. Syst.* **12**(2), 213–220 (2017). (In Chinese)
16. Xia, Q., Liu, K., Ji, Z.: Event-triggered consensus of multi-agent systems based on second-order neighbors. *CAAI Trans. Intell. Syst.* **12**(6), 833–840 (2017). (In Chinese)
17. Ren, W.: Distributed consensus in multivehicle cooperative control: Theory and applications research interests control systems & robotics. *Commun. Control Eng.* **27**(2), 71–82 (2008)



# Sliding Mode Control for Neutral-Type Systems with Stochastic Noises and Time-Delay

Qiaoyu Chen<sup>1</sup>, Wuneng Zhou<sup>1</sup>, and Dongbing Tong<sup>2(✉)</sup>

<sup>1</sup> College of Information Sciences and Technology, Donghua University,  
Shanghai 200051, China

goodluckqiaoyu@126.com, wnzhou@dhu.edu.cn

<sup>2</sup> College of Electronic and Electrical Engineering,  
Shanghai University of Engineering Science, Shanghai 201620, China  
tongdongbing@163.com

**Abstract.** This paper discusses the exponentially stable problem of stochastic neutral-type systems with Lévy noises by sliding mode control (SMC). According to the SMC method, criteria of the exponentially stable for stochastic neutral-type systems is obtained. Moreover, the update law of the control gain is provided for neutral-type systems. Finally, a simulation example is offered to show the advantages of the theoretical results.

**Keywords:** Sliding mode control · Neutral-type systems · Stochastic noises

## 1 Introduction

In the actual system, the current state of the system is often affected by the past state. In the process of network transmission, the current state and the past state may suffer from communication delay. These effects should be taken into account when mathematical models are established. And the neutral-type dynamics system is the mathematical model to describe these effect. Recently, neutral-type systems have been widely concerned by researchers. In [10], the exponential passive filter was designed for the stochastic neutral-type neural networks. In [1], the stability of neutral neural system was studied by using the improved Lyapunov function, which can be applied to more general neural network models.

Sliding mode control (SMC) is an effective control strategy for stochastic neutral-type systems because of its fast response, insensitivity to system parameters and insensitivity to external disturbances. The SMC method [6, 14] can be used to study many systems, such as neutral-type system, uncertain systems, and Markovian switching systems. In [5], the sliding mode control for neutral-type stochastic systems was studied, which can guarantee the error system and

sliding mode dynamics to be asymptotically stochastic stable. In [9], the sliding mode observer was designed for uncertain systems by the event-triggered strategy. In [9], the stabilization problems was investigated for singular Markovian systems by using the SMC method.

In the nonlinear system [3,4], the Lévy model can more accurately describe the evolution process of the system, which often happen in finance and statistics, and so on. In [11], the feedback controller was designed for stochastic neural networks with Lévy noises, and the adaptive update law was obtained. In [2], according to the  $M$ -matrix method, the adaptive state estimator was obtained for neutral-type systems with Lévy noises.

According to the above discussion, we will provide a new sliding mode control method for stochastic neutral-type systems with Lévy noises. First, it contains Lévy noises in stochastic neutral-type systems, which can more accurately reflect external noises. Second, a SMC law of stochastic systems is established. Third, according to Lyapunov theory, the sliding mode controller is designed for stochastic neutral-type systems with Lévy noises.

## 2 Preliminaries

Consider a neutral-type systems (NTSs) [8,12,15] with stochastic noises

$$\begin{aligned} & d[\eta(t) - C\eta(t - \varsigma)] \\ &= \left\{ A\eta(t) + B\eta(t - \varrho) + D(u(t) + f(\eta(t), \eta(t - \varrho), t)) \right\} dt \\ &+ M_1\sigma(\eta(t), \eta(t - \varrho))d\omega + M_2 \int_{\mathbb{R}} \mu(\eta(t), \eta(t - \varrho), z)N(dt, dz), \end{aligned} \quad (1)$$

where  $\eta(t) \in \mathbb{R}^n$ ,  $C \in \mathbb{R}^{n \times n}$ ,  $A \in \mathbb{R}^{n \times n}$ ,  $B \in \mathbb{R}^{n \times n}$ ,  $D \in \mathbb{R}^{n \times m}$ ,  $M_1 \in \mathbb{R}^{n \times n}$  and  $M_2 \in \mathbb{R}^{n \times n}$ . The control input  $u(t) \in \mathbb{R}^m$ . The nonlinear function  $f(\cdot) \in \mathbb{R}^m$ . The Brownian moment  $\omega$  belongs to  $\mathbb{R}^m$ . The noise intensity matrix  $\sigma(\cdot) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{S} \rightarrow \mathbb{R}^{n \times m}$ . The 1-D  $\mathcal{F}_t$ -adapted Poisson random measure  $N(t, z) \in [0, +\infty) \times \mathbb{R}$ .  $\mu(\cdot) : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$  is a continuous function.  $\varsigma$  and  $\varrho$  are the constant delays, and  $d = \max\{\varsigma, \varrho\}$ .

Simply, set  $\eta_{\varsigma}(t) = \eta(t - \varsigma)$ ,  $\eta_{\varrho}(t) = \eta(t - \varrho)$ ,  $f(t) = f(\eta(t), \eta(t - \varrho))$ ,  $\sigma(\eta(t), \eta(t - \varrho)) = \sigma$ , and  $\mu(\eta(t), \eta(t - \varrho), z) = \mu$ , respectively. Thus, system (1) is

$$\begin{aligned} & d[\eta(t) - C\eta_{\varsigma}(t)] \\ &= \left[ A\eta(t) + B\eta_{\varrho}(t) + D(u(t) + f(t)) \right] dt + M_1\sigma d\omega + M_2 \int_{\mathbb{R}} \mu N(dt, dz), \end{aligned} \quad (2)$$

**Definition 1** [13,17]. *System (2) is the second moment exponentially stable suppose that*

$$\lim_{T \rightarrow \infty} \mathbb{E} \int_0^T \|\eta(t; \Xi(0), \phi(0))\|^2 dt < \infty,$$

where  $\Xi \in \mathbb{L}_{\mathcal{F}_0}^2([-d, 0], \mathbb{R}^n)$ .

**Assumption 1** [16]. *The function  $f(\cdot)$  satisfies*

$$\|f(\eta(t), \eta_\varrho(t))\| \leq \bar{\theta}\|\eta(t)\| + \bar{\vartheta}\|\eta_\varrho(t)\|,$$

where  $\bar{\theta} > 0$  and  $\bar{\vartheta} > 0$ .

**Assumption 2** [12]. *There exist matrices  $Y_1 \geq 0$ ,  $Y_2 \geq 0$  meeting*

$$\text{trace}\left\{\sigma^T(\eta, \eta_\varrho) M_1^T M_1 \sigma(\eta, \eta_\varrho)\right\} \leq \eta^T Y_1 \eta + \eta_\varrho^T Y_2 \eta_\varrho.$$

**Assumption 3** [18]. *There exist matrices  $Q_1 \geq 0$ ,  $Q_2 \geq 0$  and  $Q_3 \geq 0$  meeting*

$$\begin{aligned} & \int_{\mathbb{R}} \left[ (\eta - C\eta_\varsigma + M_2\mu)^T (\eta - C\eta_\varsigma + M_2\mu) - (\eta - C\eta_\varsigma)^T (\eta - C\eta_\varsigma) \right] v(dz) \\ & \leq \eta^T Q_1 \eta + \eta_\varsigma^T Q_2 \eta_\varsigma + \eta_\varrho^T Q_3 \eta_\varrho. \end{aligned}$$

### 3 Main Results

#### 3.1 Sliding Mode Surface (SMS)

Let the SMS be

$$\epsilon(t) = G[\eta(t) - C\eta_\varsigma(t)] - \int_0^t G[(A + DK)\eta(\xi) + B\eta_\varrho(\xi)] d\xi, \quad (3)$$

where  $G \in \mathbb{R}^{m \times n}$  satisfying  $GD$  is not singular and  $G\mathfrak{D} = 0$ , and  $\mathfrak{D} = [M_1 \ M_2]$ .  $K \in \mathbb{R}^{m \times n}$  satisfying the matrix  $A + DK$  is Hurwitz (see [7]).

According to (2) and (3), it has

$$\dot{\epsilon}(t) = -GDK\eta(t) + GD[u(t) + f(t)]. \quad (4)$$

Based on the SMC theory, let  $\dot{\epsilon}(t) = 0$ , and the control input (or the equivalent control law) is got

$$u(t) = K\eta(t) - f(t). \quad (5)$$

Then, by (5) and (2), it yields

$$\begin{aligned} & d[\eta(t) - C\eta_\varsigma(t)] \\ &= \left\{ (DK + A)\eta(t) + B\eta_\varrho(t) \right\} dt + M_1\sigma d\omega + M_2 \int_{\mathbb{R}} \mu N(dt, dz). \end{aligned} \quad (6)$$

Combining  $\dot{\epsilon}(t) = 0$ , system (6) is the dynamics of sliding motion of system (2).

### 3.2 Stability Analysis

**Theorem 1.** Assume that Assumptions 2 and 3 are true. If it can find matrices  $P_1 = P_1^T > 0$ ,  $P_2 > 0$ ,  $P_3 > 0$ ,  $P_4 > 0$ , and scalars  $\rho > 0$  meeting

$$\Omega = \begin{bmatrix} \Pi_{11} & \Pi_{12} & P_1 B \\ * & \Pi_{22} - C^T P_1 B & \\ * & * & \Pi_{33} \end{bmatrix} < 0, \quad (7)$$

$$P_1 < \rho I, \quad (8)$$

$$D^T P_1 \mathfrak{D} = 0, \quad (9)$$

with  $\Pi_{11} = P_1(A + DK) + (A + DK)^T P_1 + P_2 + P_3 + \varrho P_4 + \rho Y_1 + \rho Q_1$ ,  $\Pi_{12} = -(A + DK)^T P_1 C$ ,  $\Pi_{22} = -P_2 + \rho Q_2$ ,  $\Pi_{33} = -P_3 + \rho Y_2 + \rho Q_3$ , then by designing  $G = D^T P_1$ , system (6) is the exponentially stable.

*Proof:* Let the Lyapunov function be

$$U(\eta(t) - C\eta_\varsigma(t), t) = U_1 + U_2 + U_3 + U_4, \quad (10)$$

where  $U_1 = (\eta - C\eta_\varsigma)^T P_1(\eta - C\eta_\varsigma)$ ,  $U_2 = \int_{t-\varsigma}^t \eta^T(\xi) P_2 \eta(\xi) d\xi$ ,  $U_3 = \int_{t-\varrho}^t \eta^T(\xi) P_3 \eta(\xi) d\xi$ ,  $U_4 = \int_{-\varrho}^0 \int_{t+\nu}^t \eta^T(\xi) P_4 \eta(\xi) d\xi d\nu$ .

In the light of the Itô formula [18] and (6), one gets

$$\begin{aligned} \mathcal{L}U_1 = & 2[\eta - C\eta_\varsigma]^T P_1([A + DK]\eta + B\eta_\varrho) + \text{trace}\left[\sigma^T M_1^T P_1 M_1 \sigma\right] \\ & + \int_{\mathbb{R}} \left[ (\eta - C\eta_\varsigma + M_2 \mu)^T P_1 (\eta - C\eta_\varsigma + M_2 \mu) \right. \\ & \left. - (\eta - C\eta_\varsigma)^T P_1 (\eta - C\eta_\varsigma) \right] v(dz). \end{aligned} \quad (11)$$

By (8), Assumptions 2 and 3, we have

$$\text{trace}\left[\sigma^T M_1^T P_1 M_1 \sigma\right] \leq \rho(\eta^T(t)Y_1\eta(t) + \eta_\varrho^T(t)Y_2\eta_\varrho(t)), \quad (12)$$

$$\begin{aligned} & \int_{\mathbb{R}} \left[ (\eta - C\eta_\varsigma + M_2 \mu)^T P_1 (\eta - C\eta_\varsigma + M_2 \mu) \right. \\ & \left. - (\eta - C\eta_\varsigma)^T P_1 (\eta - C\eta_\varsigma) \right] v(dz) \\ & \leq \rho(\eta^T Q_1 \eta + \eta_\varsigma^T Q_2 \eta_\varsigma + \eta_\varrho^T(t) Q_3 \eta_\varrho). \end{aligned} \quad (13)$$

According to (10), it yields

$$\mathcal{L}U_2 = \eta^T(t)P_2\eta(t) - \eta_\varsigma^T(t)P_2\eta_\varsigma(t). \quad (14)$$

Similarly, based on (10), it has

$$\mathcal{L}U_3 = \eta^T(t)P_3\eta(t) - \eta_\varrho^T(t)P_3\eta_\varrho(t). \quad (15)$$

In addition,

$$\mathcal{L}U_4 \leq \varrho\eta^T(t)P_4\eta(t) - \int_{t-\varrho}^t \eta^T(\xi)P_4\eta(\xi)d\xi. \quad (16)$$

By (11)–(16), we get

$$\mathcal{L}U = \mathcal{L}U_1 + \mathcal{L}U_2 + \mathcal{L}U_3 + \mathcal{L}U_4 \leq \begin{bmatrix} \eta^T(t) & \eta_\varsigma^T(t) & \eta_\varrho^T(t) \end{bmatrix} \Omega \begin{bmatrix} \eta(t) \\ \eta_\varsigma(t) \\ \eta_\varrho(t) \end{bmatrix}, \quad (17)$$

where

$$\Omega = \begin{bmatrix} \Pi_{11} & \Pi_{12} & P_1B \\ * & \Pi_{22} - C^TP_1B & * \\ * & * & \Pi_{33} \end{bmatrix},$$

with  $\Pi_{11} = P_1(A + DK) + (A + DK)^TP_1 + P_2 + P_3 + \varrho P_4 + \rho Y_1 + \rho Q_1$ ,  $\Pi_{22} = -P_2 + \rho Q_2$ ,  $\Pi_{33} = -P_3 + \rho Y_2 + \rho Q_3$ ,  $\Pi_{12} = -(A + DK)^TP_1C$ .

According to (7), we get  $\Omega < 0$ .

By (17), it gets

$$\mathcal{L}U \leq -\Lambda \|\eta(t)\|^2, \quad (18)$$

where  $-\Lambda = \lambda_{\max}(\Omega)$  ( $\Lambda > 0$ ),  $\lambda_{\max}(\Omega)$  shows the maximum eigenvalue of  $\Omega$ .

Based on the Dynkin's formula [18], it yields

$$-\mathbb{E} \int_0^T \mathcal{L}U dt = \mathbb{E}U_0 - \mathbb{E}U_T \leq \mathbb{E}U_0. \quad (19)$$

In the light of (18) and (19), we have

$$\mathbb{E} \int_0^T \|\eta(t)\|^2 dt \leq \frac{1}{\Lambda} \mathbb{E}U_0 < \infty.$$

Based on Definition 1, system (6) is the exponentially stable.  $\square$

*Remark 1:* Theorem 1 contains the condition  $D^TP_1\mathfrak{D} = 0$ . According to the method in [7], it can be converted to  $\text{trace}[(D^TP_1\mathfrak{D})^T(D^TP_1\mathfrak{D})] = 0$ . By the condition  $(D^TP_1\mathfrak{D})^T(D^TP_1\mathfrak{D}) < \zeta I$ , where  $\zeta > 0$  is small enough, and based on the Schur complement lemma, we have

$$\begin{bmatrix} -\zeta I & (\mathfrak{D}^TP_1D)^T \\ D^TP_1\mathfrak{D} & -I \end{bmatrix} < 0. \quad (20)$$

Then, the solution problem is changed into the following minimization problem

$$\begin{aligned} & \min \zeta, \\ & \text{subject to (7) – (8) and (20).} \end{aligned} \quad (21)$$

### 3.3 Reachability Analysis

**Theorem 2:** For system (2), the SMC law is designed by

$$u(t) = u_1(t) - u_2(t) - u_3(t), \quad (22)$$

where  $u_1(t) = K\eta(t)$ ,  $u_2(t) = (\theta(t)\|\eta(t)\| + \vartheta(t)\|\eta_\vartheta(t)\|)sign((\epsilon^T(t)GD)^T)$ ,  $u_3(t) = \frac{\varepsilon}{\|GD\|}sign((\epsilon^T(t)GD)^T)$ ,  $\varepsilon > 0$ ,  $\theta(t)$  and  $\vartheta(t)$  are devoted to estimate  $\bar{\theta}$  and  $\bar{\vartheta}$ , respectively, which are showed by

$$\begin{aligned} \tilde{\theta}(t) &= \theta(t) - \bar{\theta}, \quad \dot{\tilde{\theta}}(t) = \|\epsilon^T(t)\|\|GD\|\|\eta(t)\|, \\ \tilde{\vartheta}(t) &= \vartheta(t) - \bar{\vartheta}, \quad \dot{\tilde{\vartheta}}(t) = \|\epsilon^T(t)\|\|GD\|\|\eta_\vartheta(t)\|, \end{aligned} \quad (23)$$

$\tilde{\theta}(t)$  and  $\tilde{\vartheta}(t)$  are estimate errors.

*Proof:* Let the Lyapunov function be

$$U = \frac{1}{2}(\epsilon^T(t)\epsilon(t) + \tilde{\theta}^2(t) + \tilde{\vartheta}^2(t)). \quad (24)$$

Combining (4), (22), (23) and (24), it gets

$$\begin{aligned} \dot{U} &= \epsilon^T(t)\dot{\epsilon}(t) + \tilde{\theta}(t)\dot{\tilde{\theta}}(t) + \tilde{\vartheta}(t)\dot{\tilde{\vartheta}}(t) \\ &= \epsilon^T(t)[-GDK\eta(t) + GDU(t) + GDF(t)] + \tilde{\theta}(t)\dot{\tilde{\theta}}(t) + \tilde{\vartheta}(t)\dot{\tilde{\vartheta}}(t) \\ &\leq -\epsilon^T(t)GDU_2(t) - \epsilon^T(t)GDU_3(t) + \epsilon^T(t)GDF(t) \\ &\quad + \tilde{\theta}(t)\|\epsilon^T(t)\|\|GD\|\|\eta(t)\| + \tilde{\vartheta}(t)\|\epsilon^T(t)\|\|GD\|\|\eta_\vartheta(t)\|. \end{aligned} \quad (25)$$

Based on Assumption 1 and (23), we have

$$\begin{aligned} &- \epsilon^T(t)GDU_2(t) + \epsilon^T(t)GDF(t) \\ &+ \tilde{\theta}(t)\|\epsilon^T(t)\|\|GD\|\|\eta(t)\| + \tilde{\vartheta}(t)\|\epsilon^T(t)\|\|GD\|\|\eta_\vartheta(t)\| \\ &\leq -\|\epsilon^T(t)GD\|(\theta(t)\|\eta(t)\| + \vartheta(t)\|\eta_\vartheta(t)\|) \\ &+ \tilde{\theta}(t)\|\epsilon^T(t)\|\|GD\|\|\eta(t)\| + \tilde{\vartheta}(t)\|\epsilon^T(t)\|\|GD\|\|\eta_\vartheta(t)\| + \epsilon^T(t)GDF(t) \\ &= -\bar{\theta}\|\epsilon^T(t)\|\|GD\|\|\eta(t)\| - \bar{\vartheta}\|\epsilon^T(t)\|\|GD\|\|\eta_\vartheta(t)\| + \epsilon^T(t)GDF(t) \\ &\leq 0. \end{aligned} \quad (26)$$

By (22), (25) and (26), it obtains

$$\begin{aligned} \dot{U} &\leq -\epsilon^T(t)GDU_3(t) \\ &\leq -\|\epsilon^T(t)GD\| \cdot \frac{\varepsilon}{\|GD\|} \\ &\leq -\varepsilon\|\epsilon(t)\| < 0, \quad \text{for } \|\epsilon(t)\| \neq 0, \end{aligned} \quad (27)$$

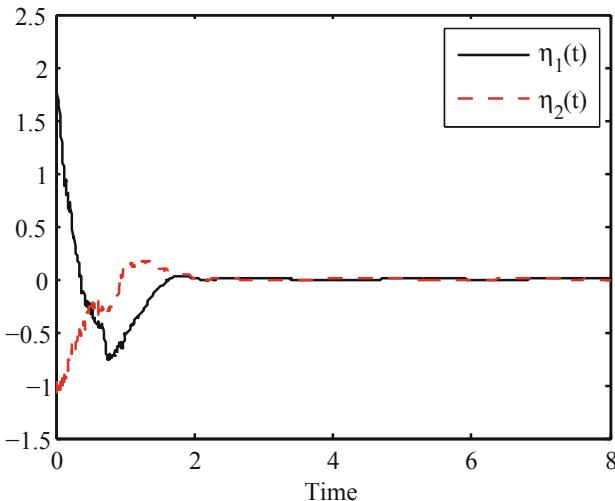
where  $\varepsilon > 0$ .

By (23),  $\dot{\theta}(t) = \dot{\tilde{\theta}}(t) = \|\epsilon^T(t)\| \|GD\| \|\eta(t)\|$  and  $\dot{\vartheta}(t) = \dot{\tilde{\vartheta}}(t) = \|\epsilon^T(t)\| \|GD\| \|\eta_\varrho(t)\|$  are greater than or equal to zero, then  $\dot{\theta}(t)$  and  $\dot{\vartheta}(t)$  are monotonous increasing. Obviously, we have  $\tilde{\theta}(t)\dot{\tilde{\theta}}(t) \geq 0$  and  $\tilde{\vartheta}(t)\dot{\tilde{\vartheta}}(t) \geq 0$ . According to (25), it has  $\epsilon^T(t)\dot{\epsilon}(t) < 0$ . Thus, the reaching condition of (3) is satisfied.  $\square$

## 4 Simulations

Consider STSs (1) with the corresponding parameters as follows

$$\begin{aligned} C &= \begin{bmatrix} 0.8 & 0 \\ 0 & 0.4 \end{bmatrix}, \quad A = \begin{bmatrix} -1.6 & 0.5 \\ -0.3 & 1.7 \end{bmatrix}, \quad B = \begin{bmatrix} -0.3 & 1.1 \\ 0.4 & -1.2 \end{bmatrix}, \\ D &= [0.4 \ -0.6]^T, \quad M_1 = \begin{bmatrix} 0.3 & -0.5 \\ 0.6 & -0.3 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 0.6 & -0.4 \\ 0.3 & -0.6 \end{bmatrix}, \\ f(t) &= 0.1 \cos(t)(\eta(t) + \eta_\varrho(t)), \quad Y_1 = \frac{1}{10}I_2, \quad Y_2 = \frac{1}{60}I_2, \\ Q_1 &= Q_2 = Q_3 = \frac{1}{40}I_2, \quad \sigma = 0.2(\eta(t) + \eta_\varrho(t)), \\ \mu &= 0.2z(\eta(t) + 0.2\eta_\varrho(t)), \quad \varsigma = \varrho = 0.5, \quad \eta(0) = [2, -1]^T, \quad K = [0.3 \ 0.6]. \end{aligned}$$



**Fig. 1.** State trajectories of system (2) under the SMC.

After calculation, Assumptions 2 and 3 can be satisfied. Based on LMIs (7)–(9), it gets

$$P_1 = \begin{bmatrix} 0.5017 & 0.0434 \\ 0.0434 & -0.4978 \end{bmatrix}, P_2 = \begin{bmatrix} -0.0320 & -0.5642 \\ -0.5642 & 0.6008 \end{bmatrix},$$

$$P_3 = \begin{bmatrix} 0.9580 & -0.7991 \\ -0.7991 & 0.8934 \end{bmatrix}, P_4 = \begin{bmatrix} -0.2264 & -0.5521 \\ -0.5521 & 0.3831 \end{bmatrix},$$

and  $\rho = 2.0147$ ,  $\zeta = 3.3082 \times 10^{-3}$ . And the SMS function  $\epsilon(t) = [0.012 \ 0.013][\eta(t) - C\eta_s(t)] - \int_0^t [-0.15 \ 0.21]\eta(\xi) - [0.01 \ 0.02]\eta_\varrho(\xi)d\xi$ . Then the SMC law (22) can be designed in Theorem 2. Therefore, NTSSs (2) is exponentially stable. Figure 1 shows the simulation results, which verifies the effectiveness of the results.

## 5 Conclusion

In this paper, according to the SMC method, the exponentially stable is studied for NTSSs with stochastic noises and Lévy noises. The dynamic characteristics of sliding mode can be obtained immediately by designing appropriate integrated SMS. In addition, the adaptive SMC law can guarantee the reachability of SMS. which can ensure that the system state is tracked.

**Acknowledgements.** This work is partially supported by the National Natural Science Foundation of China (61673257; 61573095), and the China Postdoctoral Science Foundation (2019M661322).

## References

1. Arik, S.: New criteria for stability of neutral-type neural networks with multiple time delays. *IEEE Trans. Neural Netw. Learn. Syst.* **31**(5), 1504–1513 (2020)
2. Chen, Q., Tong, D., Zhou, W., Xu, Y.: Adaptive exponential state estimation for Markovian jumping neural networks with multi-delays and Lévy noises. *Circ. Syst. Sig. Process.* **38**(7), 3321–3339 (2019)
3. Jia, Y.: Robust control with decoupling performance for steering and traction of 4WS vehicles under velocity-varying motion. *IEEE Trans. Control Syst. Technol.* **8**(3), 554–569 (2000)
4. Jia, Y.: Alternative proofs for improved LMI representations for the analysis and the design of continuous-time systems with polytopic type uncertainty: a predictive approach. *IEEE Trans. Autom. Control* **48**(8), 1413–1416 (2003)
5. Kao, Y., Xie, J., Wang, C., Karimi, H.R.: A sliding mode approach to  $H_\infty$  non-fragile observer-based control design for uncertain Markovian neutral-type stochastic systems. *Automatica* **52**, 218–226 (2015)
6. Li, F., Shi, P., Wu, L.: Control and Filtering for Semi-Markovian Jump Systems. Springer, Heidelberg (2017)
7. Li, H., Shi, P., Yao, D., Wu, L.: Observer-based adaptive sliding mode control for nonlinear Markovian jump systems. *Automatica* **64**, 133–142 (2016)

8. Li, Z.Y., Lam, J., Wang, Y.: Stability analysis of linear stochastic neutral-type time-delay systems with two delays. *Automatica* **91**, 179–189 (2018)
9. Liu, X., Su, X., Shi, P., Shen, C.: Observer-based sliding mode control for uncertain fuzzy systems via event-triggered strategy. *IEEE Trans. Fuzzy Syst.* **27**(11), 2190–2201 (2019)
10. Shi, P., Li, F., Wu, L., Lim, C.: Neural network-based passive filtering for delayed neutral-type semi-Markovian jump systems. *IEEE Trans. Neural Netw.* **28**(9), 2101–2114 (2017)
11. Sun, Y., Zhang, Y., Zhou, W., Zhou, J., Zhang, X.: Adaptive exponential stabilization of neutral-type neural network with Lévy noise and Markovian switching parameters. *Neurocomputing* **284**, 160–170 (2018)
12. Tong, D., Chen, Q., Zhou, W., Xu, Y.: Adaptive state estimation of Markov switched neural networks driven by Lévy noise. *Trans. Inst. Measur. Control* **42**(2), 330–336 (2020)
13. Tong, D., Xu, C., Chen, Q., Zhou, W.: Sliding mode control of a class of nonlinear systems. *J. Franklin Inst.* **357**(3), 1560–1581 (2020)
14. Tong, D., Xu, C., Chen, Q., Zhou, W., Xu, Y.: Sliding mode control for nonlinear stochastic systems with Markovian jumping parameters and mode-dependent time-varying delays. *Nonlinear Dyn.* **100**(2), 1343–1358 (2020)
15. Xu, C., Tong, D., Chen, Q., Zhou, W., Shi, P.: Exponential stability of Markovian jumping systems via adaptive sliding mode control. *IEEE Trans. Syst. Man Cybern. Syst.* (2019). DOI10.1109/TSMC.2018.2884565
16. Xu, Y., Zhou, W., Zhang, J., Sun, W., Tong, D.: Topology identification of complex delayed dynamical networks with multiple response systems. *Nonlinear Dyn.* **88**(4), 2969–2981 (2017)
17. Yan, X., Tong, D., Chen, Q., Zhou, W., Xu, Y.: Adaptive state estimation of stochastic delayed neural networks with fractional Brownian motion. *Neural Process. Lett.* **50**(2), 2007–2020 (2019)
18. Zhou, L., Zhu, Q., Wang, Z., Zhou, W., Su, H.: Adaptive exponential synchronization of multiscale time-delayed recurrent neural networks with Lévy noise and regime switching. *IEEE Trans. Neural Netw. Learn. Syst.* **28**(12), 2885–2898 (2017)



# Reinforcement Learning Adaptive Tracking Control for a Stratospheric Airship

Kang Wang<sup>1</sup>, Yang Liu<sup>1</sup>, Zewei Zheng<sup>1(✉)</sup>, and Ming Zhu<sup>2</sup>

<sup>1</sup> The Seventh Research Division, School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, People's Republic of China  
zeiweizheng@buaa.edu.cn

<sup>2</sup> Institute of Unmanned System, Beihang University, Beijing 100191, People's Republic of China

**Abstract.** This paper investigates the optimal performance control problem for the trajectory tracking control for a stratospheric airship with external disturbance. A reinforcement learning adaptive tracking control for a stratospheric airship is proposed. First, according to the knowledge of dynamics and kinematics, we establish the model of a stratospheric airship used in this paper. Then, to solve external disturbance problem and enhance the system performance, a controller is proposed by means of a reinforcement learning (RL) method that is primarily based on two neural networks (NNs). In the last place, the stability analysis and numerical simulations are given to verify that the designed controller is effective.

**Keywords:** Stratospheric airship · Reinforcement learning · Actor-Critic · Adaptive control · Trajectory tracking

## 1 Introduction

In recent decades, with the development of stratospheric resources, the demand for airship is increasing intensely. It owns the merits of large load capacity, long flying time and appropriate flight altitude. As a result, it can be applied to communication delay, space observation, military uses and other applications.

It is necessary to accurately track reference trajectory for fulfilling the task. Due to the characteristics of stratospheric airships such as nonlinearity and vulnerability, tracking control has become a challenge in the airship flight control.

This literature [1] applied the trajectory linearization control theory to design a controller, which can track the predefined trajectory. A dual time-scale filtering technique and backstepping method were applied to realize tracking of wheeled mobile robot, which can enhance the robustness of systems [2]. In [3], an adaptive neural network tracking control was proposed for a specific class of multiple-input-multiple-output systems. This literature [4] used backstepping approach

to design an adaptive neural network tracking controller combined with a disturbance observer, while it was only available for a specific class of strict-feedback nonlinear systems. In combination with neural network and adaptive law, the backstepping method can be a powerful control technique to ensure precisely tracking reference trajectory [5–7].

However, the optimal control problem of stratospheric airship has not been sufficient considered. Optimal control can achieve tracking control and optimize the performance of the control system simultaneously. Generally, only by solving the Hamilton-Jacobi-Bellman (HJB) equation can we seek the optimal control solution of nonlinear systems. Nevertheless, it is virtually impossible to resolve the HJB equation directly. Among many methods that are employed to solve the problem, the RL method has become a trend of technique that can solve the HJB problem to obtain the optimal control, which is in combination with actor-critic neural networks [8–10]. The literature [11] proposed an reliable reinforcement learning adaptive robust tracking control scheme for a specific class of nonlinear systems with uncertainties, which employed an improved adaptive critic technique instead of traditional techniques. The literature [12] provided a novel controller via an actor-critic design to improve the control performance for a class of nonlinear system with external disturbance. In [13], an actor-critic neural network RL scheme was designed for a nonlinear system with dead-zone input and unknown functions.

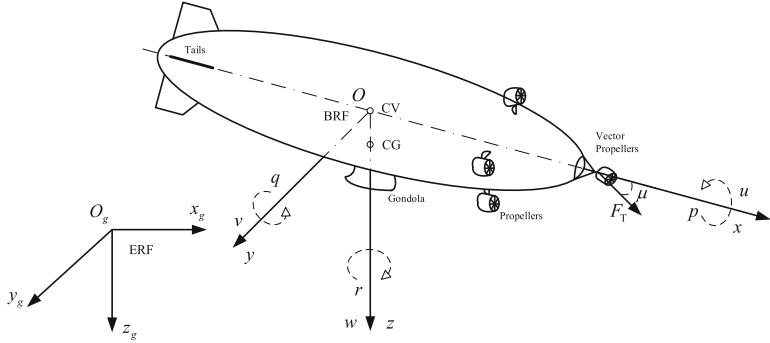
In this paper, a critic function that consist of primary and secondary critic signal is proposed. A RL method is proposed to obtain the optimal control system performance by approximating and compensating the nonlinearity of the system and unknown disturbances. Combined with six degrees of freedom kinematics and dynamics equations, a controller by means of actor-critic neural networks is designed and simulated. The organization of this paper is constructed as follows. In Sect. 2, we introduce the model of the helium-filled airships investigated throughout in the paper and establish the state space equations, at the same time some assumptions and lemmas are given. In the next section, we describe the design process in detail. In Sect. 4, numerical simulation results and some comparisons are presented. In the last section, a concise and to the point conclusion is drawn.

## 2 Problems Formulation

### 2.1 Modeling

As the Fig. 1 shows, the helium-filled airships investigated throughout the paper are composed of actuators and auxiliary devices. There are propulsion propellers installing at the front of the balloon to provide thrust and steering moment.

First, we must define the reference frames to model the stratospheric airship. The earth reference frame (ERF) is fixed to the origin  $O_g$  on the ground, which is located at a point. The z-axis points down, x-axis points north, and the y-axis points east. The body reference frame (BRF) is fixed to the origin  $O$ , which is



**Fig. 1.** The airship model

coincident with the center of the volume of the helium-filled airships. The z-axis points downward, x-axis points head, and the y-axis points right.

$\mathbf{P} = [x, y, z]^T$  and  $\boldsymbol{\Theta} = [\phi, \theta, \psi]^T$  present the airship position and attitude in ERF,  $\mathbf{v} = [u, v, w]^T$  and  $\boldsymbol{\Omega} = [p, q, r]^T$  denote the airship airspeed and angular velocity in BRF, respectively. According to the knowledge of kinematics and dynamics, we can establish the following equations for airship motion [1].

$$\begin{bmatrix} \dot{\mathbf{P}} \\ \dot{\boldsymbol{\Theta}} \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{0}_{3 \times 3} \\ \mathbf{0}_{3 \times 3} & \mathbf{K} \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\Omega} \end{bmatrix} \quad (1)$$

$$\begin{bmatrix} m\mathbf{E} + \mathbf{M}' & -mr'_C \times \\ mr'_C \times & \mathbf{I}_O + \mathbf{I}'_O \end{bmatrix} \begin{bmatrix} \dot{\mathbf{v}} \\ \dot{\boldsymbol{\Omega}} \end{bmatrix} + \begin{bmatrix} (m\mathbf{E} + \mathbf{M}') \boldsymbol{\Omega} \times \mathbf{v} + m\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}'_C) \\ \boldsymbol{\Omega} \times (\mathbf{I}_O \boldsymbol{\Omega}) + mr'_C \times (\boldsymbol{\Omega} \times \mathbf{v}) \end{bmatrix} = \begin{bmatrix} (G - B) \mathbf{R}^T \mathbf{e}_z + \mathbf{F}_a + \mathbf{F}_T \\ \mathbf{M}_G + \mathbf{M}_B + \mathbf{M}_a + \mathbf{M}_T \end{bmatrix} \quad (2)$$

where  $\mathbf{E}$ ,  $m$ ,  $\mathbf{M}'$ ,  $\mathbf{I}'_O$ ,  $\mathbf{r}'_C \times$ ,  $\mathbf{I}_O$ ,  $\mathbf{e}_z$ ,  $\mathbf{G}$ ,  $\mathbf{B}$ ,  $\mathbf{M}_B$ ,  $\mathbf{M}_G$ ,  $\mathbf{F}_a$ ,  $\mathbf{M}_a$ ,  $\mathbf{F}_T$  and  $\mathbf{M}_T$  are the detailed physical parameters and functions of the airship, which can be further understood in this article [7].

The dynamic Eqs. (1) and (2) can be deformed as

$$\begin{cases} \dot{\mathbf{P}} = \mathbf{K}\mathbf{v} \\ \dot{\mathbf{v}} = \mathbf{F}_v + \mathbf{B}_{12}\boldsymbol{\tau}_\omega + \mathbf{B}_{11}\boldsymbol{\tau}_v + \mathbf{f}_v \\ \dot{\boldsymbol{\Theta}} = \mathbf{R}\boldsymbol{\Omega} \\ \dot{\boldsymbol{\Omega}} = \mathbf{F}_\omega + \mathbf{B}_{22}\boldsymbol{\tau}_\omega + \mathbf{B}_{21}\boldsymbol{\tau}_v + \mathbf{f}_\omega \end{cases} \quad (3)$$

where  $\mathbf{F}_v$ ,  $\mathbf{F}_\omega$ ,  $\mathbf{B}_{12}$ ,  $\mathbf{B}_{11}$ ,  $\mathbf{B}_{22}$ ,  $\mathbf{B}_{21}$  are the detailed state-space functions,  $\mathbf{f}_v$  and  $\mathbf{f}_\omega$  are the disturbances and  $\boldsymbol{\tau} = [\boldsymbol{\tau}_v, \boldsymbol{\tau}_\omega]^T = [\tau_u, \tau_v, \tau_w, \tau_p, \tau_q, \tau_r]^T$  presents the system input [7].

The control aim in this paper is to design suitable control input  $\boldsymbol{\tau}$  for the helium-filled stratospheric airships to ensure that the output  $\mathbf{X}$  can accurately

track the reference trajectory  $\mathbf{X}_d$ .  $\mathbf{X} = [\mathbf{P}, \boldsymbol{\Theta}]^T = [x, y, z, \phi, \theta, \psi]^T$  represents the system output,  $\mathbf{X}_d = [\mathbf{P}_d, \boldsymbol{\Theta}_d]^T = [x_d, y_d, z_d, \phi_d, \theta_d, \psi_d]^T$  represents the desired trajectory. Moreover, the tracking error finally tends close to zero.

## 2.2 Some Assumptions and Assumptions

**Assumption 1.** *The predetermined trajectory is a known bounded smooth function.*

**Assumption 2.** *The unknown external disturbance satisfies the condition that  $\|D\| \leq \bar{D}$  always holds, where  $\bar{D}$  is a positive constant.*

**Lemma 1.** *For a given vector  $V \in \mathbf{R}^m$  and a positive constant  $\iota_v$  that is selected in advance, the following inequality are satisfied*

$$-\tanh^T(V/\iota_v)V \leq -\|v\| + mk\iota_v \quad (4)$$

where  $k$  is a constant that satisfies  $k = e^{-(k+1)}$ , i.e.,  $k = 0.2785$

**Lemma 2.** *It is known that radial basic function neural network can be applied to function approximation. For any continuous function  $\varphi(X) : \mathbf{R}^n \rightarrow \mathbf{R}$  defined on a compact set  $\Omega_z$ , it can be approximated in the following form*

$$\varphi(X) = W^T \sigma(X) + \varepsilon(X)$$

where  $X \in \Omega_z$  is the input vector,  $W \in \mathbf{R}^m$   $\sigma(X)$ ,  $\varepsilon(X)$  are the optimal weight vector, the basis function vector and the optimal approximation error. Additionally, it has been proved that  $\varepsilon(X)$  is bounded by a positive constant  $\bar{\varepsilon}$ .

## 3 Controller Design

In the beginning of this section, the tracking errors are defined in combination with the kinematics and dynamics equations. Subsequently, a continuous nonlinear complicate functions is approximated by an actor neural network for enhancing the control performance. In addition, a critical function based on a critic neural network are established to estimate and enhance the system performance and update the control action. An adaptive controller is proposed for the sake of reducing the influence of neural network reconstruction error and disturbance, which is constructed with special structure. Finally, in combination with above designed controllers, the actual control inputs are designed to ensure that the tracking errors converge close to sufficient small neighborhoods of zero and all the signals of the closed-loop system are uniformly ultimately bounded (UUB).

### 3.1 Tracking Error Definitions

In this section, the appropriate tracking errors are defined in combination with the kinematics and dynamics equations. According to the Eq.(3), we can calculate

$$\begin{aligned}\ddot{\mathbf{P}} &= \dot{\mathbf{K}}\mathbf{v} + \mathbf{K}\dot{\mathbf{v}} \\ &= \dot{\mathbf{K}}\mathbf{v} + \mathbf{K}(\mathbf{F}_v + \mathbf{B}_{12}\boldsymbol{\tau}_\omega + \mathbf{B}_{11}\boldsymbol{\tau}_v + \mathbf{f}_v) \\ &= \dot{\mathbf{K}}\mathbf{v} + \mathbf{K}\mathbf{F}_v + \mathbf{K}\mathbf{B}_{12}\boldsymbol{\tau}_\omega + \mathbf{K}\mathbf{B}_{11}\boldsymbol{\tau}_v + \mathbf{K}\mathbf{f}_v \\ &= \mathbf{F}_{v,0} + \mathbf{B}_{v,0}\boldsymbol{\tau}_v + \mathbf{d}_v\end{aligned}\quad (5)$$

where  $\mathbf{F}_{v,0} = \dot{\mathbf{K}}\mathbf{v} + \mathbf{K}\mathbf{F}_v$ ,  $\mathbf{B}_{v,0} = \mathbf{K}\mathbf{B}_{11}$ ,  $\mathbf{d}_v = \mathbf{B}_{11}\boldsymbol{\tau}_\omega + \mathbf{K}\mathbf{f}_v$ .

Define the tracking velocity error as below

$$\mathbf{e}_{fv} = \dot{\mathbf{P}}_e + \mathbf{k}_v \mathbf{P}_e \quad (6)$$

where  $\mathbf{P}_e = \mathbf{P} - \mathbf{P}_d$ ,  $\mathbf{k}_v = \text{diag}\{k_{v,x}, k_{v,y}, k_{v,z}\}$  is a positive definite constant matrix.

This following equation can be deduced

$$\begin{aligned}\dot{\mathbf{e}}_{fv} &= \ddot{\mathbf{P}}_e + \mathbf{k}_v \dot{\mathbf{P}}_e \\ &= \ddot{\mathbf{P}}_e - \ddot{\mathbf{P}}_d + \mathbf{k}_v (\dot{\mathbf{P}}_e - \dot{\mathbf{P}}_d) \\ &= \mathbf{F}_{v,0} + \mathbf{k}_v \dot{\mathbf{P}}_e + \mathbf{B}_{v,0}\boldsymbol{\tau}_v - (\ddot{\mathbf{P}}_d + \mathbf{k}_v \dot{\mathbf{P}}_d) + \mathbf{d}_v\end{aligned}\quad (7)$$

According to the Eq. (3)

$$\begin{aligned}\ddot{\boldsymbol{\Theta}} &= \dot{\mathbf{R}}\boldsymbol{\Omega} + \mathbf{R}\dot{\boldsymbol{\Omega}} \\ &= \dot{\mathbf{R}}\boldsymbol{\Omega} + \mathbf{R}(\mathbf{F}_\omega + \mathbf{B}_{22}\boldsymbol{\tau}_\omega + \mathbf{B}_{21}\boldsymbol{\tau}_v + \mathbf{f}_\omega) \\ &= \dot{\mathbf{R}}\boldsymbol{\Omega} + \mathbf{R}\mathbf{F}_\omega + \mathbf{R}\mathbf{B}_{22}\boldsymbol{\tau}_\omega + \mathbf{R}\mathbf{B}_{21}\boldsymbol{\tau}_v + \mathbf{R}\mathbf{f}_\omega \\ &= \mathbf{F}_{\omega,0} + \mathbf{B}_{\omega,0}\boldsymbol{\tau}_\omega + \mathbf{d}_\omega\end{aligned}\quad (8)$$

where  $\mathbf{F}_{\omega,0} = \dot{\mathbf{R}}\boldsymbol{\Omega} + \mathbf{R}\mathbf{F}_\omega$ ,  $\mathbf{B}_{\omega,0} = \mathbf{R}\mathbf{B}_{22}$ ,  $\mathbf{d}_\omega = \mathbf{B}_{21}\boldsymbol{\tau}_v + \mathbf{R}\mathbf{f}_\omega$ .

Give a definition of the tracking attitude error as follows

$$\mathbf{e}_{f\omega} = \dot{\boldsymbol{\Theta}}_e + \mathbf{k}_\omega \boldsymbol{\Theta}_e, \quad (9)$$

where  $\boldsymbol{\Theta}_e = \boldsymbol{\Theta} - \boldsymbol{\Theta}_d$ ,  $\mathbf{k}_\omega = \text{diag}\{k_{\omega,\phi}, k_{\omega,\theta}, k_{\omega,\psi}\}$  is a positive definite constant matrix.

This following equation can be deduced

$$\begin{aligned}\dot{\mathbf{e}}_{f\omega} &= \ddot{\boldsymbol{\Theta}}_e + \mathbf{k}_\omega \dot{\boldsymbol{\Theta}}_e \\ &= \ddot{\boldsymbol{\Theta}}_e - \ddot{\boldsymbol{\Theta}}_d + \mathbf{k}_\omega (\dot{\boldsymbol{\Theta}}_e - \dot{\boldsymbol{\Theta}}_d) \\ &= \mathbf{F}_{\omega,0} + \mathbf{k}_\omega \dot{\boldsymbol{\Theta}}_e + \mathbf{B}_{\omega,0}\boldsymbol{\tau}_\omega - (\ddot{\boldsymbol{\Theta}}_d + \mathbf{k}_\omega \dot{\boldsymbol{\Theta}}_d) + \mathbf{d}_\omega\end{aligned}\quad (10)$$

However, it is extremely intricate to work out the first-order derivation and the second-order derivation of the desired attitude  $\mathbf{P}_d$  and the desired velocity  $\boldsymbol{\Theta}_d$ . Consequently, a command filter is proposed to estimate them [7].

$$\begin{cases} \dot{\boldsymbol{\Xi}}_1 = \boldsymbol{\Xi}_2 \\ \dot{\boldsymbol{\Xi}}_2 = -2\Lambda\omega_n - \omega_n^2 \left( \boldsymbol{\Xi}_1 - \begin{bmatrix} \boldsymbol{\Theta}_d \\ \mathbf{P}_d \end{bmatrix} \right) \end{cases} \quad (11)$$

where the damping ratio  $\Lambda = \text{diag}\{\Lambda_x, \Lambda_y, \Lambda_z, \Lambda_\phi, \Lambda_\theta, \Lambda_\psi\}$  and the damping frequency  $\omega_n = \text{diag}\{\omega_{n,x}, \omega_{n,y}, \omega_{n,z}, \omega_{n,\phi}, \omega_{n,\theta}, \omega_{n,\psi}\}$  are selected. We employ  $\boldsymbol{\Xi}_1$  and  $\boldsymbol{\Xi}_2$  to take stock of the first-order derivation and the second-order derivation of the desired attitude and velocity, respectively.

Therefore, Eq. (7) and Eq. (10) can be rewritten as

$$\begin{cases} \dot{\mathbf{e}}_{f\omega} = \mathbf{F}_{\omega,0} + k_\omega \dot{\boldsymbol{\Theta}} + \mathbf{B}_{\omega,0} \boldsymbol{\tau}_\omega - (\hat{\boldsymbol{\Theta}}_d + k_\omega \hat{\dot{\boldsymbol{\Theta}}}_d) + d_\omega^* \\ \dot{\mathbf{e}}_{fv} = \mathbf{F}_{v,0} + k_v \dot{\mathbf{P}} + \mathbf{B}_{v,0} \boldsymbol{\tau}_v - (\hat{\mathbf{P}}_d + k_v \hat{\dot{\mathbf{P}}}_d) + d_v^* \end{cases} \quad (12)$$

where  $d_\omega^* = d_\omega + k_\omega (\dot{\boldsymbol{\Theta}}_d - \hat{\dot{\boldsymbol{\Theta}}}_d) + \ddot{\boldsymbol{\Theta}}_d - \hat{\ddot{\boldsymbol{\Theta}}}_d$  and  $d_v^* = d_v + k_v (\dot{\mathbf{P}}_d - \hat{\dot{\mathbf{P}}}_d) + \ddot{\mathbf{P}}_d - \hat{\ddot{\mathbf{P}}}_d$ .

*Remark 1.*  $d_\omega^*$  and  $d_v^*$  is bounded by an unknown constant according to the assumption 2.

Define  $\mathbf{e}_f = [\mathbf{e}_{fv}, \mathbf{e}_{f\omega}]^T = [e_{fx}, e_{fy}, e_{fz}, e_{f\phi}, e_{f\theta}, e_{f\psi}]^T$ , and equations can be rewritten as

$$\begin{aligned} \dot{\mathbf{e}}_f &= \begin{bmatrix} \mathbf{F}_{v,0} + k_v \dot{\mathbf{P}} + \mathbf{B}_{v,0} \boldsymbol{\tau}_v - (\hat{\mathbf{P}}_d + k_v \hat{\dot{\mathbf{P}}}_d) + d_v^* \\ \mathbf{F}_{\omega,0} + k_\omega \dot{\boldsymbol{\Theta}} + \mathbf{B}_{\omega,0} \boldsymbol{\tau}_\omega - (\hat{\boldsymbol{\Theta}}_d + k_\omega \hat{\dot{\boldsymbol{\Theta}}}_d) + d_\omega^* \end{bmatrix} \\ &= \begin{bmatrix} \dot{\mathbf{K}}\mathbf{v} + \mathbf{K}\mathbf{F}_v + k_v \dot{\mathbf{P}} + \mathbf{B}_{v,0} \boldsymbol{\tau}_v - (\hat{\mathbf{P}}_d + k_v \hat{\dot{\mathbf{P}}}_d) + d_v^* \\ \dot{\mathbf{R}}\boldsymbol{\Omega} + \mathbf{R}\mathbf{F}_\omega + k_\omega \dot{\boldsymbol{\Theta}} + \mathbf{B}_{\omega,0} \boldsymbol{\tau}_\omega - (\hat{\boldsymbol{\Theta}}_d + k_\omega \hat{\dot{\boldsymbol{\Theta}}}_d) + d_\omega^* \end{bmatrix} \\ &= \mathbf{M} + \mathbf{F} + \mathbf{B}\boldsymbol{\tau} - \mathbf{Y}_d + \mathbf{D} \end{aligned} \quad (13)$$

where  $\mathbf{M} = \begin{bmatrix} \dot{\mathbf{K}}\mathbf{v} \\ \dot{\mathbf{R}}\boldsymbol{\Omega} \end{bmatrix}$ ,  $\mathbf{F} = \begin{bmatrix} \mathbf{K}\mathbf{F}_v + k_v \dot{\mathbf{P}} \\ \mathbf{R}\mathbf{F}_\omega + k_\omega \dot{\boldsymbol{\Theta}} \end{bmatrix}$ ,  $\mathbf{B} = \begin{bmatrix} \mathbf{B}_{v,0} & \mathbf{0}_{3 \times 3} \\ \mathbf{0}_{3 \times 3} & \mathbf{B}_{\omega,0} \end{bmatrix}$ ,  $\mathbf{Y}_d = \begin{bmatrix} \hat{\mathbf{P}}_d + k_v \hat{\dot{\mathbf{P}}}_d \\ \hat{\boldsymbol{\Theta}}_d + k_\omega \hat{\dot{\boldsymbol{\Theta}}}_d \end{bmatrix}$ ,  $\mathbf{D} = \begin{bmatrix} d_v^* \\ d_\omega^* \end{bmatrix}$ .

### 3.2 Controller Design

In view of the nonlinearity and complexity of  $\mathbf{M}$ , we adopt an actor neural network to approximate the function in the following form

$$\mathbf{M} = W_a^T \sigma_a(\bar{X}) + \varepsilon_a(\bar{X}) \quad (14)$$

where  $\bar{X} = [x, y, z, \phi, \theta, \psi, x_d, y_d, z_d, \phi_d, \theta_d, \psi_d]^T$ , and  $\varepsilon_a(\bar{X})$  is the approximation error, which is bounded by a positive constant.  $W_a \in R^{12 \times 6}$  and  $\sigma_a \in R^{12 \times 1}$  denote the optimal actor neural network weight matrix and the actor neural network gaussian basis function vector.

In order to enhance the system performance and evaluate the actor neural network action, a reinforcement critic function with critic neural network is employed [11]. The critic function consists of two parts which is shown as follows

$$S_c = S + \|S\|W_c^T \sigma_c(\bar{X}) \quad (15)$$

The first part  $S = [S_x, S_y, S_z, S_\phi, S_\theta, S_\psi]^T$  plays a major role in evaluating the system performance, and its elements are designed by the following function

$$S_i = \frac{s_i}{1 + e^{-q_i e_{fi}}} - \frac{s_i}{1 + e^{q_i e_{fi}}} (i = x, y, z, \phi, \theta, \psi), \quad (16)$$

where  $q_i > 0, s_i > 0$ . From the expression, we can tell that all the value of  $S_i$  are bounded by  $s_i$ . The secondary important critic signal  $\|S\|W_c^T \sigma_c(\bar{X})$  is the second part that plays an auxiliary role in supervising the system performance, where  $W_c \in R^{12 \times 6}$  and  $\sigma_a \in R^{12 \times 1}$  represent the optimal critic neural network weight matrix and the ideal critic neural network Gaussian basis function vector, respectively.

We define  $\hat{S}_c$  as the approximation of the optimal critic neural network weight matrix  $W_c$  by designing an adaptive law to update the actual critic neural network weight matrix  $\hat{S}_c$ . Therefore, the actual critic signal is updated in the following way

$$\hat{S}_c = S + \|S\|\hat{W}_c^T \sigma_c(\bar{X}) \quad (17)$$

*Remark 2.* In combination with Eq. (16), we can easily tell that  $S_c^T S_c$  changes in the same direction as the tracking errors. Thus, when the airship tracks the predetermined trajectory, and at the same time the tracking error approaches to zero, the critic signal  $S_c^T S_c$  should decrease to zero as close as possible.

*Remark 3.* As a learning signal,  $S_c$  is more informative than the states for the controller, which can be used to obtain better control performance.

Subsequently, the adaptive update laws are proposed as follows, which are used to approximate the optimal weight matrix.

$$\dot{\hat{W}}_a = \alpha_1 \left( -L_1 \hat{W}_a + \sigma_a(\bar{X}) \hat{S}_c^T \right) \quad (18)$$

$$\dot{\hat{W}}_c = \alpha_2 \left( -L_2 \hat{W}_c + \|S\| \sigma_c(\bar{x}) \left( \hat{W}_a^T \sigma_a(\bar{X}) \right)^T \right) \quad (19)$$

where  $L_1, L_2, \alpha_1, \alpha_2$  are some positive constants. Then, it is necessary to bring up the following inequality for the sake of the design of control input

$$\begin{aligned} & \|S\| \left( \text{tr} \left( \hat{W}_a^T \sigma_a \sigma_c^T W_c \right) - \text{tr} \left( W_a^T \sigma_a \sigma_c^T \hat{W}_c \right) \right) + S^T (D + \varepsilon_a) \\ & \leq \|S\| \left( \left\| \hat{W}_a^T \right\| \bar{\sigma}_a \bar{\sigma}_c \|W_c^T\| + \|W_a^T\| \bar{\sigma}_a \bar{\sigma}_c \left\| \hat{W}_c^T \right\| + \bar{D} + \bar{\varepsilon}_a \right) \\ & \leq \|S\| \varphi \eta \end{aligned} \quad (20)$$

where

$$\begin{aligned} \varphi &= 1 + \left\| \hat{W}_a^T \right\| + \left\| \hat{W}_c^T \right\|, \eta = \max(\eta_1, \eta_2, \eta_3) \\ \eta_1 &= \bar{D} + \bar{\varepsilon}_a, \eta_2 = \bar{\sigma}_a \bar{\sigma}_c \|W_c^T\|, \eta_3 = \|W_a^T\| \bar{\sigma}_a \bar{\sigma}_c \end{aligned} \quad (21)$$

Since the parameter  $\eta$  is unknown, we have to define  $\hat{\eta}$  as the approximation of  $\eta$  by proposing an adaptive law to obtain  $\hat{\eta}$ . The adaptive update law is as follows.

$$\dot{\hat{\eta}} = \alpha_3 \frac{\|\varphi S\|^2}{\|\varphi S\| + \varepsilon_s} - \alpha_3 L_3 \hat{\eta} \quad (22)$$

where  $\alpha_3, L_3 > 0$ ,  $\varepsilon_s, \hat{\eta}(0)$  are positive constants.

For the sake of diminishing and compensating the effect of disturbance and the neural network fitting error, a special control term  $\mu_d$  combined with  $\hat{\eta}$  is designed as follows.

$$\mu_d = -\frac{\varphi^2 S}{\|\varphi S\| + \varepsilon_s} \hat{\eta} \quad (23)$$

In the end, the designed control input is presented in the following way.

$$\tau = B^{-1} \left( -\hat{W}_a^T \sigma_a (\bar{x}) - K e_f - F + Y_d + \mu_d \right) \quad (24)$$

**Theorem 1.** Consider the helium-filled airship system described by Eq. (3) under Lemmas 1–2, Assumptions 1–2, choosing the appropriate parameters  $k_x, L_1, L_2, L_3, \alpha_1, \alpha_2, \alpha_3, \varepsilon_s$ , under the action of designed adaptive laws and controllers (18), (19), (22), (23), (24). As long as the initial system states belong to one selected compact set, all the signals of closed-loop system are bounded.

*Proof.* Choosing the following Lyapunov function candidate as

$$\begin{aligned} V &= \sum_{i=x,y,z,\phi,\theta,\varphi} \frac{s_i}{q_i} \left( \ln(1 + e^{q_i e_{f_i}}) + \ln(1 + e^{-q_i e_{f_i}}) \right) + \frac{1}{2\alpha_1} \text{tr} \left( \tilde{W}_a^T \tilde{W}_a \right) \\ &\quad + \frac{1}{2\alpha_2} \text{tr} \left( \tilde{W}_c^T \tilde{W}_c \right) + \frac{1}{2\alpha_3} \tilde{\eta}^2 \end{aligned} \quad (25)$$

where  $\tilde{W}_i = W_i - \hat{W}_i$  ( $i = a, c$ ) and  $\tilde{\eta} = \eta - \hat{\eta}$ .

Then, we have the derivative of  $V$

$$\begin{aligned}
\dot{V} = & S_x \dot{e}_{fx} + S_y \dot{e}_{fy} + S_z \dot{e}_{fz} + S_\phi \dot{e}_{f\phi} + S_\theta \dot{e}_{f\theta} + S_\psi \dot{e}_{f\psi} \\
& + \frac{1}{\alpha_1} \text{tr} \left( \tilde{W}_a^T \dot{\tilde{W}}_a \right) + \frac{1}{\alpha_2} \text{tr} \left( \tilde{W}_c^T \dot{\tilde{W}}_c \right) + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}} \\
= & S^T \dot{e}_f + \frac{1}{\alpha_1} \text{tr} \left( \tilde{W}_a^T \dot{\tilde{W}}_a \right) + \frac{1}{\alpha_2} \text{tr} \left( \tilde{W}_c^T \dot{\tilde{W}}_c \right) + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}} \\
= & S^T \left( -\hat{W}_a^T \sigma_a - K e_f + \mu_d + W_a^T \sigma_a + D + \varepsilon_a \right) \\
& + \frac{1}{\alpha_1} \text{tr} \left( \tilde{W}_a^T \dot{\tilde{W}}_a \right) + \frac{1}{\alpha_2} \text{tr} \left( \tilde{W}_c^T \dot{\tilde{W}}_c \right) + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}} \\
= & -S^T \tilde{W}_a^T \sigma_a - S^T K e_f + S^T (\mu_d + D + \varepsilon_a) \\
& + \frac{1}{\alpha_1} \text{tr} \left( \tilde{W}_a^T \dot{\tilde{W}}_a \right) + \frac{1}{\alpha_2} \text{tr} \left( \tilde{W}_c^T \dot{\tilde{W}}_c \right) + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}} \\
= & -S^T \tilde{W}_a^T \sigma_a - S^T K e_f + S^T (\mu_d + D + \varepsilon_a) + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}} \\
& + \text{tr} \left( \tilde{W}_a^T \left( \sigma_a \hat{S}_c^T - L_1 \hat{W}_a \right) \right) - \text{tr} \left( \tilde{W}_c^T \left( L_2 \hat{W}_c^T + \|S\| \sigma_c \left( \hat{W}_a^T \sigma_a \right)^T \right) \right) \\
= & -S^T \tilde{W}_a^T \sigma_a - S^T K e_f - \text{tr} \left( \tilde{W}_c^T \left( L_2 \hat{W}_c + \|S\| \sigma_c \left( \hat{W}_a^T \sigma_a \right)^T \right) \right) \\
& + S^T (\mu_d + D + \varepsilon_a) + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}} + \text{tr} \left( \tilde{W}_a^T \left( \sigma_a \left( S + \|S\| \hat{W}_c^T \sigma_c \right)^T - L_1 \hat{W}_a \right) \right)
\end{aligned} \tag{26}$$

Since  $\text{tr} \left( \tilde{W}_a^T \sigma_a S^T \right) = S^T \tilde{W}_a^T \sigma_a$ , the equation can simplified by eliminating these terms.

$$\begin{aligned}
\dot{V} = & \|S\| \left( \text{tr} \left( \hat{W}_a^T \sigma_a \sigma_c^T W_c \right) - \text{tr} \left( W_a^T \sigma_a \sigma_c^T \hat{W}_c \right) \right) \\
& - \text{tr} \left( \tilde{W}_a^T L_1 \hat{W}_a \right) - \text{tr} \left( \tilde{W}_c^T L_2 \hat{W}_c \right) - S^T K e_f + S^T (\mu_d + D + \varepsilon_a) + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}}
\end{aligned} \tag{27}$$

Substituting  $\tilde{W}_a = \hat{W}_a - W_a$ ,  $\tilde{W}_c = \hat{W}_c - W_c$  into Eq. (27), we can reach the following deduction

$$\begin{aligned}
\dot{V} = & \|S\| \left( \text{tr} \left( \left( \hat{W}_a^T - W_a \right) \sigma_a \sigma_c^T \hat{W}_c \right) - \text{tr} \left( \tilde{W}_c^T \sigma_c \sigma_a^T \hat{W}_a \right) \right) \\
& - \text{tr} \left( \tilde{W}_a^T L_1 \hat{W}_a \right) - \text{tr} \left( \tilde{W}_c^T L_2 \hat{W}_c \right) - S^T K e_f + S^T (\mu_d + D + \varepsilon_a) + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}} \\
= & \|S\| \left( \text{tr} \left( \hat{W}_a^T \sigma_a \sigma_c^T W_c \right) - \text{tr} \left( W_a^T \sigma_a \sigma_c^T \hat{W}_c \right) \right) \\
& - \text{tr} \left( \tilde{W}_a^T L_1 \hat{W}_a \right) - \text{tr} \left( \tilde{W}_c^T L_2 \hat{W}_c \right) - S^T K e_f + S^T (\mu_d + D + \varepsilon_a) + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}}
\end{aligned} \tag{28}$$

Using inequality (20)

$$\dot{V} \leq -S^T K e_f - \text{tr} \left( \tilde{W}_a^T L_1 \hat{W}_a \right) - \text{tr} \left( \tilde{W}_c^T L_2 \hat{W}_c \right) + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}} + \|S\| \varphi \eta + S^T \mu_d \quad (29)$$

Equation (16) can be rewritten as  $S_i = \frac{s_i(e^{q_i e_{fi}} - e^{-q_i e_{fi}})}{2 + e^{-q_i e_{fi}} + e^{q_i e_{fi}}} (i = x, y, z, \phi, \theta, \psi)$ . Therefore, it can be deduced that  $S_i < s_i \tanh(q_i e_{fi})$ , then we make the following definition.

$$q_x = \min(q_i), S_x = \min(s_i), k_x = \lambda_{\min}(K), \quad (30)$$

Therefore, we have

$$\begin{aligned} \dot{V} \leq & -k_x S_x \tanh^T(e_f/q_x) e_f - \text{tr} \left( \tilde{W}_a^T L_1 \hat{W}_a \right) - \left( \tilde{W}_c^T L_2 \hat{W}_c \right) \\ & + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}} + \|S\| \varphi \eta + S^T \mu_d \end{aligned} \quad (31)$$

Based on Lemma 1

$$-\tanh^T(e_f/q_x) e_f \leq -\|e_f\| + m k_\tau q_x \quad (32)$$

where  $k_\tau$  is a constant that satisfies  $k_\tau = e^{-(k_\tau+1)}$ , i.e.,  $k_\tau = 0.2785$

Therefore, we can obtain

$$\begin{aligned} \dot{V} \leq & k_x S_x (-\|e_f\| + m k_\tau q_x) - \text{tr} \left( \tilde{W}_a^T L_1 \hat{W}_a \right) \\ & - \text{tr} \left( \tilde{W}_c^T L_2 \hat{W}_c \right) + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}} + \|S\| \varphi \eta + S^T \mu_d \end{aligned} \quad (33)$$

Combining with the two inequalities  $-\text{tr}(\tilde{W}_a^T \hat{W}_a) \leq \frac{1}{2} \|W_a\|^2 - \frac{1}{2} \|\tilde{W}_a\|^2$

and  $-\text{tr}(\tilde{W}_c^T \hat{W}_c) \leq \frac{1}{2} \|W_c\|^2 - \frac{1}{2} \|\tilde{W}_c\|^2$ , then we can deduce as follows

$$\begin{aligned} \dot{V} \leq & k_x S_x (-\|e_f\| + m k_\tau q_x) + \frac{1}{2} L_1 \|W_a\|^2 - \frac{1}{2} L_1 \|\tilde{W}_a\|^2 + \frac{1}{2} L_2 \|W_c\|^2 \\ & - \frac{1}{2} L_1 \|\tilde{W}_c\|^2 + \frac{1}{\alpha_3} \tilde{\eta} \dot{\tilde{\eta}} + \|S\| \varphi \eta + S^T \mu_d \\ \leq & k_x S_x (-\|e_f\| + m k_\tau q_x) + \frac{1}{2} L_1 \|W_a\|^2 - \frac{1}{2} L_1 \|\tilde{W}_a\|^2 + \frac{1}{2} L_2 \|W_c\|^2 \\ & - \frac{1}{2} L_1 \|\tilde{W}_c\|^2 + \tilde{\eta} \frac{\|\varphi S\|^2}{\|\varphi S\| + \varepsilon_s} - \tilde{\eta} L_3 \hat{\eta} - S^T \frac{\varphi^2 S}{\|\varphi S\| + \varepsilon_s} \hat{\eta} + \|S\| \varphi \eta \\ \leq & k_x S_x (-\|e_f\| + m k_\tau q_x) + \frac{1}{2} L_1 \|W_a\|^2 - \frac{1}{2} L_1 \|\tilde{W}_a\|^2 + \frac{1}{2} L_2 \|W_c\|^2 \\ & - \frac{1}{2} L_1 \|\tilde{W}_c\|^2 + \frac{\|\varphi S\|^2 (\tilde{\eta} - \hat{\eta})}{\|\varphi S\| + \varepsilon_s} - \tilde{\eta} L_3 \hat{\eta} + \|S\| \varphi \eta \\ \leq & k_x S_x (-\|e_f\| + m k_\tau q_x) + \frac{1}{2} L_1 \|W_a\|^2 - \frac{1}{2} L_1 \|\tilde{W}_a\|^2 + \frac{1}{2} L_2 \|W_c\|^2 \\ & - \frac{1}{2} L_1 \|\tilde{W}_c\|^2 - \tilde{\eta} L_3 \hat{\eta} + \frac{\|\varphi S\| \eta \varepsilon_s}{\|\varphi S\| + \varepsilon_s} \end{aligned} \quad (34)$$

Combining with  $-\tilde{\eta}\hat{\eta} \leq \frac{1}{2}\eta^2 - \frac{1}{2}\tilde{\eta}^2$ ,  $0 < \|\varphi S\| / (\|\varphi S\| + \varepsilon_s) < 1$ , we have:

$$\dot{V} \leq -k_x S_x \|e_f\| - \frac{1}{2} L_1 \|\tilde{W}_a\|^2 - \frac{1}{2} L_1 \|\tilde{W}_c\|^2 - \frac{1}{2} L_3 \tilde{\eta}^2 + \Phi \quad (35)$$

where  $\Phi = \frac{1}{2} L_1 \|W_a\|^2 + \frac{1}{2} L_2 \|W_c\|^2 + \frac{1}{2} L_3 \eta^2 + k_x S_x m k_\tau q_x + \eta \varepsilon_s$ .

So if the condition is satisfied that  $\|e_f\| > \Phi/k_x S_x$  or  $\|\tilde{W}_a\| > \sqrt{(2\Phi/L_1)}$  or  $\|\tilde{W}_c\| > \sqrt{(2\Phi/L_2)}$  or  $\tilde{\eta} > \sqrt{(2\Phi/L_3)}$ ,  $\dot{V} < 0$  can be obtained. Therefore, the earlier analysis demonstrates that the tracking errors  $e_f$ , the weight estimation errors of critic NN and actor NN, and  $\hat{\eta}$  are UUB. Thus, all signals of this closed-loop system are proved to be UUB in terms of Lyapunov's second method.

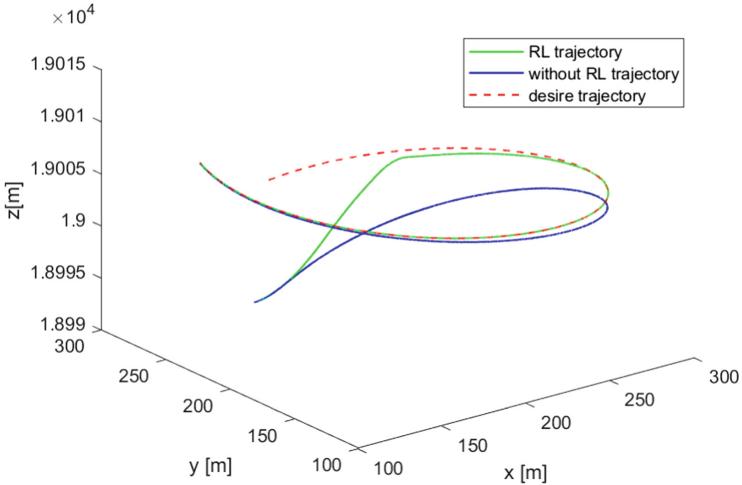
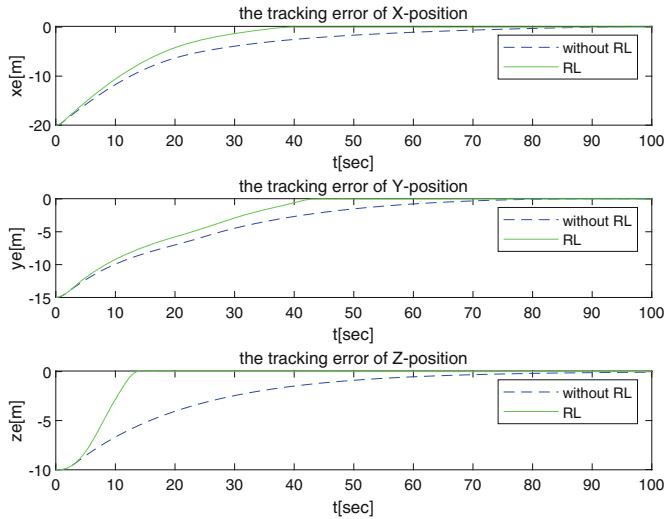
*Remark 4.* Combined with some lemmas and inequalities, a special Lyapunov function is introduced and proved to be positive and decreasing, which can guarantee the stability of the airship system.

## 4 Simulations

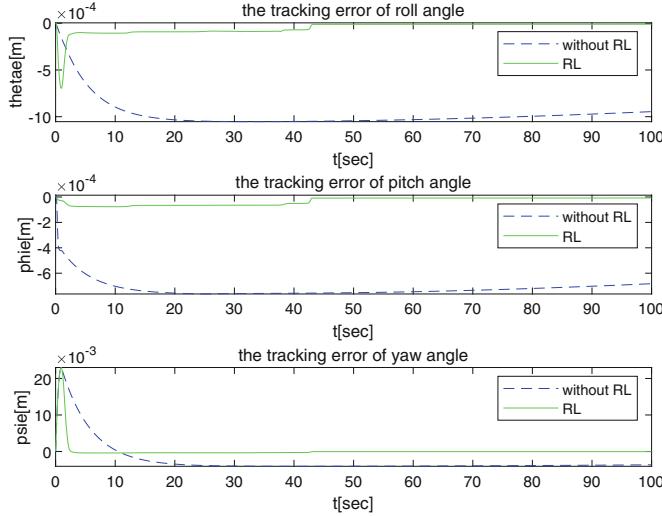
In this section, simulations for trajectory tracking are presented with the simulation step size as 0.01 s to verify the proposed control algorithm. The helium-filled airship investigated throughout the study is the same as the one in [7] and the physical parameters can be referred to the article. The initial states are  $\mathbf{P} = [180, 285, 18990]^T \text{m}$ ,  $\boldsymbol{\Theta} = [0, 0, 0]^T \text{rad}$ ,  $\mathbf{v} = [5, 0, 0]^T \text{m/s}$  and  $\boldsymbol{\Omega} = [0, 0, 0]^T \text{rad/s}$ . The desired trajectory is determined by  $\mathbf{P}_d = [x_d, y_d, z_d]^T = [200 + 100 \times \sin(0.05t), 200 + 100 \times \cos(0.05t), 0.1t + 19000]^T \text{m}$ . The desired attitude  $\boldsymbol{\Theta}_d = [\phi_d, \theta_d, \psi_d]^T = [0, \arctan 2(\dot{z}_d, \sqrt{\dot{x}_d^2 + \dot{y}_d^2}), \arctan 2(\dot{y}_d, \dot{x}_d)]^T \text{rad}$ . The parameters for the RL trajectory tracking controller is presented in the Table 1.

**Table 1.** Parameters for the RL trajectory tracking controller

$\Lambda$	1.5	$\omega_n$	25
$\eta(0)$	1	$\alpha_1$	1
$\alpha_2$	1	$\alpha_3$	1
$L_1$	0.5	$L_2$	0.5
$L_3$	1	$q_i$	2
$\mathbf{k}_\omega$	diag {15, 20, 15}	$s_i$	1
$\mathbf{k}_v$	diag {15, 15, 2}	$\varepsilon_s$	0.5
$\mathbf{K}$	$10^{-2} \times \text{diag } \{5, 5, 5, 20, 20, 20\}$		

**Fig. 2.** The tracking trajectory**Fig. 3.** The tracking position errors

All simulation results are shown in Figs. 2, 3, and 4. The stratospheric airship trajectory is shown in Fig. reffig2, where the red curve shows the desired trajectory, the blue curve shows the trajectory under the influence of the controller without RL algorithm and the green curve shows the trajectory under the RL adaptive controller, respectively. It is obvious that the trajectory tracking controllers can accurately track the reference trajectory with initial position and attitude errors. In addition, under the action of reinforcement learning adaptive controller, the trajectory of the helium-filled stratospheric airship is more in line with the desired trajectory.



**Fig. 4.** The tracking attitude errors

The tracking errors are shown in Fig. 3 and 4, where the blue curve shows the tracking errors under the effect of the controller without RL algorithm and the green curve shows the tracking errors with the RL controller. According to the comparison results of tracking errors, we can draw a conclusion that, compared to the controller without RL method, the RL adaptive controller can track the reference trajectory faster and better.

## 5 Conclusions

A RL adaptive tracking control combined with neural networks has been designed for a stratospheric airship with external disturbance, which can enhance the tracking performance and accurately track the desired trajectory. All signals of the system are bounded and the stability of the system can be guaranteed in terms of Lyapunov's second method.

**Acknowledgments.** This work was supported by Beijing Natural Science Foundation (No.4202038) and Fundamental Research Funds for the Central Universities (No. YWF-20-BJ-J-419).

## References

1. Zheng, Z., Huo, W., Wu, Z.: Trajectory tracking control for underactuated stratospheric airship. *Adv. Space Res.* **50**(7), 906–917 (2012)
2. Sun, W., Tang, S., Gao, H., Zhao, J.: Two time-scale tracking control of nonholonomic wheeled mobile robots. *IEEE Trans. Control Syst. Technol.* **24**(6), 2059–2069 (2016)

3. Li, R., Chen, M., Wu, Q.: Adaptive neural tracking control for uncertain nonlinear systems with input and output constraints using disturbance observer. *Neurocomputing* **235**(Apr 26), 27–37 (2017)
4. Aguiar, A.P., Hespanha, J.P.: Logic-based switching control for trajectory-tracking and path-following of underactuated autonomous vehicles with parametric modeling uncertainty. *Proc. Am. Control Conf.* **4**, 3004–3010 (2004)
5. Liu, L., Wang, D., Peng, Z.: ESO-based line-of-sight guidance law for path following of underactuated marine surface vehicles with exact sideslip compensation. *IEEE J. Ocean. Eng.* **42**(2), 477–487 (2017)
6. Wen, G., Chen, P., Ge, S.S., Yang, H., Liu, X.: Optimized adaptive nonlinear tracking control using actor-critic reinforcement learning strategy. *IEEE Trans. Ind. Inf.* **15**, 4969–4977 (2019)
7. Chen, T., Zhu, M., Zheng, Z.: Asymmetric error-constrained path-following control of a stratospheric airship with disturbances and actuator saturation. *Mech. Syst. Sig. Process.* **119**(Mar 15), 501–522 (2019)
8. Lee, J.Y., Park, J.B., Choi, Y.H.: Integral reinforcement learning for continuous-time input-affine nonlinear systems with simultaneous invariant explorations. *IEEE Trans. Neural Netw. Learn. Syst.* **26**(5), 916–932 (2015)
9. Guo, X., Yan, W., Cui, R.: Integral reinforcement learning-based adaptive NN control for continuous-time nonlinear MIMO systems with unknown control directions. *IEEE Trans. Syst. Man Cybernet. Syst.* **50**(7), 1–10 (2019)
10. Sun, L., Zheng, Z.: Nonlinear adaptive trajectory tracking control for a stratospheric airship with parametric uncertainty. *Nonlinear Dyn.* **82**(3), 1419–1430 (2015)
11. Wang, D., Mu, C.: Adaptive-critic-based robust trajectory tracking of uncertain dynamics and its application to a spring-mass-damper system. *IEEE Trans. Ind. Electron.* **65**, 654–663 (2017)
12. Pane, Y.P., Nageshrao, S.P., Babuška, R.: Actor-critic reinforcement learning for tracking control in robotics. In: Decision & Control. IEEE (2016)
13. Tang, L., Liu, Y.J., Tong, S.: Adaptive neural control using reinforcement learning for a class of robot manipulator. *Neural Comput. Appl.* **25**(1), 135–141 (2013)



# Model Reference Adaptive Control with Output Constraints

Yu Hua, Tianping Zhang<sup>(✉)</sup>, Manfei Lin, and Weiwei Deng

College of Information Engineering, Yangzhou University, Yangzhou 225127, China  
tpzhang@yzu.edu.cn

**Abstract.** This article focuses on the issue of model reference adaptive control (MRAC) for first-order linear time-invariant systems (LTIS) with output constraints and unknown gain sign. The completely unknown control coefficient is handled based on Nussbaum function. The output constraints are effectively dealt with by introducing barrier Lyapunov function (BLF). With the help of the Lyapunov synthesis approach, the updating laws of unknown parameters are determined. With the aid of Babalat's lemma and the property of BLF, the tracking error is proved to converge asymptotically to zero, and output restrictions are not triggered. Simulation findings are used to verify the effectiveness of the proposed MRAC algorithm.

**Keywords:** Unknown gain sign · Output restrictions · Nussbaum type function · Barrier Lyapunov function · Model reference adaptive control

## 1 Introduction

As we know, MRAC and self-tuning control (STC) are two traditional adaptive design approaches. But, the proposed schemes usually do not consider gain sign. In addition, output constraints have not been studied in the control objective. When the gain sign was unknown, Nussbaum function was introduced in [1–4]. The issue of K-filters based adaptive control was discussed for output-feedback systems with unknown control coefficients in [5]. In [6], adaptive neural control was presented for uncertain pure-feedback plants subject to full state restrictions. Backstepping based adaptive control was discussed in [7]. But, completely unknown control coefficient and output constraints had not been addressed on MRAC strategies in the present textbooks [7, 8]. In [9], observer-based adaptive control was developed by the aid of barrier Lyapunov function for output-feedback nonlinear systems with output restrictions. The problem of MRAC was discussed for LTIS with unknown gain sign in [10]. However, the considered plant did not contain the output constraints, and the following error did not tend to zero in the simulation curve in [10].

In this article, a novel MRAC approach is investigated for output constrained LTIS under the condition that control gain is completely unknown including its sign and magnitude. The updating algorithms of two unknown constants and

Nussbaum parameter are proposed based on the Lyapunov synthesis method. The completely unknown control coefficient is disposed of based on Nussbaum function. By theoretical analysis, the tracking error is proved to converge asymptotically to zero, the boundedness of all signals in the adaptive control system is shown, in addition, output constraints are not triggered.

## 2 Problem Description and Preliminaries

Suppose our discussing system is the first-order LTIS, and its transformation function is

$$P(s) = \frac{Y(s)}{U(s)} = \frac{k_p}{s + a_p} \quad (1)$$

where  $k_p, a_p$  are unknown constants.

Consider the reference model as follows:

$$M(s) = \frac{Y_m(s)}{R(s)} = \frac{k_m}{s + a_m} \quad (2)$$

where  $k_m$  and  $a_m > 0$  are two known constants,  $Y_m(s), R(s)$  stand for the Laplace transformations of the output  $y_m$  and the input  $r$  of (2), respectively.

The control goal will design the system input  $u$  for plant (1) such that  $y$  of (1) tracks  $y_m$  of (2), and  $y \in \Omega_y = \{y : |y| < k_c\}$  with known positive constant  $k_c$ .

**Assumption 1.**  $k_p$  is completely unknown including its sign and magnitude.

**Assumption 2.**  $r(t) \in L_\infty$ , and  $|y_m| < B_0 < k_c$  with known positive constant  $B_0$ , where  $L_\infty$  denotes the set of all bounded function.

A function  $N(\xi)$  is said to be a Nussbaum function if the following two properties hold:

- (i)  $\lim_{\ell \rightarrow \infty} \sup \frac{1}{\ell} \int_0^\ell N(\xi) d\xi = +\infty$
- (ii)  $\lim_{\ell \rightarrow \infty} \inf \frac{1}{\ell} \int_0^\ell N(\xi) d\xi = -\infty$

It is easy to know that  $\xi^2 \cos(\xi)$ ,  $\xi^2 \sin(\xi)$ ,  $\exp(\xi^2) \cos(\frac{\pi}{2}\xi)$  are Nussbaum functions in [1,3,4]. In this article, select  $N(\xi) = \exp(\xi^2) \cos(\frac{\pi}{2}\xi)$ .

**Lemma 1** [8]. *If function  $g(\tau)$  is uniformly continuous for  $\tau \geq 0$ , and  $|g(\tau)| \in L_1$ , then  $g(\tau) \rightarrow 0$  as  $\tau \rightarrow \infty$ .*

**Lemma 2** [2]. *For any nonnegative differentiable function  $V(\cdot)$  and any differentiable function  $\xi(\cdot)$  over  $[0, t_f]$ , and even smooth Nussbaum function  $N(\xi(\cdot))$*

and non-zero constant  $g$ , if there exists a proper constant  $c_0$  such that the following inequality holds:

$$V(t) \leq c_0 + \int_0^t [gN(\xi(\tau)) + 1] \dot{\xi}(\tau) d\tau, \quad \forall t \in [0, t_f] \quad (3)$$

then  $V(t)$ ,  $\xi(t)$  and  $\int_0^t [gN(\xi(\tau)) + 1] \dot{\xi}(\tau) d\tau$  must be bounded on  $[0, t_f]$ .

**Lemma 3** [9]. For any positive constant  $k_b$ , let  $X = \{x \in R : |x| < k_b\} \subset R$  and  $D = R^l \times X \subset R^{l+1}$  be open sets,  $l$  is a positive integer. Consider the system:

$$\dot{Z} = \hbar(t, Z) \quad (4)$$

where  $Z = [z^T, x]^T \in D$ ,  $z \in R^l$ , and  $\hbar$  is piecewise continuous in  $t$  and locally Lipschitz in  $Z$ , uniformly in  $t$ , on  $D$ . Assume that there exist two nonnegative differentiable functions  $V_1(z), V_2(x)$  such that the following properties hold.

$$V_2(x) \rightarrow \infty, x \rightarrow k_b \text{ or } x \rightarrow -k_b \quad (5)$$

$$\ell_1(\|z\|) \leq V_1(z) \leq \ell_2(\|z\|) \quad (6)$$

where  $\ell_1(\cdot)$  and  $\ell_2(\cdot)$  are class  $K_\infty$  functions. Define  $V(Z) = V_1(z) + V_2(x)$ ,  $x(0) \in X$  and  $z(0) \in R^l$ . If the derivative of  $V$  satisfies

$$\dot{V} = \frac{\partial V}{\partial Z} \hbar \leq -\mu V + \lambda \quad (7)$$

then  $x \in (-k_b, k_b), \forall t \in [0, +\infty)$ , and  $z \in L_\infty$ , where  $\mu > 0$  and  $\lambda > 0$ .

**Lemma 4** [9]. For  $k_b > 0$  and any real variable  $x$ , if  $|x| < k_b$ , then the following inequality holds:

$$\log \frac{k_b^2}{k_b^2 - x^2} < \frac{x^2}{k_b^2 - x^2} \quad (8)$$

### 3 MRAC Design and Main Results

According to (1) and (2), we obtain

$$\dot{y} = -a_p y + k_p u \quad (9)$$

$$\dot{y}_m = -a_m y_m + k_m r \quad (10)$$

Define the following tracking error

$$e = y - y_m \quad (11)$$

Its derivative is

$$\begin{aligned} \dot{e} &= -a_p y + a_m y_m + k_p u - k_m r \\ &= -a_m e + (a_m - a_p)y + k_p u - k_m r \end{aligned} \quad (12)$$

Let  $V_1 = \frac{1}{2|k_p|} \ln \frac{k_b^2}{k_b^2 - e^2}$ , where  $k_b = k_c - B_0$ . Therefore, we have

$$\begin{aligned}\frac{dV_1}{dt} &= -\frac{a_m}{|k_p|} \frac{e^2}{(k_b^2 - e^2)} + \frac{e}{(k_b^2 - e^2)} \left[ \frac{(a_m - a_p)}{|k_p|} y_p + \frac{k_p}{|k_p|} u - \frac{k_m}{|k_p|} r \right] \\ &= -\frac{a_m}{|k_p|} \frac{e^2}{(k_b^2 - e^2)} + \frac{k_p}{|k_p|} \frac{e}{(k_b^2 - e^2)} eu \\ &\quad + \frac{e}{(k_b^2 - e^2)} \frac{(a_m - a_p)}{|k_p|} y_p - \frac{k_m}{|k_p|} \frac{e}{(k_b^2 - e^2)} r\end{aligned}\quad (13)$$

Construct the control law  $u$  and the adjusting law of Nussbaum parameter  $\xi$  as follows:

$$u = N(\xi)[-c_0(t)r - d_0(t)y] \quad (14)$$

$$\dot{\xi} = \frac{e}{k_b^2 - e^2} [-c_0(t)r - d_0(t)y] \quad (15)$$

where  $c_0(t), d_0(t)$  are the estimates of  $c_0^* = \frac{k_m}{|k_p|}$  and  $d_0^* = \frac{a_p - a_m}{|k_p|}$ , respectively.

Define  $\phi(t)$  as follows:

$$\phi(t) = \begin{bmatrix} \phi_r(t) \\ \phi_y(t) \end{bmatrix} = \begin{bmatrix} c_0^* \\ d_0^* \end{bmatrix} - \begin{bmatrix} c_0(t) \\ d_0(t) \end{bmatrix} \quad (16)$$

From (14) and (15), we have

$$eu = N(\xi)\dot{\xi} \quad (17)$$

Substituting (17) into (13), we obtain

$$\begin{aligned}\frac{dV_1}{dt} &= -\frac{a_m}{|k_p|} \frac{e^2}{(k_b^2 - e^2)} + \frac{k_p}{|k_p|} N(\xi)\dot{\xi} + \dot{\xi} - \dot{\xi} + \frac{e}{(k_b^2 - e^2)} [-c_0^*r - d_0^*y] \\ &= -\frac{a_m}{|k_p|} \frac{e^2}{(k_b^2 - e^2)} + \left[ \frac{k_p}{|k_p|} N(\xi) + 1 \right] \dot{\xi} - \phi_r \frac{er}{(k_b^2 - e^2)} - \phi_y \frac{ey}{(k_b^2 - e^2)}\end{aligned}\quad (18)$$

The adaptive laws of  $c_0(t)$  and  $d_0(t)$  are designed as follows:

$$\dot{c}_0(t) = -\eta \frac{er}{(k_b^2 - e^2)} \quad (19)$$

$$\dot{d}_0(t) = -\eta \frac{ey}{(k_b^2 - e^2)} \quad (20)$$

**Theorem 1.** For plant (1) and reference model (2), if Assumptions 1 and 2 are true, the control law is determined by (14), and the updating laws are given by (15), (19) and (20), then for any bounded initial conditions, all signals  $y, y_m, u, c_0(t), d_0(t), \xi$  and  $e$  are bounded, and  $e(t) \rightarrow 0$  as  $t \rightarrow \infty$  and output restrictions are not triggered, i.e.,  $y \in \Omega_y$ .

*Proof.* Define the nonnegative differentiable function  $V$  as follows:

$$V = V_1 + \frac{1}{2\eta} (\phi_r^2 + \phi_y^2) \quad (21)$$

where  $\eta > 0$ .

Using (18) and (21), it yields

$$\dot{V} = -\frac{a_m}{|k_p|} \frac{e^2}{(k_b^2 - e^2)} + \left( \frac{k_p}{|k_p|} N(\xi) + 1 \right) \dot{\xi} \quad (22)$$

Integrating (22) over  $[0, t]$ , we have

$$\begin{aligned} V(t) &= V(0) + \int_0^t \frac{k_p}{|k_p|} (N(\xi) + 1) \dot{\xi} d\tau - \int_0^t \frac{a_m}{|k_p|} \frac{e^2}{(k_b^2 - e^2)} d\tau \\ &\leq V(0) + \int_0^t \frac{k_p}{|k_p|} (N(\xi) + 1) \dot{\xi} d\tau \end{aligned} \quad (23)$$

From Lemma 2, it yields  $V(t) \in L_\infty$ , furthermore, we have  $\phi_y, \phi_r, \xi, \int_0^t \frac{k_p}{|k_p|} (N(\xi) + 1) \dot{\xi} d\tau$  are bounded, and  $y, \dot{\xi}, u, \dot{e}$  are also bounded. From (23), we obtain

$$\int_0^t \frac{a_m}{|k_p|} \frac{e^2}{k_b^2} d\tau \leq \int_0^t \frac{a_m}{|k_p|} \frac{e^2}{(k_b^2 - e^2)} d\tau = V(0) - V(t) + \int_0^t k_p [N(\xi) + 1] \dot{\xi} d\tau < +\infty$$

i.e.,  $e \in L_2$ . Due to  $e, \dot{e} \in L_\infty$ , it yields  $\lim_{t \rightarrow \infty} e(t) = 0$  according to Lemma 1. Based on the previous analysis, we obtain all the signals in the closed-loop system are bounded. Furthermore, from (22) and Lemma 4, we have

$$\dot{V} \leq -2a_m V + \mu \quad (24)$$

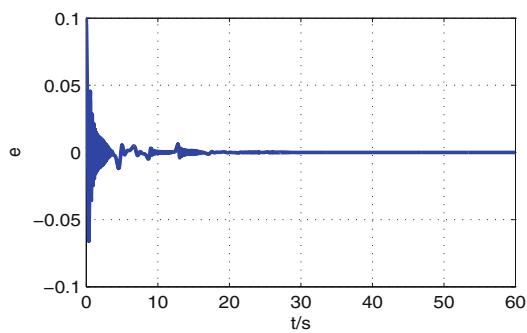
where  $\mu > 0$ . According to Lemma 3 and (23), we have  $|e| < k_b$ . It implies that  $|y| \leq |e| + |y_m| < k_b + B_0 = k_c$ .

## 4 Simulation Results

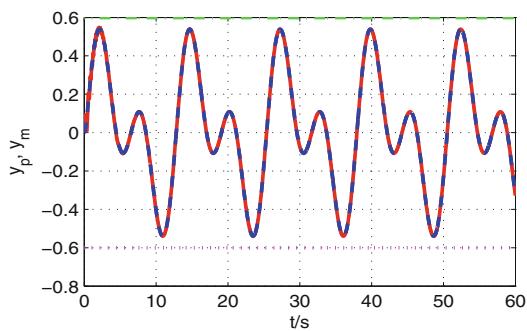
To verify the proposed theoretical findings, a numerical example is provided as follows:

$$\begin{aligned} M(s) &= \frac{2.5}{s+4} \\ P(s) &= -\frac{3.5}{s+2} \end{aligned}$$

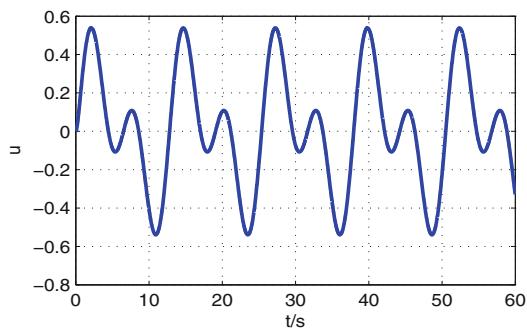
i.e.,  $\dot{y}_m = -4y_m + 2.5r, \dot{y} = -2y - 3.5u$ . Choose the reference input  $r(t) = 0.5[\sin(\frac{1}{2}t) + \sin(t)]$ . Our design goal is to construct  $u(t)$  such that  $y(t)$  asymptotically follows  $y_m(t)$ , and  $|y| < k_b$  with  $k_b = 0.2$ . Select  $\dot{\xi} = e[-c_0(t)r - d_0(t)y]$ ,  $N(\xi) = e^{\xi^2} \cos(\frac{\pi}{2}\xi)$ , the updating laws are designed as  $\dot{c}_0(t) = \eta er/(k_b^2 - e^2)$ ,  $\dot{d}_0(t) = \eta ey/(k_b^2 - e^2)$ ,  $u = N(\xi)[-c_0(t)r - d_0(t)y]$ . Select  $y(0) = 0.1, y_m(0) = 0, [c_0(0), d_0(0)] = [0.1, 0.1], \xi(0) = 0.55, \eta = 25$ . The simulation curves are plotted by Matlab in Figs. 1, 2, 3, 4 and 5. From Figs. 1, 2, 3, 4 and 5, we know that all signals  $y, y_m, e, \dot{\xi}, c_0(t), d_0(t)$  are bounded, and the tracking error asymptotically converges to zero.



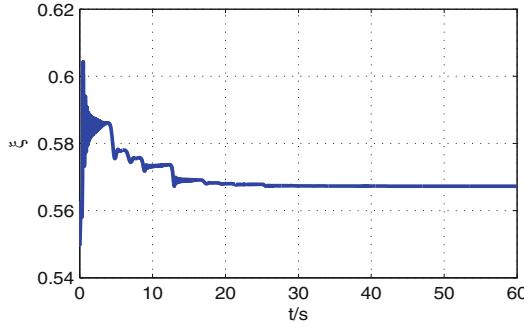
**Fig. 1.** Tracking error  $e$



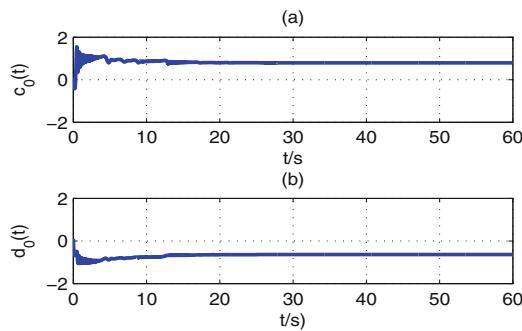
**Fig. 2.** Curve of  $y_p$  (solid line) and  $y_m$ (dashed line) for system and reference model



**Fig. 3.** Curve of input  $u$



**Fig. 4.** Nussbaum parameter  $\xi$



**Fig. 5.** (a) Tuning parameter  $c_0(t)$ ; (b) tuning parameter  $d_0(t)$

## 5 Conclusions

A new MRAC strategy is presented in this paper. Unknown high frequency gain sign is disposed of by means of Nussbaum type function. Using the Lyapunov synthesis approach, the updating laws are developed for Nussbaum parameter and two ideal constants. All the signals in the closed-loop system are proved to be bounded, and tracking error asymptotically converges to zero as time  $t$  tends to infinity with the help of Babalat's lemma. Simulation findings illustrate the effectiveness of the proposed control method.

**Acknowledgements.** This work was partially supported by the National Natural Science Foundation of China (61573307), the Natural Science Foundation of Jiangsu Province (BK20181218) and Yangzhou University Top-level Talents Support Program (2016).

## References

1. Nussbaum, R.D.: Some remarks on the conjecture in parameter adaptive control. *Syst. Control Lett.* **3**(3), 243–246 (1983)
2. Ye, X.D., Jiang, J.P.: Adaptive nonlinear design without a priori knowledge of control directions. *IEEE Trans. Autom. Control* **43**(11), 1617–1621 (1998)
3. Ge, S.S., Hong, F., Lee, T.H.: Adaptive neural control of nonlinear time-delay system with unknown virtual control coefficients. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **34**(1), 499–516 (2004). <https://doi.org/10.1109/TSMCB.2003.817055>
4. Ryan, E.P.: A universal adaptive stabilizer for a class of nonlinear systems. *Syst. Control Lett.* **16**(3), 209–218 (1991)
5. Xia, X.N., Zhang, T.P.: Adaptive output feedback dynamic surface control of nonlinear systems with unmodeled dynamics and unknown high-frequency gain sign. *Neurocomputing* **143**, 312–321 (2014). <https://doi.org/10.1016/j.neucom.2014.05.061>
6. Zhang, T.P., Xia, M.Z., Yi, Y., Shen, Q.K.: Adaptive neural dynamic surface control of pure-feedback nonlinear systems with full state constraints and dynamic uncertainties. *IEEE Trans. Syst. Man Cybern. Syst.* **47**(8), 2378–2387 (2017). <https://doi.org/10.1109/TSMC.2017.2675540>
7. Krstic, M., Kanellakopoulos, I., Kokotovic, P.V.: *Nonlinear and Adaptive Control Design*. Wiley, New York (1995)
8. Han, C.J.: *Adaptive Control*. Qinghua University Press, Beijing (1990)
9. Ren, B.B., Ge, S.S., Tee, K.P., Lee, T.H.: Adaptive neural control for output feedback nonlinear systems using a barrier Lyapunov function. *IEEE Trans. Neural Netw.* **21**(8), 1339–1345 (2010). <https://doi.org/10.1109/TNN.2010.2047115>
10. Liu, H.Q., Zhang, T.P., Wu, Z.W., Hua, Y.: Model reference adaptive control with unknown gain sign. In: *Proceedings of 2019 Chinese Intelligent Systems Conference-Lecture Notes in Electrical Engineering*, Springer, Singapore, vol. 592, pp. 508–514 (2019). [https://doi.org/10.1007/978-981-32-9682-4\\_53](https://doi.org/10.1007/978-981-32-9682-4_53)



# Low-Dose CT Image Denoising Using a Generative Adversarial Network Based on U-Net Network Structure

Yuan Fang<sup>1</sup>, Guoli Wang<sup>2</sup>, Xianhua Dai<sup>1</sup>, and Xuemei Guo<sup>2(✉)</sup>

<sup>1</sup> School of Electronics and Information Technology, Sun Yat-sen University,  
Guangzhou 510006, China

<sup>2</sup> School of Data and Computer Science, Sun Yat-sen University,  
Guangzhou 510006, China  
guoxuem@mail.sysu.edu.cn

**Abstract.** Low-dose Computed Tomography (CT) technology is widely used because it can greatly reduce the harm of scanning radiation to human body. However, due to the reduction of radiation, the projection data will be polluted and the reconstructed CT image will eventually have a lot of noise and artifacts. Therefore, how to improve the quality of CT images on the premise of reducing the radiation dose of CT scan has become a hot topic in the field of CT imaging. This article will combine the generative adversarial network (GAN) with U-net encoding and decoding structure applied in the low-dose CT (LDCT) image denoising study. Compared with other network structures, this network can extract image features more effectively and retain image details. At the same time design of stationary wavelet transform loss function, the realization of low dose CT effectively improve the quality of the images.

**Keywords:** Low dose CT denoising · Generative adversarial network · U-net · Stationary wavelet transform

## 1 Introduction

In recent years, with the continuous development of medical imaging technology, CT scanning technology has been more and more widely used. However, the high radiation generated during CT scanning can pose a potential danger to human body. Therefore, reducing the radiation dose while ensuring that the image quality meets the needs of clinical diagnosis has become an important direction in the field of medical imaging. In 1990, Naidich [12] et al. proposed the concept of low-dose CT. Which is to reduce the radiation dose by reducing the tube current while other scanning parameters are unchanged [11]. As the tube current decreases, the number of photons received by the detector decreases, resulting in the projection data being polluted by noise. And the reconstructed CT image has obvious noise and streak artifacts, which will adversely affect clinical diagnosis. A large number of experimental research results show that the

statistical distribution of quantum noise in LDCT images approximately obeys the Poisson distribution [17]. In response to these problems, many algorithms have been proposed to improve the quality of LDCT images, which can be divided into projection domain denoising algorithms, image reconstruction algorithms, and image domain denoising algorithms.

The idea of the projection domain denoising algorithm is to denoise the projected data directly. Of which the typical methods include bilateral filtering, adaptive balanced mean filtering [8], adaptive convolution filtering [18], and penalty weighted least squares [15] (PWLS). Chen [20] et al. apply an adaptive weighted non-local prior model proposed in low-dose CT image reconstruction. The advantage of this model is that it can adaptively select the global information of the image and achieve a good balance between maintaining resolution and removing noise. Wang [21] et al. first propose a new non-local filtering algorithm, in which the weight of weighted average is related not only to the reconstructed image after denoising of projected data, but also to the reconstructed image before denoising of projected data. The advantage of this type of algorithm is that it can make full use of the statistical law of the noise distribution in the projection domain. But the disadvantage is that data inconsistencies may occur during the noise reduction in the projection domain, and it is easy to generate new noise or artifacts in the reconstructed image [13].

The most representative of CT image reconstruction algorithms is the filtered back projection (FBP) algorithm [10]. It has the advantages of high resolution and fast imaging speed, and is currently the most widely used reconstruction algorithm. In recent years, some improved FBP algorithms and iterative reconstruction algorithms such as adaptive statistical iterative reconstruction (ASIR), model-based iterative reconstruction (MBIR) have appeared one after another. Xu [22] et al. are inspired by compression perception and incorporated the sparse constraint of redundant dictionary into the objective function of statistical iterative algorithm. This method not only effectively reduced the influence of noise, but also well protected the detailed features of the image. Zhang [25] et al. propose an adaptive sorting strategy for the fractal-order model based on pixels on the basis of statistical iterative reconstruction model, which can effectively improve the performance of the original model. But while it improves the quality of CT image reconstruction, it also increases the complexity of the algorithm and the time-consuming calculation.

The image domain denoising algorithm [5,9] doesn't depend on the projection data and can directly denoise the reconstructed CT image. Thus becoming the current research focus in the field of LDCT image denoising. In recent years, the deep learning [3] method has greater advantages than traditional methods when removing the complex noise of LDCT images. Xu [2] et al. propose a convolutional neural network (RED-CNN). Experimental results show that, the algorithm shows absolute advantages in objective evaluation indicators such as peak-to-noise ratio [1] (PSNR) and structural similarity [16] (SSIM) compared with traditional methods. However, although the RED-CNN algorithm has obvious denoising effect, the processed image is too smooth and the details are missing.

Yi [24] et al. train a reverse training network and a sharpness detection network to guide the training process in combination with conditional generation antagonistic network (cGAN). The results of this method showed a small resolution loss, and the validity of this model was verified through experiments. Yang [23] et al. proposed a generation countermeasures network based on Wasserstein distance and perception loss, which can not only effectively remove the noise and artifacts of CT images, but also can protect the details of images better than traditional models. In this paper, the U-net network structure is incorporated into the generative adversarial network, and the details of the image are better restored in addition to removing the noise of LDCT images.

## 2 Methods

### 2.1 Noise Reduction Model

First, assuming that  $X$  is a LDCT image,  $Y$  is a corresponding normal dose CT (NDCT) image and  $R$  is the noise in  $X$ . Then the relationship of the three can be expressed as:

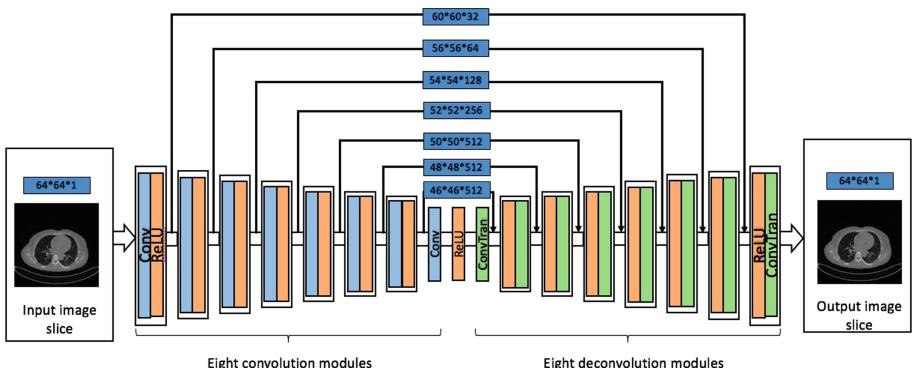
$$X = Y + R \quad (1)$$

Nest, Implementing an end-to-end mapping from  $X$  to  $Y$ . The problem can be transformed to seek a function  $f$ :

$$\underset{f}{\operatorname{argmin}} \|f(X) - Y\|_2^2 \quad (2)$$

### 2.2 Generative Adversarial Network

The generative adversarial network [4] (GAN) consists of two parts: generator network  $G$  and discriminator network  $D$ . The  $G$  network generates synthesized



**Fig. 1.** U-net generator network structure

images according to the input images, and the  $D$  network is responsible for determining whether the images given are synthetic or real. There is an adversarial relationship between the  $G$  network and the  $D$  network. Because the purpose of the  $G$  network training is to make the synthesized images to cheat the  $D$  network as much as possible, while the purpose of the  $D$  network training is to identify the fake images generated by the  $G$  network. With the progress of the training, the  $G$  network composite images can reach the degree of real images, and make it difficult for the  $D$  network to discriminate the true and false. The target formula of GAN is expressed as follows:

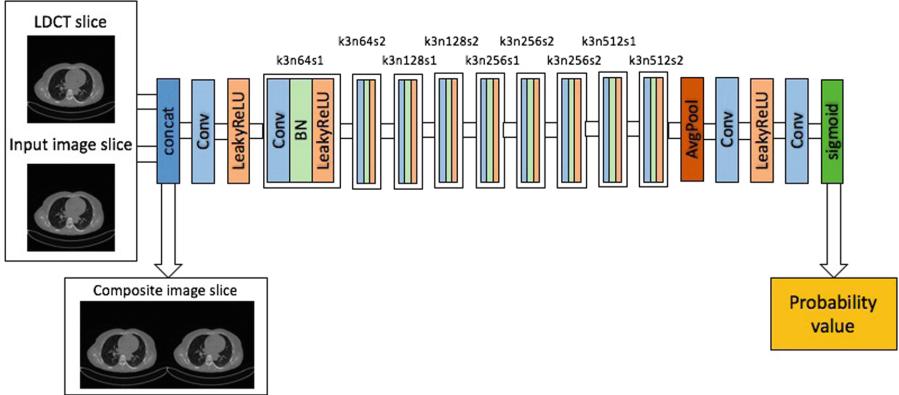
$$L_{GAN}(G, D) = E_y[\log D(y)] + E_x[\log(D(G(x)))] \quad (3)$$

where  $E(\cdot)$  denotes the expectation operator,  $x$  are LDCT images as the input,  $y$  are NDCT images.  $D(\cdot)$  is the output vector of the  $D$  network, which represents the probability that the input samples are from real samples. And  $G(\cdot)$  is the output of the  $G$  network, which is the synthesized sample vector. The goal of the  $G$  Network is to minimize the formula, and the goal of the  $D$  network is to maximize the formula. Therefore, this is actually a min-max problem.

### 2.3 Network Structures

For convenience, we name this network Unet-GAN, which consists two parts. The first part is the generator network  $G$ , which is a convolutional neural network with 8 encoder units and 8 decoder units. The structure of the encoder module is “convolution + activation function”, and the structure of the decoder module is “activation function + deconvolution”. The first decoder unit does not include the activation function. All activation functions use the relu activation function. In addition, the size of the convolution kernel in the first two encoder unit modules and the corresponding two decoder unit modules is  $5*5$ , the size of all the other convolution kernels is  $3*3$ , and the convolution step is 1. Therefore, the downsampling reduction factor in encoder units and the upsampling amplification factor in decoder units are both non-integer magnifications and are not equal. The specific model structure is shown in Fig. 1.

The second part of the network is the discriminator  $D$ , as show in Fig. 2. The discriminator network connects the original low dose image block and the discriminant image block through concat operation. The original LDCT image blocks were used as conditional input to assist the discriminant network. The basic structure of the network refers to the design of VGG [14] network structure. There are 12 convolution layers, and the structure of the 9 convolution modules from the second convolution layer is “convolution + BatchNorm + activation function”. The first and last two convolution layers don’t contain the BatchNorm [7] module. In addition, except the convolution kernel size of the last two convolution is  $1 * 1$ , the size of all convolution nuclei is  $3 * 3$ . The activation function uses the leaky-relu activation function with a slope of 0.2. In the middle 9 convolution modules, the convolution step length is kept 1 and 2 alternating with each other. Next, with the deepening of the convolutional layer, the number of



**Fig. 2.** Conditional auxiliary discrimination network structure

feature channels increases to 512 layers and remains unchanged. At the same time, an average pooling layer is added after 9 convolution modules, which can avoid model overfitting. The Sigmoid activation function is adopted in the last layer, and the final output is the probability value of judging the real image, which is between 0 and 1.

## 2.4 Loss Function

The loss function in this paper is divided into two parts, and the first part is the loss function of the generator network. In this paper, the three-stage stationary wavelet transform [19] is integrated into the loss function, and the high-frequency information of the image is further learned. The loss function of the generator network consists of three parts: the MSE pixel loss item  $l_{SWT\_MSE}$  the VGG feature loss item  $l_{SWT\_VGG}$  and the adversarial loss item  $l_{GAN}$ .

$$Loss_{SWT\_G} = \alpha l_{SWT\_MSE} + \beta l_{SWT\_VGG} + \gamma l_{GAN} \quad (4)$$

Among them, the three training parameters of  $\alpha$ ,  $\beta$  and  $\gamma$  were set as 1.0, 0.5 and 0.005 in this experiment.  $l_{SWT\_MSE}$  loss item represents the minimum mean square deviation of pixel space after two images are decomposed by stationary wavelet.  $l_{SWT\_VGG}$  loss item refers to the euclidean distance represented by the feature of Relu activation layer of the 19-layer VGG network after two images are decomposed by stationary wavelet decomposition. The  $l_{GAN}$  loss item represents the probability value of the output of the composite image in the discriminator. The specific expression is shown in formula 5, formula 6 and formula 7.

$$l_{SWT\_MSE} = \frac{1}{WH} \sum_{x=1}^W \sum_{y=1}^H (SWT\_9(I_{x,y}^{NDCT}) - SWT\_9(G(I^{LDCT})_{x,y}))^2 \quad (5)$$

$$l_{SWT.VGG} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j} SWT\_9((I^{NDCT})_{x,y}) - \phi_{i,j} SWT\_9((G(I^{LDCT}))_{x,y}))^2 \quad (6)$$

$$l_{GAN} = \sum_{n=1}^N logD(G(I^{LDCT}), I^{NDCT}) \quad (7)$$

where,  $I^{LDCT}$  is the input LDCT image, and  $I^{NDCT}$  is the true NDCT image. The size of the image is the same,  $W, H$  refers to the length and width of the image respectively.  $\phi_{i,j}$  represents the feature map obtained by convolution of layer  $j$  after activation before the maximum pooling layer of layer  $j$  in VGG19 network.  $W_{i,j}$  and  $H_{i,j}$  respectively represent the dimension size of the corresponding feature graph.  $SWT\_9(\cdot)$  function extracts 9 high frequency images generated by three level stationary wavelet transform.

### 3 Experiments

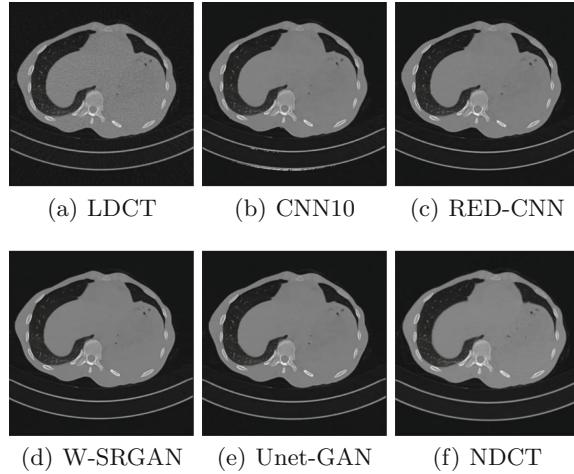
#### 3.1 Experiments Datasets

The images of the training data set used in this paper were from 7,000 two-dimensional NDCT scan images of 10 patients provided by the first people's hospital in kashgar, xinjiang. And the tube current set by the scanner is 150 mA. In this paper, 2000 NDCT images were randomly selected from 7000 NDCT images in 2d scan as the label data set in the training data set. At the same time, the label data set is converted into the corresponding LDCT images by the specified system to constitute the training data set. In addition, this article uses two test sets. The first test set are the 230 abdominal LDCT images (belly230) was provided to the first people's hospital of kashgar, xinjiang, which have corresponding NDCT images. The second are the 120 Lung LDCT images (lung120) which were randomly selected from 50 patients in ELCAP [6] (International Early Lung CAncer Program, i-elcap) Public Lung Image Database, which haven't corresponding NDCT images.

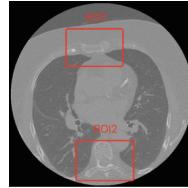
#### 3.2 Network Training

This experiment is implemented on a server with an operating system of Ubuntu 16.04. The server is configured with four Titan Xp Pascal graphics cards. In addition, the development language version is python2.7 and the neural network framework is Pytorch 0.4.0.

All network training uses the Adam optimizer. The size of the input image block is  $size = 64 * 64$ , the initial learning rate is  $l_r = 0.0001$ , the weight decay value is  $l_{decay} = 0.00005$ , and the batch size is  $batch\_size = 16$ . At last, the network iteratively trains 200 epochs.



**Fig. 3.** Experimental comparison of belly230 low dose CT image denoising



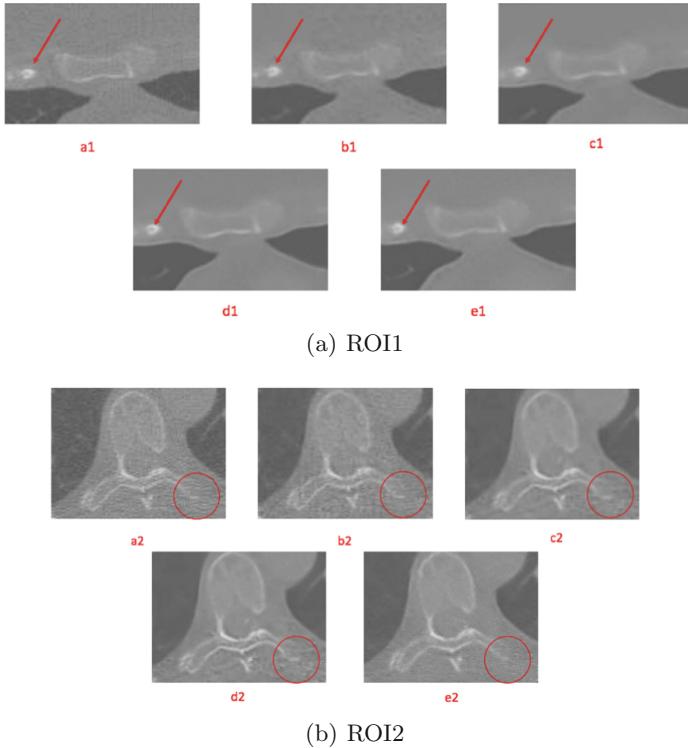
**Fig. 4.** Map of the local area of the image

### 3.3 Denoising Results

In order to verify the denoising effect of the Unet-GAN model proposed in this paper on LDCT images, three algorithms, namely CNN10, RED-CNN and W-SRGAN, are selected for experimental comparison in the field of LDCT image denoising. Figure 3 shows the denoising effect of four models for a randomly selected test image in belly230. It can be found that the Unet-GAN model and other comparison models proposed in this paper have obvious denoising effects on LDCT images, and even the quality of the denoised CT images is better than the real NDCT images in some parts, which reflects the effectiveness of the model.

In order to further compare the advantages of Unet-GAN model over other models, Fig. 4 is a test image in lung120 test set randomly selected. In the figure, two regions of local interest are marked with red rectangle boxes. Figure 5 shows the enlarged results of denoising the marked areas by four models.

The letters a, b, c, d and e represent the original LDCT image, CNN10, RED-CNN, W-SRGAN and the Unet-GAN model respectively. The numbers 1 and 2 correspond to the local areas ROI1 and ROI2 in the Fig. 4 respectively. First, observing the results in the ROI1 region. It can be found that the four



**Fig. 5.** Zooming out of the local area processed by each model

models have obvious effects on noise processing. However, the reconstruction of the white bone in the image treated with the Unet-GAN model is more clear visually in the last image. Secondly, observing the results in the ROI2 region. It can be found that the CNN10 model has the worst effect for the removal of artifacts in the image, while the Unet-GAN model has the most obvious effect. From the above two observations, we believe that the Unet-GAN model in this paper has more advantages in preserving image details compared with other comparison models.

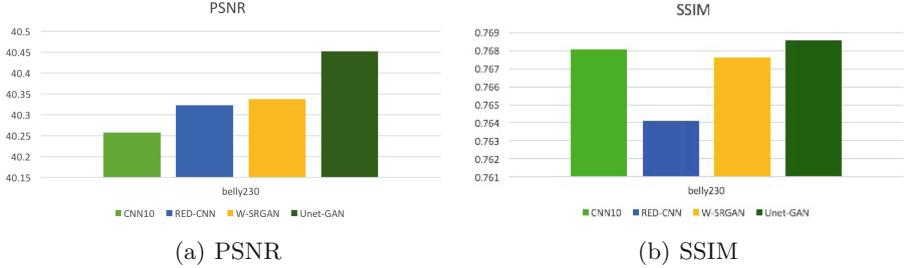
### 3.4 Quantitative Analysis

In order to objectively analyze the denoising effect of the model, two all-reference image evaluation indexes, peak signal-to-noise ratio (PSNR) and structure similarity (SSIM) were used for calculation and analysis of the belly230 test set. The specific data are shown in Table 1. In order to observe the differences of the results of each model more intuitively, the corresponding histogram is drawn, as shown in Fig. 6.

First, we can find that the Unet-GAN model has a great advantage over other models in the objective index value. In the belly230 test set, the mean value

**Table 1.** Average value of image evaluation index of test set belly230

Test set name	Evaluation index	CNN10	RED-CNN	W-SRGAN	Unet-GAN
belly230	PSNR	40.2577	40.3227	40.3373	<b>40.4524</b>
	SSIM	0.7681	0.7641	0.7676	<b>0.7686</b>

**Fig. 6.** Test set image indicator bar chart

of PSNR increased by 0.1151 dB compared with the W-SRGAN model. And compared with the CNN10 model it even increased by 0.1947 dB, the increase is relatively large. From the above data, it can be concluded that the Unet-GAN model has certain advantages over other comparison models in the denoising of LDCT images.

## 4 Conclusion

To sum up, a new generative adversarial network model (Unet-GAN) is proposed to solve the problem of LDCT image denoising in this paper. The generator network adopts a network structure similar to U-net encoding and decoding, and at the same time uses the skip design. Compared with other network structures, image feature information can be extracted more effectively. The discriminator network introduces cGAN model on the basic structure of VGG network. The LDCT images were used as conditional auxiliary input to the discriminator network to further improve the accuracy of the discriminant network. Finally, a new joint loss function based on stationary wavelet transform is proposed to preserve the details of the image.

**Acknowledgments.** This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61772574 and in part by the Key Program of the National Social Science Fund of China with Grant No. 18ZDA308. In addition, this work was supported in part by the Natural Science Foundation (NSF) of Guangdong Province (2014A030308014).

## References

1. Avcıbaş, I., Sankur, B., Sayood, K.: Statistical evaluation of image quality measures. *J. Electron. Imaging* **11**(2), 206–223 (2002)
2. Chen, H., et al.: Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans. Med. Imaging* **36**(12), 2524–2535 (2017)
3. Deng, L., Yu, D., et al.: Deep learning: methods and applications. *Found. Trends® Sig. Process.* **7**(3–4), 197–387 (2014)
4. Goodfellow, I.: NIPS 2016 tutorial: generative adversarial networks. arXiv preprint [arXiv:1701.00160](https://arxiv.org/abs/1701.00160) (2016)
5. Green, M., Marom, E.M., Kiryati, N., Konen, E., Mayer, A.: Efficient low-dose CT denoising by locally-consistent non-local means (LC-NLM). In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 423–431. Springer (2016)
6. Henschke, C.I., McCauley, D.I., Yankelevitz, D.F., Naidich, D.P., McGuinness, G., Miettinen, O.S., Libby, D., Pasmantier, M., Koizumi, J., Altorki, N., et al.: Early lung cancer action project: a summary of the findings on baseline screening. *Oncologist* **6**(2), 147–152 (2001)
7. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint [arXiv:1502.03167](https://arxiv.org/abs/1502.03167) (2015)
8. Jiang, H.: Adaptive streak artifact reduction in computed tomography resulting from excessive x-ray photon noise. *Med. Phys.* **25**(11) (1998)
9. Kang, D., Slomka, P., Nakazato, R., Woo, J., Berman, D.S., Kuo, C.C.J., Dey, D.: Image denoising of low-radiation dose coronary CT angiography by an adaptive block-matching 3D algorithm. In: Medical Imaging 2013: Image Processing, vol. 8669, p. 86692G. International Society for Optics and Photonics (2013)
10. Katsevich, A.: Theoretically exact filtered backprojection-type inversion algorithm for spiral CT. *SIAM J. Appl. Math.* **62**(6), 2012–2026 (2002)
11. Mori, I., Machida, Y., Osanai, M., Iinuma, K.: Photon starvation artifacts of x-ray CT: their true cause and a solution. *Radiol. Phys. Technol.* **6**(1), 130–141 (2013)
12. Naidich, D.P., Marshall, C.H., Gribbin, C., Arams, R.S., McCauley, D.I.: Low-dose CT of the lungs: preliminary observations. *Radiology* **175**(3), 729–731 (1990)
13. Ramani, S., Fessler, J.A.: A splitting-based iterative algorithm for accelerated statistical x-ray CT reconstruction. *IEEE Trans. Med. Imaging* **31**(3), 677–688 (2011)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
15. Wang, J., Li, T., Lu, H., Liang, Z.: Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose x-ray computed tomography. *IEEE Trans. Med. Imaging* **25**, 1272–1283 (2006)
16. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
17. Whiting, B.R.: Signal statistics in x-ray computed tomography. In: Medical Imaging 2002: Physics of Medical Imaging, vol. 4682, pp. 53–60. International Society for Optics and Photonics (2002)
18. Zeng, D., Huang, J., Bian, Z., Niu, S., Zhang, H., Feng, Q., Liang, Z., Ma, J.: A simple low-dose x-ray CT simulation from high-dose scan. *IEEE Trans. Nuclear Sci.* **62**(5), 2226–2233 (2015)
19. Nongpiur, R.C., Shpak, D.J.: Impulse-noise suppression in speech using the stationary wavelet transform. *J. Acoust. Soc. Am.* **133**(2), 866 (2013)

20. Chen, Y., Gao, D., Nie, C., Luo, L., Chen, W., Yin, X., Lin, Y.: Bayesian statistical reconstruction for low-dose x-ray computed tomography using an adaptive-weighting nonlocal prior. *Comput. Med. Imaging Graphics* **33**(7), 495–500 (2009)
21. Wang, Y., Fu, S., Li, W., Zhang, C.: An adaptive nonlocal filtering for low-dose CT in both image and projection domains. *J. Comput. Des. Eng.* **2**(2), 113–118 (2015)
22. Xu, Q., Yu, H., Mou, X., Zhang, L., Hsieh, J., Wang, G.: Low-dose x-ray CT reconstruction via dictionary learning. *IEEE Trans. Med. Imaging* **31**(9), 1682–1697 (2012)
23. Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M.K., Zhang, Y., Sun, L., Wang, G.: Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans. Med. Imaging* **37**(6), 1348–1357 (2018)
24. Yi, X., Babyn, P.: Sharpness-aware low-dose CT denoising using conditional generative adversarial network. *J. Digit. Imaging* **31**(5), 655–669 (2018)
25. Zhang, Y., Wang, Y., Zhang, W., Lin, F., Pu, Y., Zhou, J.: Statistical iterative reconstruction using adaptive fractional order regularization. *Biomed. Opt. Express* **7**(3), 1015–1029 (2016)



# Robust Monocular Visual-Inertial SLAM Using Nonlinear Optimization

Jingyun Duo<sup>1</sup>, Lei Ji<sup>1</sup>, and Long Zhao<sup>1,2(✉)</sup>

<sup>1</sup> School of Automation Science and Electrical Engineering, Beihang University,  
Beijing 100191, China  
[flylong@buaa.edu.cn](mailto:flylong@buaa.edu.cn)

<sup>2</sup> Science and Technology on Aircraft Control Laboratory, Beihang University,  
Beijing 100191, China

**Abstract.** In this paper, a robust monocular visual-inertial SLAM based on nonlinear optimization is proposed. In our method, visual feature points are assigned different information matrices according to the image pyramid layers at which the features are extracted. IMU pre-integration strategy is adopted to avoid repeated IMU integration caused by initial states change in optimization. Meanwhile, we adopted the strategies of sliding window and marginalization in order to yield higher precision of states estimation and restrict the computational complexity. Experiments are designed to compare our algorithm with MSCKF and VINS on EuRoC dataset, and the results show that our method can effectively estimate the motion and sparse map.

**Keywords:** Monocular visual · Inertial · SLAM · State estimation · Nonlinear optimization

## 1 Introduction

SLAM (Simultaneous Localization and Mapping) is a computational method to create an incremental map and at the same time determine the carrier position in the map [1]. In recent years, owing to the development of computers and robotics, SLAM becomes a flourishing research hotspot and shows important application value in fields of unmanned robot system, virtual reality and augmented reality.

SLAM focuses on both localization and mapping problems, and localization and mapping are strongly correlated. That is to say the two problems are premises of each other, and neither of them can be obtained independently. Visual sensors commonly used in the research of SLAM include monocular camera, binocular camera and depth camera. Due to small size, low cost and ease of installation, monocular visual SLAM systems have attracted wide attention in academic and industrial fields in recent years [2,3]. However, there are two defects of the monocular visual SLAM system. Firstly, monocular visual SLAM system is scale ambiguous, which severely limits its application in many scenarios. Secondly, monocular visual SLAM system is obviously restricted by the

environment conditions, especially in the weak texture environment, it is impossible to estimate the camera motion effectively. The IMU( Inertial Measurement Unit) can provides acceleration and angular velocity measurements, which is considered to be highly complementary to the camera and can effectively overcomes the above two defects.

Visual-inertial tight combination SLAM fuses the raw measurements of the IMU and the camera, and the precision is much higher than the loose combination method. Visual-inertial tight combination SLAM can generally be divided into two categories: filter based method [4,5] and nonlinear optimization based method [6,7]. MSCKF(Multi-state Constraint Kalman Filter) [4] is the most popular method of visual-inertial tight combination SLAM based on EKF(Extended Kalman Filter). The states of MSCKF contain multiple camera poses, which adopt observations of the same visual features in multiple cameras to construct multi-constraint updates. However, MSCKF only updates the states once, and there is a large error in the linearization process. The nonlinear optimization based visual-inertial tight combination SLAM considers all the “historical” sensor poses as well as all the visual features observed by the cameras. In order to ensure the efficiency, the sliding window is usually used to optimize the latest multiple states, and the remaining states are marginalized. It has been proved that the optimization based method is more accurate than filter based method under the same computational [8]. Among many optimization-based visual-inertial SLAM method, VINS( Visual-inertial Navigation system) [9] is considered as one of the most classic systems. In VINS, all of the features are assigned the same information matrices, but in fact features extracted from different image pyramid layers have different weights in nonlinear optimization, therefore the accuracy of states estimation is limited.

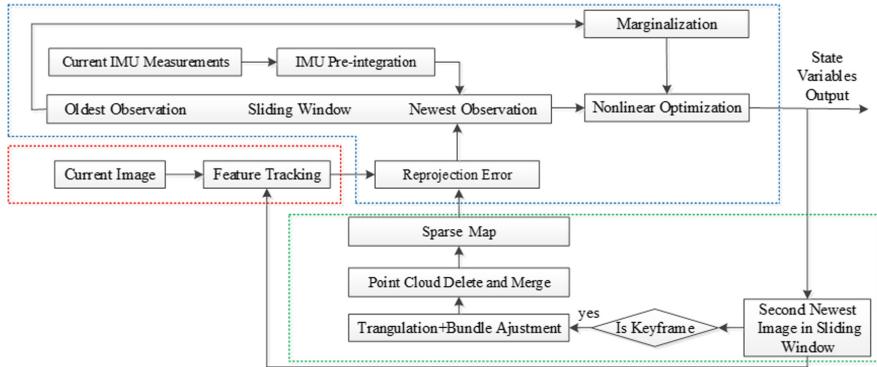
In this paper, a monocular visual-inertial SLAM based on nonlinear optimization is proposed. The cost function is constructed by reprojection errors of visual feature points and IMU pre-integration error. In our method, features are assigned different information matrices according to the image pyramid layers at which the features are extracted. This strategy greatly improves the accuracy of visual observation in the optimization. In order to improve the efficiency, IMU pre-integration strategy is adopted, which can avoid repeated IMU integration caused by initial states change in optimization. Meanwhile, we introduced the strategies of sliding window and marginalization, so as to yield higher precision of state estimation and restrict the computational complexity.

## 2 Overview

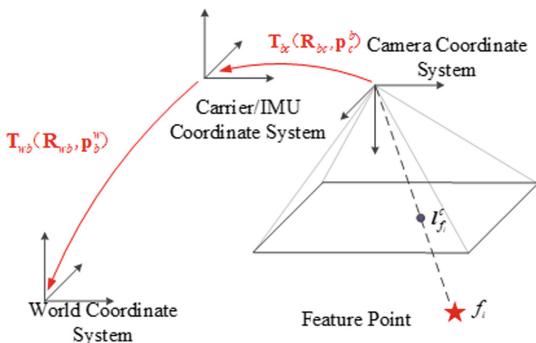
The basic framework of our proposed monocular visual-inertial SLAM is shown in Fig. 1, which is mainly divided into three parallel threads: features tracking (see the red box in Fig. 1), state estimation (see the blue box in Fig. 1) and local mapping (see the green box in Fig. 1).

As shown in Fig. 2, the proposed monocular visual-inertial SLAM sensor system includes a monocular camera and a 6-Dof IMU. We employ the following

coordinate systems and notations throughout this paper.  $(\cdot)^w$  represents projection of the state in the world coordinate system, and the  $z$  axis of the world coordinate system coincides with the direction of gravity.  $(\cdot)^b$  represents projection of the state in the carrier coordinate system, and the IMU coordinate system coincides with the carrier coordinate system.  $(\cdot)^c$  represents projection of the state in the camera coordinate system.  $\mathbf{l}_{f_i}^c$  denotes the pixel coordinate of the feature point  $f_i$  in the image coordinate system.  $\mathbf{R}_{bc}$  and  $\mathbf{p}_c^b$  denote the rotation matrix and translation from the camera coordinate system to carrier coordinate system respectively.  $\mathbf{R}_{wb}$  and  $\mathbf{p}_b^w$  denote the rotation matrix and translation from the carrier coordinate system to the world coordinate system respectively.



**Fig. 1.** Basic framework of the proposed monocular visual-inertial SLAM



**Fig. 2.** Coordinate systems of the proposed monocular visual-inertial SLAM

### 3 Method

Monocular visual-inertial SLAM problem can be simplified to a nonlinear optimization problem. The cost function is constructed by reprojection errors of visual feature points and IMU pre-integration error, and the Gauss-Newton method is used to iteratively optimize the state variables. The cost function of sliding window based monocular visual-inertial SLAM can be expressed as

$$\begin{aligned} J(\chi) = & \sum_{k=1}^n \sum_{i=1}^m \mathbf{r}(\hat{l}_{f_i}^{c_k}, \chi)^T \mathbf{W}_{\mathbf{r}(\hat{l}_{f_i}^{c_k}, \chi)} \mathbf{r}(\hat{l}_{f_i}^{c_k}, \chi) \\ & + \sum_{k=1}^{n-1} \mathbf{r}(\hat{z}_{b_{k+1}}^{b_k}, \chi)^T \mathbf{W}_{\mathbf{r}(\hat{z}_{b_{k+1}}^{b_k}, \chi)} \mathbf{r}(\hat{z}_{b_{k+1}}^{b_k}, \chi) + J_{mar}(\chi) \end{aligned} \quad (1)$$

where  $\chi$  denotes the set of state variables in the sliding window,  $\chi = [\mathbf{x}_{b_1}, \mathbf{x}_{b_2}, \dots, \mathbf{x}_{b_n}, \mathbf{p}_{f_1}^w, \mathbf{p}_{f_2}^w, \dots, \mathbf{p}_{f_m}^w]$ ,  $n$  and  $m$  denote the frame number and the number of visual feature points observed in the sliding window respectively, The state variable corresponding to the frame  $k$  in the sliding window is  $\mathbf{x}_{b_k} = [\mathbf{p}_{b_k}^w, \mathbf{v}_{b_k}^w, \mathbf{R}_{wb_k}, \mathbf{b}_{ab_k}, \mathbf{b}_{gb_k}]$ .  $\mathbf{p}_{b_k}^w, \mathbf{v}_{b_k}^w$  and  $\mathbf{R}_{wb_k}$  denote the translation, velocity and rotation matrix of the IMU system relative to the world system,  $\mathbf{b}_{ab_k}$  and  $\mathbf{b}_{gb_k}$  denote the accelerometer bias and gyro drift respectively,  $\mathbf{p}_{f_i}^w$  denotes the position of feature point  $f_i$  in world coordinates,  $\mathbf{r}(\hat{l}_{f_i}^{c_k}, \chi)$  denotes the reprojection error of feature point  $f_i$  on frame  $k$  in the sliding window,  $\mathbf{r}(\hat{z}_{b_{k+1}}^{b_k}, \chi)$  denotes the IMU pre-integration error corresponding to the frame  $k$  to frame  $k+1$  in the sliding window,  $\mathbf{W}_{\mathbf{r}(\hat{l}_{f_i}^{c_k}, \chi)}$  and  $\mathbf{W}_{\mathbf{r}(\hat{z}_{b_{k+1}}^{b_k}, \chi)}$  denote the information matrices corresponding to  $\mathbf{r}(\hat{l}_{f_i}^{c_k}, \chi)$  and  $\mathbf{r}(\hat{z}_{b_{k+1}}^{b_k}, \chi)$  respectively,  $J_{mar}(\chi)$  denotes the prior information generated by the marginalization.

#### 3.1 Reprojection Error of Feature Point

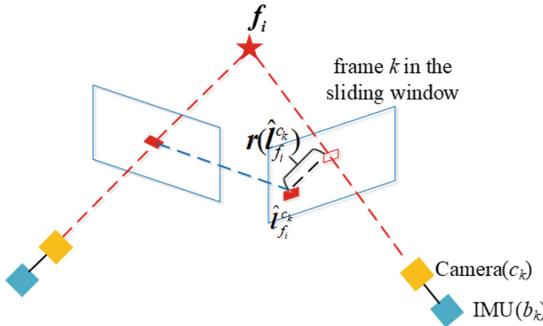
The reprojection errors of feature points based on the pinhole imaging model are adopted to construct visual observations. As shown in Fig. 3, the reprojection error of feature point  $f_i$  on frame  $k$  in the sliding window can be expressed as

$$\mathbf{r}(\hat{l}_{f_i}^{c_k}, \chi) = \pi(\mathbf{R}_{bc}^{-1}(\mathbf{R}_{wb_k}^{-1}(\mathbf{p}_{f_i}^w - \mathbf{p}_{b_k}^w) - \mathbf{p}_c^b)) - \hat{l}_{f_i}^{c_k} \quad (2)$$

where  $\hat{l}_{f_i}^{c_k}$  denotes the observation of  $f_i$  on frame  $k$  in the sliding window, which can be acquired by feature tracking.  $\mathbf{R}_{bc}$  and  $\mathbf{p}_c^b$  denote the rotation matrix and translation from the camera coordinate system to IMU coordinate system respectively, which can be acquired by camera/IMU external parameter calibration. The information matrix  $\mathbf{W}_{\mathbf{r}(\hat{l}_{f_i}^{c_k}, \chi)}$  is related to the layer of the image pyramid at which the feature  $f_i$  is extracted. Image pyramid is built by downsampling the original image, so in this paper the information matrix of the reprojection error is defined as

$$\mathbf{W}_{\mathbf{r}(\hat{l}_{f_i}^{c_k}, \chi)} = \begin{bmatrix} \left(\frac{1}{\sigma^d}\right)^2 & 0 \\ 0 & \left(\frac{1}{\sigma^d}\right)^2 \end{bmatrix} \quad (3)$$

where  $\sigma$  denotes the downsampling coefficient of the image pyramid, and we set  $\sigma = 1.2$  in this paper,  $d$  is the layer number at which the feature  $f_i$  is extracted.



**Fig. 3.** An illustration of the reprojection error

### 3.2 IMU Pre-integration Error

Monocular visual-inertial SLAM based on nonlinear optimization needs to update the state variables through multiple iterations. Commonly, IMU uses a recursive method to implement states estimation. In order to avoid repeated IMU integration caused by initial states change in optimization, the IMU pre-integration strategy [10] is adopted. IMU pre-integration error corresponding to the frame  $k$  to frame  $k + 1$  in the sliding window can be expressed as

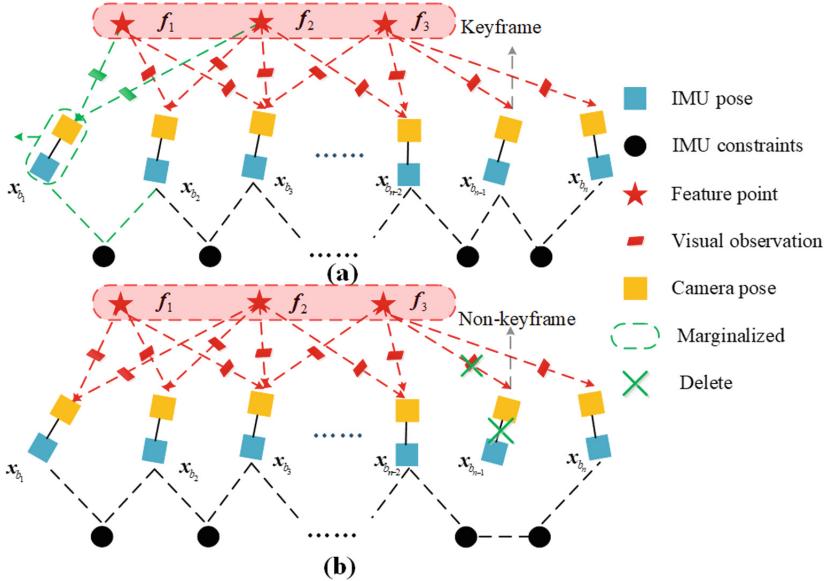
$$r(\hat{z}_{b_{k+1}}^{b_k}, \chi) = \begin{bmatrix} R_{wb_k}^{-1} (p_{b_{k+1}}^w - p_{b_k}^w - v_{b_k}^w \Delta t_{b_{k+1}b_k} + \frac{1}{2} g^w \Delta t_{b_{k+1}b_k}^2) - \Delta \hat{p}_{b_k b_{k+1}} \\ R_{wb_k}^{-1} (v_{b_{k+1}}^w - v_{b_k}^w + g^w \Delta t_{b_{k+1}b_k}) - \Delta \hat{v}_{b_k b_{k+1}} \\ 2[\Delta \hat{q}_{b_k b_{k+1}}^{-1} \otimes q_{wb_k}^{-1} \otimes q_{wb_{k+1}}]_{xyz} \\ \hat{b}_{ab_{k+1}} - b_{ab_{k+1}} \\ b_{gb_{k+1}} - b_{gb_{k+1}} \end{bmatrix} \quad (4)$$

where  $[\cdot]_{xyz}$  denotes the real part of a quaternion. IMU pre-integration defines a motion increment, which is independent of the initial states and the gravity vector and can be solved by inertial measurements only.

### 3.3 Marginalization

In monocular visual-inertial SLAM, the number of the states will steadily increase over time. If we concern all states and solve the full-state SLAM, which will not meet the real-time requirement. If only concern the latest states, the correlation between the states will be ignored, which significantly reduces the accuracy of the whole system. In order to yield higher precision of state estimation and restrict the computational complexity, we introduced the strategies of

sliding window and marginalization. As shown in Fig. 4, if the second latest frame is a keyframe, the oldest keyframe is marginalized by using the Schur complement [7], and the corresponding measurements are converted to prior of the error terms. Conversely, if the second latest frame is a non-keyframe, that is to say it contains less effective information, and most of the observed feature points can also be observed by other frames. Therefore, the measurements corresponding to this frame are directly deleted.

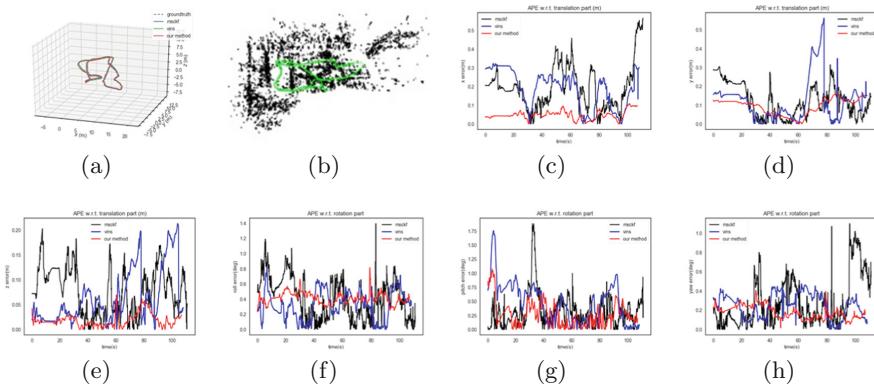


**Fig. 4.** An illustration of the marginalization strategy (a) the second latest frame is a keyframe (b) the second latest frame is a non-keyframe

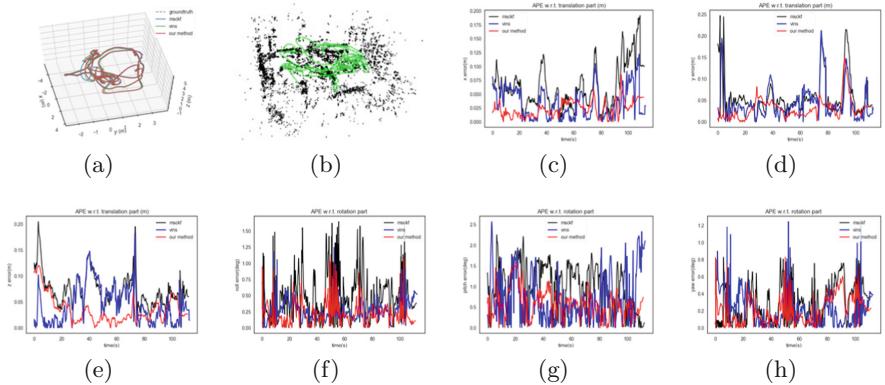
## 4 Experimental Results

In this section, experiments are designed to compare our algorithm with MSCKF [4] and VINS [9] on EuRoC (European Robotics Challenge) dataset, and the superiority of our algorithm is verified by quantitative analysis. MSCKF and VINS are currently recognized as excellent open source visual-inertial systems, among which MSCKF is implemented based on filtering, while VINS is implemented based on nonlinear optimization. The EuRoC dataset provides image data and IMU data with accurate timestamps, which are collected by a visual-inertial sensor system based on the AscTec (Unmanned Aerial Vehicle) platform. This dataset has been widely used by developers of visual SLAM algorithm to evaluate the algorithm performance. Due to limited space, we only show

the results on “MH\\_05\\_difficult” and “V2\\_02\\_medium” dataset. The trajectory curve, sparse point cloud map and APE (Absolute Pose Error) are shown in Fig. 5 and 6. The statistical results of pose absolute error are shown in Table 1 and 2. It can be seen from Fig. 5(a) and Fig. 6(a) that MSCKF, VINS and our algorithm all can effectively estimate the carrier pose without significant drift. On “MH\\_05\\_difficult” dataset, the position root mean square errors of MSCKF, VINS and our algorithm are 0.288415 m, 0.280874 m and 0.117506 m, and the attitude root mean square errors respectively are  $0.756401^\circ$ ,  $0.761376^\circ$  and  $0.570830^\circ$ . On “V2\\_02\\_medium” dataset, the position root mean square errors of the above three algorithms are 0.131211 m, 0.094537 m and 0.065146 m respectively, and the root mean square errors of attitude are  $1.334333^\circ$ ,  $1.002012^\circ$  and  $0.783835^\circ$  respectively. It can be seen that the accuracy of our algorithm is obviously better than the other two algorithms for the following two reasons. Firstly, we assign different weights to the features on different pyramid layers when using nonlinear optimization to estimate the states, while the other two algorithms set the same weight to all features. Therefore, the visual measurement accuracy and robustness of the proposed algorithm are better than the other two algorithms. Secondly, MSCKF and VINS do not effectively process the map information. In this paper, sparse 3D map is constructed, and point clouds are deleted, merged and optimized by map thread, which improves the algorithm accuracy to some extent.



**Fig. 5.** Experimental results on “MH\\_05\\_difficult” dataset: (a) trajectory, (b) trajectory and map, (c)-(e) absolute position error in x,y and z axis, (f)-(h) absolute roll, pitch and yaw error



**Fig. 6.** Experimental results on “V2\_02\_medium” dataset: (a) trajectory, (b) trajectory and map, (c)-(e) absolute position error in x,y and z axis, (f)-(h) absolute roll, pitch and yaw error

**Table 1.** Statistical results of absolute position error

Data file	Algorithm	Root mean square error (meter)
MH_05_difficult	MSCKF	0.288415
	VINS	0.280874
	Our algorithm	0.117506
V2_02_medium	MSCKF	0.131211
	VINS	0.094537
	Our algorithm	0.065146

**Table 2.** Statistical results of absolute attitude error

Data file	Algorithm	Root mean square error (degree)
MH_05_difficult	MSCKF	0.756401
	VINS	0.761376
	Our algorithm	0.570830
V2_02_medium	MSCKF	1.334333
	VINS	1.002012
	Our algorithm	0.783835

## 5 Conclusion

In this paper, we present a robust monocular visual-inertial SLAM based on nonlinear optimization. Our algorithm can effectively solve the scale ambiguous problem in monocular vision and can also provide the short-time IMU state

estimation in the weak texture environment where the visual measurements are not reliable. Experiments are designed on public dataset, and the results show that our method can effectively estimate the motion and sparse map.

**Acknowledgements.** This work is supported by the National Natural Science Foundation of China (Grant No. 41874034), the National Science and Technology Major Project of the National Key R&D Program of China (Grant No. 2016YFB0502102), the Beijing Natural Science Foundation (Grant No. 4202041), the Aeronautical Science Foundation of China.

## References

1. Davison, A.J., Reid, I.D., Molton, N.D., et al.: Monoslam: real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 1052–1067 (2007)
2. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Trans. Robot.* **31**(5), 1147–1163 (2015)
3. Engel, J., Schps, T., Cremers, D.: LSD-SLAM: large-scale direct monocular slam. In: 13th European Conference on Computer Vision, ECCV 2014, Zurich, Switzerland, 6–12 September 2014 (2014)
4. Li, M., Mourikis, A.I.: High-precision, consistent EKF-based visual-inertial odometry. *Int. J. Robot. Res.* **32**(6), 690–711 (2013)
5. Bloesch, M., Omari, S., Hutter, M., et al.: Robust visual inertial odometry using a direct EKF-based approach. In: The 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015, Hamburg, Germany, 28 September–03 October 2015 (2015)
6. Shen, S., Michael, N., Kumar, V.: Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft MAVs. In: The 2015 IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, America, 26–30 May 2015 (2015)
7. Yang, Z., Shen, S.: Monocular visual-inertial state estimation with online initialization and camera-imu extrinsic calibration. *IEEE Trans. Automation Sci. Eng.* **14**(1), 1–13 (2016)
8. Strasdat, H., Montiel, J.M.M., Davison, A.J.: Visual SLAM: why filter? *Image Vis. Comput.* **30**(2), 65–77 (2012)
9. Qin, T., Li, P., Shen, S.: VINS-Mono: a robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.* **34**(4), 1004–1020 (2018)
10. Forster, C., Carlone, L., Dellaert, F., et al.: On manifold preintegration for real-time visual-inertial odometry. *IEEE Trans. Robot.* **33**(1), 1–21 (2017)



# Research on Aerodynamically Assisted Orbit Maneuver Method Based on Feature Model Correction

Yue Lin<sup>1,2(✉)</sup>, Yingmin Jia<sup>2</sup>, and Songtao Fan<sup>1</sup>

<sup>1</sup> Beijing Institute of Control Engineering, 100094 Beijing, China  
linyue371@163.com

<sup>2</sup> BUAA 7th Research Division, 100191 Beijing, China

**Abstract.** For the hypersonic vehicle to use aerodynamic assistance to carry out orbit maneuver, because of the considerable uncertainty in the aerodynamic parameters during the re-entry process, how to carry out precise orbit maneuver with aerodynamic assistance is an important part of the research of this method. Using predictive correction guidance and control methods based on feature models, fixed-cycle track prediction and guidance correction during flight can greatly improve the robustness and adaptive ability of the system and eliminate the control deviation caused by uncertain aerodynamic parameters and allow the vehicle to accurately determine the orbit. A model with parametric bias was established for system simulation. The results show that the accuracy of the guidance control still meets the system requirements and the system has good self-adaptability through this method with aerodynamic parameter bias of 40%.

**Keywords:** Model-based · Orbit maneuver · Aerodynamically assisted · Adaptive control

## 1 Introduction

The multifunctional hypersonic vehicle is quite different from the traditional spacecraft, and can use its large lift-to-drag ratio feature to perform aerodynamically assisted orbit maneuver tasks [1].

In the process of using aerodynamics for orbital maneuvers, due to the high speed of the spacecraft (generally above Mach 20), there is great uncertainty in the aerodynamic parameters [2, 3]. If a conventional guidance control algorithm is used, it is likely to be due to a large target orbit error caused by parameter uncertainty.

Therefore, a guidance algorithm with strong adaptive control capability is needed to complete the aerodynamically assisted orbit maneuver task in a complex and uncertain environment.

In order to deal with the parameters, the end and the complex objects that are difficult to describe with accurate mathematical models for effective high-performance control, the method of feature modeling and full-coefficient adaptive control came into being. This method can effectively improve the technical approach and effect in the practical engineering field.

In this paper, the high-speed flying vehicle used the prediction and correction method based on the feature model to control its target orbit during the orbit maneuver process of hypersonic vehicles through the aerodynamically-assisted method [4]. A large-scale bias test was carried out for aerodynamic parameters, and the robustness and adaptability of the method were studied in depth.

## 2 Aerodynamically-Assisted Guidance Modeling

In order for the hypersonic vehicle to complete deorbit and reentry tasks through maneuver, it is necessary to ensure the orbital shape of the spacecraft before deorbit. To carry out advance control and correction, the ultimate goal of using aerodynamics to carry out auxiliary orbit change is to convert the above-mentioned indicators to target values with limited propellant.

The main effects of aerodynamics on the spacecraft can be divided into two parts:  $D$  and  $L$ .  $D$  is the spacecraft resistance, and the resistance directly affects the spacecraft's mechanical energy [5], so that the orbital energy of the spacecraft decreases monotonously;  $L$  is the lift force, which is perpendicular to the speed direction of the spacecraft, and therefore, it mainly changes the shape of the orbit, but does not affect the mechanical energy of the spacecraft [6]. Therefore, in the process of space and reentry joint guidance, it is necessary to estimate and control the change of the spacecraft's mechanical energy to ensure that the spacecraft has enough energy to for the predetermined orbital shape after the maneuver [7,8]. The ability of  $L$  and  $D$  to change the orbital shape of the spacecraft is mainly studies in this article, and their ability to change the orbital phase will be discussed in another article.

### 2.1 Spacecraft Dynamics Modeling

Based on the space and the reentry section, the dynamic model of the spacecraft is given comprehensively. The habit of reentry vehicle dynamics modeling is mainly adopted in this article, the aerodynamic force of the space section is deemed as zero and this section is handled in a unified approach.

$$\begin{cases} \dot{r} = v * \sin(\gamma) \\ \dot{\theta} = \frac{v * \cos(\gamma) * \sin(\psi)}{r * \cos(\varphi)} \\ \dot{\varphi} = \frac{v * \cos(\gamma) * \cos(\psi)}{r} \\ \dot{\gamma} = -D - g * \sin(\gamma) \\ \dot{\psi} = \frac{1}{v} * \left( L * \cos(\sigma) + \left( \frac{v^2}{r} - g \right) * \cos(\gamma) \right) \\ \dot{\Psi} = \frac{1}{v} \left( \frac{L * \sin(\sigma)}{\cos(\gamma)} + \frac{v^2}{r} \cos(\gamma) \sin(\psi) \tan(\varphi) \right) \end{cases} \quad (1)$$

where  $r$  is the distance from the spacecraft to the center of the earth,  $v$  is the speed of the spacecraft,  $(\theta)$  is the geocentric longitude of the spacecraft,  $\varphi$  is the geocentric latitude of the spacecraft,  $\gamma$  is the angle between the spacecraft speed and the local horizontal plane, that is, the flight path angle,  $\psi$  is the angle between the spacecraft speed and the local North Pole direction, which is the azimuth of the flight path, and  $\sigma$  is the rolling angle of the spacecraft body around the speed direction, also known as the bank angle, which is also the angle between the direction of the aerodynamic lift force generated by the spacecraft and the local vertical line.

$$g = \frac{\mu_{\text{earth}}}{r^2} \quad (2)$$

where  $g$  is the local gravity acceleration of the spacecraft and  $\mu_{\text{earth}}$  is the earth's gravitational constant.

$$\begin{cases} D = \frac{1}{2} * \frac{\rho * v^2 * C_D * S}{m} \\ L = \frac{1}{2} * \frac{\rho * v^2 * C_L * S}{m} \\ K = \frac{L}{D} = \frac{C_L}{C_D} \end{cases} \quad (3)$$

where  $D$  and  $L$  represent aerodynamic drag acceleration and aerodynamic lift acceleration (the spacecraft mass is normalized during the study), and  $K$  is the spacecraft lift-to-drag ratio.  $K$  is treated as a constant under the limited working conditions in this paper.

$$\begin{cases} D = \frac{1}{2} * \frac{\rho * v^2 * C_D * S}{m} \\ L = \frac{1}{2} * \frac{\rho * v^2 * C_L * S}{m} \\ K = \frac{L}{D} = \frac{C_L}{C_D} \end{cases} \quad (4)$$

## 2.2 Orbit-Related Parameters

In addition to dynamic modeling, it is necessary to establish an orbital root number model of the spacecraft, and formulate a joint guidance strategy based on the orbital root number model.

In the current problem, the key orbit parameters include the semi-major axis (including apogee and perigee), eccentricity, orbit inclination, and orbit mechanical energy, as shown below:

$$\begin{cases} a = \frac{r_p + r_a}{2} = \frac{\mu_{\text{earth}}}{v_p * v_a} \\ e = 1 - \frac{r_1 * v_a^2}{\mu_{\text{earth}}} \\ h = \text{cross}(r, v) = [h_x; h_y; h_z] \\ \cos(i) = \frac{h_z}{h} \\ \varepsilon = \frac{v^2}{2} - \frac{\mu_{\text{earth}}}{r} = -\frac{\mu_{\text{earth}}}{2a} = -\frac{\mu_{\text{earth}}}{r_p + r_a} \end{cases} \quad (5)$$

where  $a$  is the semi-major axis of the spacecraft,  $r_a$  and  $r_p$  are the apogee and perigee distances of the spacecraft,  $v_a$  and  $v_p$  are the apogee and perigee speeds

of the spacecraft,  $e$  is the eccentricity of the spacecraft orbit,  $h$  is the moment of momentum of the spacecraft orbit,  $i$  is the inclination angle of the spacecraft orbit,  $\Omega$  is the right ascension of the ascending node of the spacecraft, and  $\varepsilon$  is the mechanical energy of the spacecraft.

### 2.3 Parameter Association

The spacecraft has been flying in a near-circular orbit for a long time ( $r_p \approx r_a$ ). In the process of the aerodynamically assisted orbit maneuver,  $L$  has no effect on the spacecraft speed, so it does not affect the mechanical energy of the spacecraft  $\varepsilon$ , and  $D$  will monotonically decrease the spacecraft speed  $v$ . The decrease is:

$$\frac{d\varepsilon}{dt} = v * \frac{dv}{dt} \quad (6)$$

So, the same speed variation  $\Delta v$  has a greater impact on the mechanical energy in the high-speed section of the spacecraft than in the low-speed section, and the speed of the spacecraft at the perigee ( $r_p$ ) is higher than that at the apogee ( $r_a$ ).

For moment of momentum  $h$ , if the speed increment is the normal direction of the orbit surface, it will not affect the size of the  $h$  scalar. The right ascension of ascending node  $\Omega$  can be calculated according to formula (4):

$$\frac{d \tan(\Omega)}{dt} = \frac{d h_x}{-d h_y} \quad (7)$$

Under the condition of a limited speed increment, the following approximate conclusions can be drawn:

$$\begin{cases} \frac{d\Omega}{dt} = \frac{dh_x}{-dh_y} * \frac{1}{\tan(\Omega)^2 + 1} \\ \frac{di}{dt} = \frac{dh_z}{dt} * \frac{1}{(-1)*\sin(i)*h} \\ \frac{d\Omega}{dt} \approx \frac{\text{norm}(r)*dv}{\text{norm}(h)*\sin(i)} \end{cases} \quad (8)$$

It can be seen that, with a certain flight orbit (that is, a certain orbit inclination angle) and the spacecraft mechanical energy, the amount of change in the right ascension of the ascending node is linearly related to the normal speed increment, while the normal speed increment (produced by  $L$ ) is directly proportional to the loss of the spacecraft mechanical energy (produced by  $D$ ).

Therefore, to complete the rapid precession of the corresponding right ascension of the ascending node, as long as the lost orbital mechanical energy is replenished in advance, the spacecraft can maintain its original orbital shape after completing the maneuver (the semi-major axis, the eccentricity, the orbital inclination angle and the perigee amplitude remain unchanged).

### 3 Aerodynamically Assisted Guidance Based on Feature Model

#### 3.1 Guidance Process

The spacecraft deorbit mission consists of the following three main targets:

- (1) Complete the phase adjustment for the spacecraft so that the track of the sub-satellite points meets the requirements of the re-entry trajectory and enters the deorbit standby orbit;
- (2) The spacecraft brakes, adjusts the spacecraft orbit, and enters the deorbit target orbit;

Target 1 mainly aims to change the orbit cycle by changing the semi-major axis of the spacecraft, so as to achieve the purpose of increasing the westward speed of the sub-satellite point track for each orbiting cycle. Target 2 is usually to enter the deorbit orbit in one go through the deorbit braking pulse after the phase adjustment is completed. The apogee of the deorbit orbit is the same as that of the original near-circular orbit, and the perigee is generally lower than the sea level.

To complete this task, the traditional workflow (scheme 1) is as follows:

- (1) Enter the transition orbit through the pulse increment  $\Delta V_1$  (the time required is 1 orbital cycle), increase the current semi-major axis, and increase the current westward speed of the sub-satellite point for each orbiting cycle;
- (2) According to the phase difference of the sub-satellite point of the deorbit standby orbit, wait several orbiting cycles (waiting time  $T_1$ ) to make the current sub-satellite point track coincide with the deorbit standby orbit of the sub-satellite point track;
- (3) Reduce the semi-major axis of the orbit through pulse increment  $\Delta V_2$  (the time required is 1 orbiting cycle) and enter the deorbit standby track;
- (4) Wait several orbiting cycles (waiting time  $T_2$ ), and when the deorbit point is reached, the spacecraft will brake  $\Delta V_3$ . After the braking is completed, the spacecraft will enter the deorbit target orbit.

#### 3.2 Scheme Tradeoffs

From the perspective of time, the greater  $\Delta V_1$  and  $\Delta V_2$ , and the faster the westward speed of the transition orbit sub-satellite point, the less time the deorbit transition takes ( $1+T_1+1+T_2$ ), but the more propellant will be consumed. In terms of energy,  $\Delta V_2$  and  $\Delta V_3$  both decelerate the spacecraft and reduce the semi-major axis of the orbit. If aerodynamic assistance can be used to carry out this work, time and propellant consumption can be significantly optimized. Not only can the response speed for spacecraft missions be improved but also the propellant consumption can be reduced and the on-orbit platform capability for multiple missions can be enhanced.

The optimized task flow (Scheme 2) is as follows:

- (1) Make mission planning in advance. It is required that after a period of  $1 + T1 + 1$  since the mission launch, the spacecraft right enters the deorbit target orbit;
- (2) Enter the transition orbit (ellipse) through pulse increment  $\Delta V1$ , increase the semi-major axis of the current orbit, and increase the westward speed of the sub-satellite point of each current orbiting cycle;
- (3) According to the phase difference of the sub-satellite point of the deorbit standby orbit, wait several orbiting cycles (waiting time  $T1$ ), and after 1 orbiting cycle, meet the deorbit target orbit conditions;
- (4) Reduce the orbit perigee to 85 km through pulse increment  $\Delta V2'$  at the apogee of the current orbit;
- (5) When the spacecraft enters the perigee atmosphere, use the adaptive prediction and correction guidance method based on the feature model to plan the spacecraft's rolling angle, and make the spacecraft's apogee coincide with that of the deorbit standby orbit;
- (6) At the intersection between the current orbit and the deorbit target orbit, use pulse increment  $\Delta V3'$  to make the spacecraft fully enter the deorbit target orbit. Complete the deorbit task.

By comparing the two different deorbit schemes, the following preliminary analysis results are obtained:

- (1) In terms of time consumption, scheme 2 does not save  $T2$  time, but when aerodynamic assistance is used to reduce the apogee height, this part of energy can be converted into the precession of the ascension of the spacecraft ascending intersection point to improve the trajectory of the westward speed of the spacecraft's sub-satellite track;
- (2) In terms of propellant consumption, scheme 2 can completely avoid the consumption of  $\Delta V2$  through aerodynamically assisted orbit change, and at the same time, propellant consumption  $\Delta V3'$  during the final entry into the deorbit target orbit is less than  $\Delta V3$  of deorbit transition orbit braking;
- (3) In terms of observability of the spacecraft orbit, the use of aerodynamics for orbital maneuver and deorbit makes the maneuver dynamics much greater than that through use of propellant, so the spacecraft is also less likely to be tracked and observed, increasing its invisibility.

### 3.3 Feature-Model Based Guidance Method

In order to minimize  $\Delta V3'$  required in step 6 of scheme 2, it is necessary to make the apogee of the spacecraft reach that of the deorbit standby orbit in step 5.

At this time, the aircraft uses aerodynamically assisted apogee maneuver method, and because of the uncertainty of the aerodynamic characteristics and atmospheric parameters of the spacecraft, a fast-response adaptive guidance algorithm is needed to ensure the spacecraft's reaching the predetermined apogee altitude.

During the flight mission, the guidance cycle (generally N seconds) must first be determined. In each guidance cycle, a dynamic equation is used to predict the apogee height of the detector after it leaves the atmosphere. If there is a difference between the predicted apogee height and the target apogee height, the adaptive prediction and correction method will be used to correct and update the guidance rate.

In this paper, the difference between the predicted apogee height and the target apogee height is set as the state quantity  $y(t)$ , the correction value of the spacecraft rolling angle is set as the control variable  $u(t)$ , and the first-order feature model are used for simplified simulation according to the analysis of the spacecraft's orbital motion feature, and the following discretized feature model is given:

$$(k+1) = A(k) * y(k) + B(k) * u(k) \quad (9)$$

where  $y(k)$  is the difference between the current predicted apogee altitude and the target apogee altitude, and  $u(k)$  is the value of the correction of the spacecraft's rolling angle evenly assigned to each time point. The main goal of the algorithm is to continuously modify  $A(k)$  and  $B(k)$  during operation, and finally make  $y$  approach zero. The gradient method is used for parameter identification.

$$\begin{cases} \alpha(k) = [y(k); u(k)] \\ \beta(k) = [A(k); B(k)] \end{cases} \quad (10)$$

Define  $\hat{\beta}(k)$  as the estimated value of  $\beta(k)$ , then use the gradient method to estimate  $\beta(k)$ .

$$\begin{cases} \hat{\beta}(k) = [\hat{A}(k); \hat{B}(k)] \\ \hat{\beta}(k+1) = \hat{\beta}(k) + \frac{(y(k) - \text{dot}(\alpha(k), \beta(k)))}{\text{dot}(\alpha(k), \alpha(k)) + \lambda_2} * (\lambda_1 * \alpha(k)) \end{cases} \quad (11)$$

where  $\lambda_1$  and  $\lambda_2$  are positive constants. After completing the estimation of  $\hat{\beta}(k)$  in each guidance cycle, the value of the rolling angle control (correction amount) is given through linear negative feedback:

$$u(k) = -L * \frac{\hat{A}(k) * y(k)}{\hat{B}(k) + \lambda_3} \quad (12)$$

where  $L$  is a gain parameter greater than 0, and  $\lambda_3$  and  $\hat{B}(k)$  have the same sign to ensure that the absolute value of the denominator will not be too small and cause an unstable state.

## 4 Simulation Verification

### 4.1 Simulation Parameters and Analysis

With the given hypersonic vehicle model used for simulation, the spacecraft flies on a near-circular orbit with a height of 500km. Based on the mission

**Table 1.** Simulation parameter list

Orbital characteristics	Value
Spacecraft mass (Mass)	1000 kg
Spacecraft lift-to-drag ratio (K)	5.0
Propellant specific impulse	3000 N.s/kg
Initial orbit (near-circular)-apogee height	500 km
Deorbit standby orbit-apogee height	300 km
Orbit-inclination angle i	45°
Atmospheric bank angle	-89 +89°
Phase adjustment orbit-apogee height	870 km
Phase adjustment westward advance angle	7.7°

requirements, it completes the westward advance of the 8-degree sub-satellite point track phase, and then enters 500 km (apogee)/0 km (perigee) deorbit target orbit. Based on this, the following simulation parameters are provided (Table 1):

According to the analysis of simulation conditions, a speed increment of about 100 m/s is required to enter the phase adjustment orbit from the initial orbit. On the phase adjustment orbit, the westward phase angle of each orbit cycle is 0.94°. To complete the 7.7-degree phase angle adjustment, it takes about 9 circles.

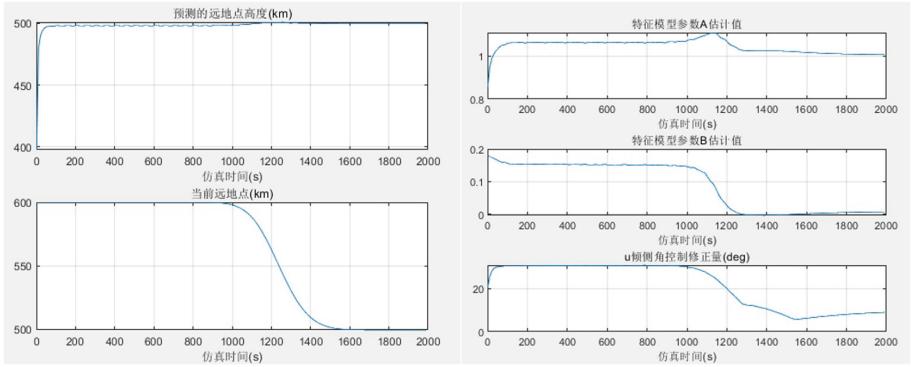
If during the 7th circle of the phase adjustment, the aircraft's aerodynamic assistance is used to adjust the ascending node, according to the analysis of [6], on this orbit, a (100–27) m/s speed increment (in order to adjust the spacecraft apogee to about 600 km for deorbit) can achieve ascending node precession of about 0.87°, which can reduce the phase adjustment time by one circle, and the spacecraft can end the phase adjustment process during the eighth cycle.

## 4.2 Adaptive Guidance Strategy

After the eighth circle, through the above-mentioned aerodynamically assisted ascending node adjustment, the apogee of the spacecraft orbit will be at about 600 km, and it is necessary to adaptively control the rolling angle at the perigee during the next circle to make the apogee reach 500 km of the target orbit. Through feature modeling and adaptive guidance algorithm control, after entry into the atmosphere following the 9th circle, the final apogee deviation is -0.38037 km.

## 4.3 Entering the Deorbit Target Orbit

After completing the control in Sect. 4.2, the apogee height of the spacecraft is less than or equal to that of the deorbit target orbit, and the perigee height is greater than that of the deorbit target orbit. Therefore, according to the ellipse



**Fig. 1.** Adaptive orbit control for off tracking

formula, the intersection of these two coplanar orbits can be calculated and the transfer from the current track to the deorbit target orbit can be completed through one pulse at the intersection. According to the above-mentioned conditions, the projections of the two sets of orbit velocity vectors at the intersection point on the orbit plane through simulation calculation are (Fig. 1):

$$V1 = [-4505.225; 6619.250] \quad (13)$$

$$V2 = [-4491.380; 6542.073] \quad (14)$$

The required speed increment is:

$$\Delta V3' = \text{norm}(V1 - V2) = 78.41 \text{ m/s} \quad (15)$$

#### 4.4 Resource Effect Analysis

In a task with given parameters, if scheme 1 is adopted, the total resources required are:

$$\begin{cases} T1 = 1 + 9 + 1 + N + 112\text{cycles} \\ \Delta V = \Delta V1 + \Delta V2 + \Delta V3 = 100 + 100 + 198 = 398 \text{ m/s} \end{cases} \quad (16)$$

The total amount of time and the total amount of propellant consumption decreased by about 18%, and better maneuverability and prevention from being continuously observed were realized, which increased the resource utilization and mission agility of the spacecraft.

## 5 Conclusion

Aiming at the deorbit mission of hypersonic aircraft, an aerodynamically assisted method is proposed to improve the spacecraft maneuverability during deorbit,

and reduce propellant consumption, so that the spacecraft can complete the deorbit mission more quickly and economically, and leave more time and resources for subsequent tasks.

## References

1. Ying, N.A.N.: The progress of reentry trajectory and control of space vehicle. *Mis-siles Space veh.* **5**, 1–11 (1994)
2. Hu, J.: Adaptive predictive guidance: a unified guidance method. *Aerosp. Control Appl.* **45**(4), 53–63 (2010)
3. Bing-nan, K.A.N.G.: Optimization of hypersonic glide trajectory based on nonlinear planning. *Flight Dyn.* **26**(3), 49–53 (2008)
4. Nai-gang, C.U.I.: A review of on-orbit servicing. *J. Astronaut.* **28**(4), 805–811 (2007)
5. Mao-mao, L.I.: An adaptive predictor-corrector method of mars entry phase. *J. Astronaut.* **38**(5), 506–515 (2017)
6. Sengupta, P., Vadali, S.R., Alfriend, K.T.: Modeling and control of satellite formations in high eccentricity orbits. *J. Astronaut. Sci.* **52**(1–2), 149–168 (2004)
7. Wong, H., Pan, H., Kapila, V.: Output feedback control for spacecraft formation flying with coupled translation and attitude dynamics. In: Proceedings of the American Control Conference, Portland, OR, June 2005
8. Chen, G., Kang, X.-W., Yan, G.-R., Chen, S.-L.: Real time robust adaptive reentry guidance law based on pseudo-spectral method. *J. Syst. Simul.* **20**(20), 5623–5626, 5634 (2008)



# Fault Estimation of Switched Linear Systems with Actuator and Sensor Faults

Chunying Su<sup>1</sup>, Jianting Lyu<sup>1(✉)</sup>, Xin Wang<sup>1</sup>, and Dai Gao<sup>2</sup>

<sup>1</sup> School of Mathematical Science, Heilongjiang University,  
Harbin, People's Republic of China

[lvjianting@hlju.edu.cn](mailto:lvjianting@hlju.edu.cn)

<sup>2</sup> School of Mechatronics Engineering, Harbin Institute of Technology,  
Harbin, People's Republic of China

**Abstract.** This paper addresses the multiple fault estimation problem for the switched linear systems with actuator and sensor faults. By introducing an intermediate variable and augmenting state and sensor faults into a new vector, an fault estimation approach is proposed to contemporaneously estimate the state, the actuator and sensor faults. It is shown that under switching signals with average dwell time closed-loop signals remain uniformly ultimately bounded. Finally, a case study is given to illustrate the effectiveness of the proposed scheme.

**Keywords:** Fault estimation · Switched linear systems · Estimator · Average dwell time

## 1 Introduction

Switched systems are a branch of hybrid systems which consist of multiple subsystems, either continuous subsystems or discrete ones, that are coordinated with switching rules, have attracted remarkable attentions and been widely applied in diverse areas such as aircraft control, communication systems and water-quality control system [1–4]. Aiming at the stability problem in time-varying switched nonlinear systems under restricted switching, authors in [5] presented the stability of the nominal-like part of system which requires a weak Lyapunov function. Besides, quite a few researchers concentrated on the stability analysis of switched systems, see [6–10]. On the other hand, actuator and/or sensor failures may cause the poor performance, even the instability of the system. Therefore, because of its theoretical and practical significance, fault detection and isolation (FDI) and fault estimation (FE) also call a great interests of scholars.

Due to the high-leveled security and reliability requirements, FDI and FE has been adopted to perform fault detection, fault source location, fault influence estimation and thus by fault-tolerant controller designing, to guarantee the system stability [11–14]. Particularly, FE can provide more accurate information of the fault occurrence from the perspective of time and space scales. Several efforts were devoted to the FE of nonlinear system [15–20]. For linear systems

with uncertainties, sliding mode observers was adopted to deduce a robust actuator fault reconstruction method, in which not only the fault was detected and isolated, but also an estimate of the fault was provided [21]. In [22], aiming at the nonlinear multi-agent system with actuator faults, a design method of bidirectional interactive fault estimation and fault-tolerant control based on hierarchical structure was proposed. For a class of Markovian jump systems, contemporaneous estimations of states, system fault and sensor fault was investigated in [23]. Recently some studies on switched systems about FE have also been reported. For the discrete-time switched system with finite-frequency, an observer design method for actuator fault estimation was considered in [24]. FE for descriptor switched systems was addressed and meanwhile the precise state estimations and fault predictions were also achieved in [25].

Motivated by the above-mentioned researches, this note investigated the FE problem of the switched linear systems with sensor and actuator faults. Based on observer design and average dwell time approach, an estimator is developed for the switched linear systems with multiple faults. The bounds of the faults and their derivatives are not needed here. To estimate the actuator and sensor faults simultaneously, we introduce the intermediate variable and augment state and sensor faults into a new vector. It is shown that under the proposed approach the closed-loop signals are uniformly ultimately bounded.

The rest of the paper is arranged as follows. In Sect. 2, the necessary preliminaries are introduced and the issue of concern is stated. Design of the intermediate estimator is presented in Sect. 3. Section 4 gives the simulation results and the effectiveness discussions, and finally the conclusions are drawn in Sect. 5.

## 2 Preliminaries and Problem Statement

A switched linear system is investigated and described as follows

$$\dot{x}(t) = A_{\sigma(t)}x(t) + B_{\sigma(t)}u(t) + E_{\sigma(t)}s_a(t), \quad (1)$$

$$\varpi(t) = C_{\sigma(t)}x(t) + D_{\sigma(t)}s_f(t), \quad (2)$$

where  $\sigma(t) : R_+ \rightarrow \mathcal{P} = \{1, 2, \dots, P\}$  is a switching signal that is also considered as a piecewise continuous function of time;  $x(t) \in R^n$ ,  $u(t) \in R^m$  and  $\varpi(t) \in R^p$  are the state variable, the control input and the output measurement, respectively; the unknown time-varying fault  $s_a(t)$  satisfies  $\|\dot{s}_a(t)\| \leq \theta$  with  $\theta \geq 0$ , and the unknown time-varying sensor fault  $s_f(t)$  satisfies  $\|\dot{s}_f(t)\| \leq \eta$  with  $\eta \geq 0$ ,  $s_a(t) \in R^r$  and  $s_f(t) \in R^q$  are the fault signals which represent the process ones or actuator ones (when  $E_{\sigma(t)} = B_{\sigma(t)}$ ) and the sensor faults, respectively;  $A_{\sigma(t)}, B_{\sigma(t)}, C_{\sigma(t)}, D_{\sigma(t)}$  and  $E_{\sigma(t)}$  are real constant matrices with suitable dimensions, and  $(A_{\sigma(t)}, C_{\sigma(t)})$  is assumed to be observable,  $E_{\sigma(t)}, D_{\sigma(t)}$  is of full column rank, i.e.,  $\text{rank } (E_{\sigma(t)}) = r$ ,  $\text{rank } (D_{\sigma(t)}) = q$ .

**Definition 1.** Switching signal  $\sigma(t)$  has average dwell time  $\tau_a$  if two positive numbers  $N_0$  and  $\tau_a$  exist and satisfy that

$$N_{\sigma(t)}(T, t) \leq N_0 + \frac{T - t}{\tau_a}, \forall T \geq t \geq 0, \quad (3)$$

in which,  $N_{\sigma(t)}(T, t)$  counts the switches that occur within the interval  $[t, T]$ .

**Assumption 1.** When  $t \in [t_i, t_{i+1})$ ,  $\sigma(t) = i$ ,  $i \in P$ , that is, the  $i$ th subsystem is active. Also, we assume  $i \neq i+1$  for all  $i$ .

The objective of this paper is to deal with the fault estimation problem of the switched linear systems by designing the fault estimator.

### 3 Main Result

For the augmented systems depicted as (1) and (2), an intermediate estimator is constructed to solve the multi-fault estimation problem. Let  $\bar{x}(t) = [x(t)^T s_f(t)^T]^T$ , then the systems (1) and (2) will be revoiced in the following augmented form as

$$\dot{\bar{x}}(t) = \bar{A}_{\sigma(t)} \bar{x}(t) + \bar{B}_{\sigma(t)} u(t) + \bar{E}_{\sigma(t)} s_a(t) + \bar{M}_{\sigma(t)} \dot{s}_f(t), \quad (4)$$

$$\varpi(t) = \bar{C}_{\sigma(t)} \bar{x}(t), \quad (5)$$

in which  $\bar{A}_{\sigma(t)} = \begin{bmatrix} A_{\sigma(t)} & 0 \\ 0 & 0 \end{bmatrix}$ ,  $\bar{B}_{\sigma(t)} = \begin{bmatrix} B_{\sigma(t)} \\ 0 \end{bmatrix}$ ,  $\bar{E}_{\sigma(t)} = \begin{bmatrix} E_{\sigma(t)} \\ 0 \end{bmatrix}$ ,  $\bar{M}_{\sigma(t)} = \begin{bmatrix} 0 \\ I \end{bmatrix}$ ,  $\bar{C}_{\sigma(t)} = [C_{\sigma(t)} \ D_{\sigma(t)}]$ .

We present the following assumption.

**Assumption 2.** For every complex number  $\lambda$  with non-negative real part

$$\text{rank} \begin{bmatrix} \bar{A}_{\sigma(t)} + \lambda I & \bar{E}_{\sigma(t)} \\ \bar{C}_{\sigma(t)} & 0 \end{bmatrix} = n + q + \text{rank}(\bar{E}_{\sigma(t)}). \quad (6)$$

Define an intermediate variable as  $\gamma(t) = s_a(t) - K_{\sigma(t)} \bar{x}(t)$ , where  $K_{\sigma(t)}$  will be given later.

Using (4), it follows that  $\gamma(t)$  satisfies the following dynamics

$$\begin{aligned} \dot{\gamma}(t) = & \dot{s}_a(t) - K_{\sigma(t)} (\bar{A}_{\sigma(t)} \bar{x}(t) \bar{B}_{\sigma(t)} u(t) + \bar{E}_{\sigma(t)} \gamma(t) + \bar{E}_{\sigma(t)} K_{\sigma(t)} \bar{x}(t) \\ & + \bar{M}_{\sigma(t)} \dot{s}_f(t)). \end{aligned} \quad (7)$$

Then we propose the following estimator as

$$\dot{\hat{x}}(t) = \bar{A}_{\sigma(t)} \hat{x}(t) + \bar{B}_{\sigma(t)} u(t) + \bar{E}_{\sigma(t)} \hat{s}_a(t) + L_{\sigma(t)} (\varpi(t) - \hat{\varpi}(t)), \quad (8)$$

$$\dot{\hat{\gamma}}(t) = -K_{\sigma(t)} \bar{E}_{\sigma(t)} \hat{\gamma}(t) - K_{\sigma(t)} (\bar{A}_{\sigma(t)} \hat{x}(t) + \bar{B}_{\sigma(t)} u(t) + \bar{E}_{\sigma(t)} K_{\sigma(t)} \hat{x}(t)), \quad (9)$$

where  $\hat{\varpi}(t) = \bar{C}_{\sigma(t)} \hat{x}(t)$ ,  $\hat{s}_a(t) = \hat{\gamma}(t) + K_{\sigma(t)} \hat{x}(t)$ ,  $\hat{s}_f(t) = \tilde{C}_{\sigma(t)} \hat{x}(t)$ ,  $\tilde{C}_{\sigma(t)} = [0, I_q]$ ,  $\hat{x}(t)$ ,  $\hat{\gamma}(t)$ ,  $\hat{\varpi}(t)$ ,  $\hat{s}_f(t)$  and  $\hat{s}_a(t)$  are the estimations of  $\bar{x}(t)$ ,  $\gamma(t)$ ,  $\varpi(t)$ ,  $s_f(t)$  and  $s_a(t)$ , respectively.

Define  $\xi(t) = \bar{x}(t) - \hat{x}(t)$ ,  $\delta(t) = \gamma(t) - \hat{\gamma}(t)$  and  $\iota(t) = s_a(t) - \hat{s}_a(t)$ . Then the error systems can be obtained by

$$\dot{\xi}(t) = (\bar{A}_{\sigma(t)} - L_{\sigma(t)}\bar{C}_{\sigma(t)})\xi(t) + \bar{E}_{\sigma(t)}\iota(t) + \bar{M}_{\sigma(t)}\dot{s}_f(t), \quad (10)$$

$$\dot{\delta}(t) = \dot{s}_a(t) - K_{\sigma(t)}\bar{E}_{\sigma(t)}\delta(t) - K_{\sigma(t)}((\bar{A}_{\sigma(t)} + \bar{E}_{\sigma(t)}K_{\sigma(t)}))\xi(t) + \bar{M}_{\sigma(t)}\dot{s}_f(t)). \quad (11)$$

**Theorem 1.** Consider the error systems (10) and (11) under Assumptions (1)–(2). Then the estimator (8)–(9) ensures the states of the error systems are uniformly ultimately bounded if for given scalars  $\omega_{\sigma(t)} > 0$ ,  $\varepsilon > 0$ , there exist positive definite matrices  $Q_{\sigma(t)}$ ,  $H_{\sigma(t)}$  and  $Z_{\sigma(t)}$  such that

$$II = \begin{bmatrix} \Pi_{11\sigma(t)} & \Pi_{12\sigma(t)} & Q_{\sigma(t)}\bar{M}_{\sigma(t)} & 0 & 0 \\ * & \Pi_{22\sigma(t)} & 0 & \omega_{\sigma(t)}Z_{\sigma(t)}\bar{E}_{\sigma(t)}^T\bar{M}_{\sigma(t)} & Z_{\sigma(t)} \\ * & * & -\frac{1}{\varepsilon}I & 0 & 0 \\ * & * & * & -\frac{1}{\varepsilon}I & 0 \\ * & * & * & * & -\frac{1}{\varepsilon}I \end{bmatrix} < 0, \quad (12)$$

where  $\Pi_{11\sigma(t)} = (\bar{A}_{\sigma(t)} - L_{\sigma(t)}\bar{C}_{\sigma(t)})^TQ_{\sigma(t)} + Q_{\sigma(t)}(\bar{A}_{\sigma(t)} - L_{\sigma(t)}\bar{C}_{\sigma(t)}) + \omega_{\sigma(t)}Q_{\sigma(t)}\bar{E}_{\sigma(t)}\bar{E}_{\sigma(t)}^T + \omega_{\sigma(t)}\bar{E}_{\sigma(t)}\bar{E}_{\sigma(t)}^TQ_{\sigma(t)}$ ,  $\Pi_{12\sigma(t)} = Q_{\sigma(t)}\bar{E}_{\sigma(t)} - \omega_{\sigma(t)}\bar{A}_{\sigma(t)}^T\bar{E}_{\sigma(t)}Z_{\sigma(t)} - \omega_{\sigma(t)}^2\bar{E}_{\sigma(t)}\bar{E}_{\sigma(t)}^T\bar{E}_{\sigma(t)}Z_{\sigma(t)}$ ,  $\Pi_{22\sigma(t)} = \frac{1}{\varepsilon}I - \omega_{\sigma(t)}(Z_{\sigma(t)}\bar{E}_{\sigma(t)}^T\bar{E}_{\sigma(t)} + \bar{E}_{\sigma(t)}^T\bar{E}_{\sigma(t)}Z_{\sigma(t)})$ , design  $K_{\sigma(t)} = \omega_{\sigma(t)}\bar{E}_{\sigma(t)}^T$ ,  $\tau_a > \frac{\ln\mu}{a_0}$  for given  $a_0 > 0$  and  $\mu = \max\{\frac{\lambda_{\max}V_k(\tilde{e}(t))}{\lambda_{\min}V_l(\tilde{e}(t))}, k, l \in P\}$ , and the estimator gain can be obtained as  $L_{\sigma(t)} = Q_{\sigma(t)}^{-1}H_{\sigma(t)}$ .

Proof: Using the following Lyapunov function candidate

$$V_{\sigma(t)}(t) = \xi^T(t)Q_{\sigma(t)}\xi(t) + \delta^T(t)Z_{\sigma(t)}\delta(t). \quad (13)$$

When  $t \in [t_i, t_{i+1})$ ,  $\sigma(t) = p$ ,  $p \in P$ , taking the derivative of  $V_p(t)$  gives

$$\begin{aligned} \dot{V}_p(t) &= \xi^T(t)((\bar{A}_p - L_p\bar{C}_p)^TQ_p + Q_p(\bar{A}_p - L_p\bar{C}_p))\xi(t) + 2\xi^T(t)Q_p\bar{E}_p\iota(t) \\ &\quad + 2\xi^T(t)Q_p\bar{M}_p\dot{s}_f(t) + 2\delta^T(t)Z_p\dot{s}_a(t) - 2\omega_p\delta^T Z_p(t)\bar{E}_p^T\bar{E}_p\delta(t) - 2\omega_p\delta^T Z_p(t) \\ &\quad \times \bar{E}_p^T\bar{A}_p\xi(t) - 2\omega_p^2\delta^T Z_p(t)\bar{E}_p^T\bar{E}_p\bar{E}_p^T\xi(t) - 2\omega_p\delta^T(t)Z_p\bar{E}_p^T\bar{M}_p\dot{s}_f(t). \end{aligned} \quad (14)$$

Using  $\iota(t) = \delta(t) + \omega_p\bar{E}_p^T\xi(t)$  into (14) gives

$$\begin{aligned} \dot{V}_p(t) &= \xi^T(t)((\bar{A}_p - L_p\bar{C}_p)^TQ_p + Q_p(\bar{A}_p - L_p\bar{C}_p))\xi(t) + 2\xi^T(t)Q_p\bar{E}_p\delta(t) \\ &\quad + 2\omega_p\xi^T(t)Q_p\bar{E}_p\bar{E}_p^T\xi(t) + 2\xi^T(t)Q_p\bar{M}_p\dot{s}_f(t) + 2\delta^T Z_p(t)\dot{s}_a(t) - 2\omega_p \\ &\quad \times \delta^T(t)Z_p\bar{E}_p^T\bar{E}_p\delta(t) - 2\omega_p\delta^T(t)Z_p\bar{E}_p^T\bar{A}_p\xi(t) - 2\omega_p^2\delta^T(t)Z_p\bar{E}_p^T\bar{E}_p \\ &\quad \times \bar{E}_p^T\xi(t) - 2\omega_p\delta^T(t)Z_p\bar{E}_p^T\bar{M}_p\dot{s}_f(t). \end{aligned} \quad (15)$$

The follows shall be noted

$$2\delta^T(t)Z_p\dot{s}_a(t) \leq \varepsilon\delta^T(t)Z_pZ_p^T\delta(t) + \frac{1}{\varepsilon}\theta^2, \quad (16)$$

$$2\xi^T(t)Q_p\bar{M}_p\dot{s}_f(t) \leq \varepsilon\xi^T(t)Q_p\bar{M}_p\bar{M}_p^TQ_p\xi(t) + \frac{1}{\varepsilon}\eta^2, \quad (17)$$

and

$$-2\omega_p \delta^T(t) Z_p \bar{E}_p^T \bar{M}_p \dot{s}_f(t) \leq \varepsilon \omega_p^2 \delta^T(t) Z_p \bar{E}_p^T \bar{M}_p \bar{M}_p^T \bar{E}_p Z_p^T \delta(t) + \frac{1}{\varepsilon} \eta^2. \quad (18)$$

From (15)–(18), one has

$$\begin{aligned} \dot{V}_p(t) &\leq \xi^T(t)((\bar{A}_p - L_p \bar{C}_p)^T Q_p + Q_p(\bar{A}_p - L_p \bar{C}_p))\xi(t) + 2\xi^T(t)Q_p \bar{E}_p \delta(t) \\ &\quad + 2\omega_p \xi^T(t)Q_p \bar{E}_p \bar{E}_p^T \xi(t) + \varepsilon \xi^T(t)Q_p \bar{M}_p \bar{M}_p^T Q_p \xi(t) + \frac{1}{\varepsilon} \eta^2 + \varepsilon \delta^T(t) \\ &\quad \times Z_p Z_p^T \delta(t) + \frac{1}{\varepsilon} \theta^2 - 2\omega_p \delta^T(t) Z_p \bar{E}_p^T \bar{E}_p \delta(t) - 2\omega_p \xi^T(t) Z_p \bar{A}_p^T \bar{E}_p \delta(t) \\ &\quad - 2\omega_p^2 \xi^T(t) Z_p \bar{E}_p \bar{E}_p^T \bar{E}_p \delta(t) + \varepsilon \omega_p^2 \delta^T(t) Z_p \bar{E}_p^T \bar{M}_p \bar{M}_p^T \bar{E}_p Z_p^T \delta(t) + \frac{1}{\varepsilon} \eta^2. \end{aligned} \quad (19)$$

Denote  $\tilde{e} = [\xi^T(t) \ \delta^T(t)]^T$ , (19) can be represented as

$$\dot{V}_p(t) \leq \tilde{e}^T \varrho_{1p} \tilde{e} + \frac{1}{\varepsilon} \theta^2 + \frac{2}{\varepsilon} \eta^2, \quad (20)$$

where

$$\varrho_{1p} = \begin{bmatrix} \varrho_{11p} & \varrho_{12p} \\ * & \varrho_{22p} \end{bmatrix}, \quad (21)$$

and  $\varrho_{11p} = (\bar{A}_p - L_p \bar{C}_p)^T Q_p + Q_p(\bar{A}_p - L_p \bar{C}_p) + \omega_p Q_p \bar{E}_p \bar{E}_p^T + \omega_p \bar{E}_p \bar{E}_p^T Q_p + \varepsilon Q_p \bar{M}_p \bar{M}_p^T Q_p$ ,  $\varrho_{12p} = P_p \bar{E}_p - \omega_p Z_p \bar{A}_p^T \bar{E}_p - \omega_p^2 Z_p \bar{E}_p \bar{E}_p^T \bar{E}_p$ ,  $\varrho_{22p} = \varepsilon Z_p Z_p^T - 2\omega_p Z_p \bar{E}_p^T \bar{E}_p - \varepsilon \omega_p^2 Z_p \bar{E}_p^T \bar{M}_p \bar{M}_p^T \bar{E}_p Z_p^T$ .

It then follows from (13) that

$$\begin{aligned} V_p(t) &\leq \lambda_{max}(Q_p) \|\xi(t)\|^2 + \lambda_{max}(Z_p) \|\delta(t)\|^2 \\ &\leq \max[\lambda_{max}(Q_p), \lambda_{max}(Z_p)](\|\xi(t)\|^2 + \|\delta(t)\|^2). \end{aligned} \quad (22)$$

Let  $\varrho_{2p} = -\varrho_{1p}$ , and if  $\varrho_{1p} < 0$ , i.e.,  $\varrho_{2p} > 0$ , then it can be deduced that

$$\begin{aligned} \dot{V}_p(t) &\leq \lambda_{max}(\varrho_{1p})(\|\xi(t)\|^2 + \|\delta(t)\|^2) \\ &= -\lambda_{min}(\varrho_{2p})(\|\xi(t)\|^2 + \|\delta(t)\|^2) \\ &\leq -\rho_p V_p(t) + \varsigma_p, \end{aligned} \quad (23)$$

where

$$\rho_p = \frac{\lambda_{min}(\varrho_{2p})}{\max[\lambda_{max}(Q_p), \lambda_{max}(Z_p)]}, \varsigma_p = \frac{1}{\varepsilon} \theta^2 + \frac{2}{\varepsilon} \eta^2. \quad (24)$$

Let

$$\kappa = \min\{\rho_p\}, \alpha = \max\{\varsigma_p\}. \quad (25)$$

From (25), we have

$$\dot{V}_p(t) \leq -\kappa V_p(t) + \alpha. \quad (26)$$

Consider the function as

$$\phi(t) = \exp\{\kappa t\} V_{\sigma(t)}(\tilde{e}(t)). \quad (27)$$

On each interval  $[t_i, t_{i+1})$ , from (27) one has

$$\dot{\phi}(t) \leq \alpha \exp\{\kappa t\}, t \in [t_i, t_{i+1}). \quad (28)$$

Note that  $V_k(\tilde{e}(t)) \leq \mu V_l(\tilde{e}(t)), \forall k, l \in P$ , then we have

$$\phi(t_{i+1}) \leq \mu [\phi(t_i) + \int_{t_i}^{t_{i+1}} \alpha \exp\{\kappa t\} dt]. \quad (29)$$

Furthermore, from (29) we have

$$\begin{aligned} \phi(T^-) &\leq \mu^{N_\sigma(T,0)} [\phi(0) + \sum_{i=0}^{N_\sigma(T,0)-1} \mu^{-i} \int_{t_i}^{t_{i+1}} \alpha \exp\{\kappa t\} dt \\ &\quad + \mu^{-N_\sigma(T,0)-1} \int_{t_{N_\sigma(T,0)}}^T \alpha \exp\{\kappa t\} dt]. \end{aligned} \quad (30)$$

For any  $\psi \in (0, \kappa - (\frac{\ln \mu}{\tau_a}))$ , one has  $\tau_a > \frac{\ln \mu}{\kappa - \psi}$ . It then follows from (3) that

$$N_\sigma(T, t) \leq N_0 + \frac{(\kappa - \psi)(T, t)}{\ln \mu}, \forall T \geq t \geq 0. \quad (31)$$

Since  $\psi < \kappa$ , we have

$$\int_{t_i}^{t_{i+1}} \alpha \exp\{\kappa t\} dt \leq \exp\{(\kappa - \psi)t_{i+1}\} \int_{t_i}^{t_{i+1}} \alpha \exp\{\psi t\} dt. \quad (32)$$

From (30) and (32), one has

$$\phi(T^-) \leq \mu^{N_\sigma(T,0)} \phi(0) + \mu^{1+N_0} \exp\{(\kappa - \psi)T\} \int_0^T \alpha \exp\{\psi t\} dt. \quad (33)$$

It is easy to see that there exist  $\underline{\varphi}, \bar{\varphi}$ , such that  $\underline{\varphi}(\tilde{e}) \leq V_k(\tilde{e}) \leq \bar{\varphi}(\tilde{e})$ , which indicates that

$$\begin{aligned} \underline{\varphi}(\|\tilde{e}(T)\|) &\leq \exp\{N_0 \ln \mu\} \exp\left\{\left(\frac{\ln \mu}{\tau_a} - \kappa\right)T\right\} \\ &\times \bar{\varphi}(\|\tilde{e}(0)\|) + \mu^{1+N_0} \frac{\alpha}{\psi}, \forall T > 0. \end{aligned} \quad (34)$$

Furthermore, for any given constant  $\vartheta > 0$ , can obtain  $\mu^{1+N_0} \frac{\alpha}{\psi} \leq \frac{1}{2} \vartheta^2$ , then can obtain  $\lim_{t \rightarrow \infty} \tilde{e}^2(t) \leq \vartheta^2$ , therefore,  $[\xi^T(t), \delta^T(t)]^T$  is uniformly ultimately bounded and  $\iota(t)$  is uniformly bounded. This proof ends here.

Based on the above analysis, the observer's designing process can be reduced to the following.

---

**Algorithm 1.** Fault estimator Design
 

---

- 1) States and faults are augmented to a new vector, and the augmented systems are constructed as (4)–(5);
  - 2) Introduce an intermediate variable  $\gamma(t)$  and propose a designed estimator as (8)–(9);
  - 3) Find positive definite matrix  $Q_p$ , and multiply  $Q_p$  by both sides of (12) to get (21);
  - 4) Select the observer gains  $L_p = Q_p^{-1}H_p$ , and calculate the average dwell time  $\tau_a$ , then implement the observer in (8)–(9).
- 

*Remark 1.* Observer that if  $s_a(t) = 0$  ( $or s_f(t) = 0$ ), then multiple faults estimation problem will reduce to a single fault estimation problem. With the similar analysis the sufficient condition for the single fault estimation problem can be achieved.

*Remark 2.* Here an fault estimation approach for the switched linear systems is given to estimate the actuator and sensor faults by introducing an intermediate variable and augmenting state and sensor faults into a new vector. The fault estimators can be further utilized for compensation in fault-tolerant control problem.

## 4 Simulation Illustration

In this section, a typical instance is applied to verify the significance of the proposed algorithm. Inspect the switched linear system (1)-(2) with two subsystems and the system parameters are selected as

$$A_1 = \begin{bmatrix} -1.35 & 0.98 \\ 17.1 & -1.85 \end{bmatrix}, B_1 = \begin{bmatrix} -0.13 & -0.013 \\ 14.2 & 10.75 \end{bmatrix}, C_1 = [7 \ 0], E_1 = [0.01 \ 0.01],$$

$$A_2 = \begin{bmatrix} -1.87 & 0.98 \\ 12.6 & -2.63 \end{bmatrix}, B_2 = \begin{bmatrix} -0.16 & -0.005 \\ -29.2 & 21.3 \end{bmatrix}, C_2 = [3 \ 0], E_2 = [0.1 \ 0.1],$$

$$D_1 = 1, D_2 = 2.$$

The actuator fault  $s_a(t)$  and sensor fault  $s_f(t)$  are considered as Table 1.

The system state starts from  $x_1(0) = x_2(0) = [0 \ 0]^T$ , and there defines the input as  $u(t) = k_j y$  ( $j = 1, 2$ ) with  $k_1 = [1 \ 1]$  and  $k_2 = [1.3 \ 1.3]$ . Selecting  $\omega_1 = 1, \omega_2 = 1$ , can obtains

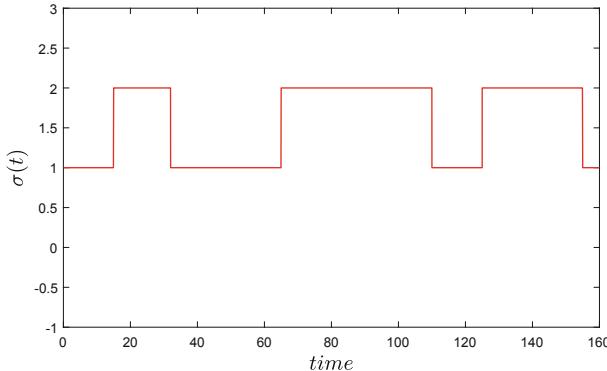
$$Q_1 = \begin{bmatrix} 1.7224 & -1.6924 & -1.3301 \\ -1.6924 & 1.6871 & 1.3295 \\ -1.3301 & 1.3295 & 1.0484 \end{bmatrix}, L_1 = \begin{bmatrix} 216.3369 \\ 225.2579 \\ -9.3048 \end{bmatrix},$$

$$Q_2 = \begin{bmatrix} 0.4797 & -0.4671 & -0.9285 \\ -0.4671 & 0.4643 & 0.9292 \\ -0.9285 & 0.9292 & 1.8641 \end{bmatrix}, L_2 = \begin{bmatrix} 307.6830 \\ 507.2608 \\ -98.4601 \end{bmatrix}.$$

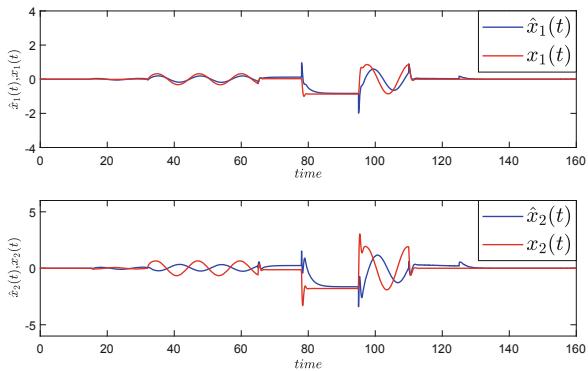
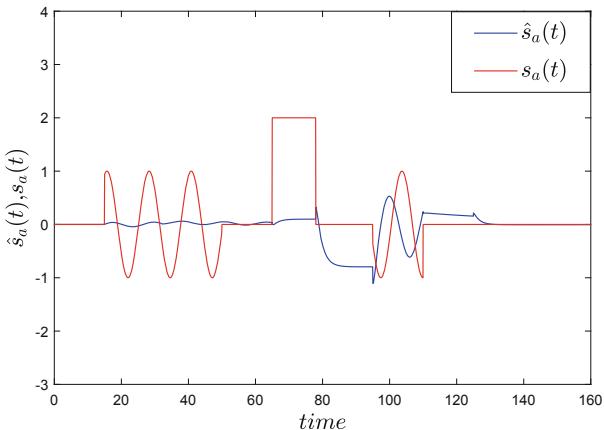
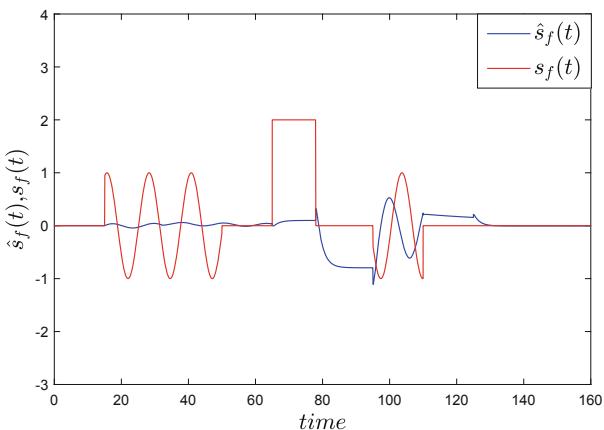
The switching signal is shown in Fig. 1. The states and their estimation results are illustrated in Fig. 2, the fault  $s_a(t)$  and its estimation result is illustrated in Fig. 3, the fault  $s_f(t)$  and its estimation result is illustrated in Fig. 4. It can be seen from Figs. 2, 3 and 4 that satisfactory estimation performance has been achieved.

**Table 1.** FAULTS  $s_a(t)$  and  $s_f(t)$

$t$	[0s,15s)	[15s,32s)	[32s,50s)	[50s,65s)	[65s,78s)
$s_a(t)$	0	$\sin(t)$	$\sin(t)$	0	2
$t$	[78s,95s)	[95s,110s)	[110s,125s)	[125s,155s)	[155s,160s]
$s_a(t)$	0	$\sin(t)$	0	0	0
$t$	[0s,15s)	[15s,32s)	[32s,50s)	[50s,65s)	[65s,78s)
$s_f(t)$	0	$\sin(t)$	$\sin(t)$	0	2
$t$	[78s,95s)	[95s,110s)	[110s,125s)	[125s,155s)	[155s,160s]
$s_f(t)$	0	$\sin(t)$	0	0	0



**Fig. 1.** Switching signal

**Fig. 2.** System state and estimate**Fig. 3.** Actuator fault and estimate**Fig. 4.** Sensor fault and estimate

## 5 Conclusion

The fault estimation problem for the switched linear systems has been developed in this paper. Considering the case that actuator and sensor failures, we propose an estimator, where an intermediate variable is introduced and state and sensor faults are augmented. With the proposed approach, it is proved the state, the actuator and sensor faults can be estimated simultaneously and the closed-loop signals are uniformly ultimately bounded. Future work will focus on fault Tolerant control for the switched linear systems.

**Acknowledgments.** This work was supported in part by the National Natural Science Foundation of China (Grant nos. 61873087 and 61703148), the Natural Science Foundation of Heilongjiang Province (Grant nos. F2017023).

## References

1. He, X., Zhao, J.: Multiple Lyapunov functions with blending for induced  $L_2$ -norm control of switched LPV systems and its application to an F-16 aircraft model. *Asian J. Control* **16**(1), 149–161 (2014)
2. Ma, D., Zhao, J.: Stabilization of networked switched linear systems: an asynchronous switching delay system approach. *Syst. Control Lett.* **77**, 46–54 (2015)
3. Zong, G., Wang, R., Zheng, W., Hou, L.: Finite-time  $H_\infty$  control for discrete-time switched nonlinear systems with time delay. *Int. J. Robust Nonlinear Control* **25**(6), 914–936 (2015)
4. Yang, X., Lu, J.: Finite-time synchronization of coupled networks with Markovian topology and impulsive effects. *IEEE Trans. Autom. Control* **61**(8), 2256–2261 (2016)
5. Daafouz, J., Riedinger, P., Iung, C.: Stability analysis and control synthesis for switched systems: a switched Lyapunov function approach. *IEEE Trans. Autom. Control* **47**(11), 1883–1887 (2002)
6. Chatterjee, D., Liberzon, D.: Stability analysis of deterministic and stochastic switched systems via a comparison principle and multiple Lyapunov functions. *SIAM J. Control Optim.* **45**(1), 174–206 (2006)
7. Kang, Y., Zhai, D., Liu, G., Zhao, Y.: On input-to-state stability of switched stochastic nonlinear systems under extended asynchronous switching. *IEEE Trans. Cybern.* **46**(5), 1092–1105 (2016)
8. Mancilla-Aguilar, J., Haimovich, H., Garca, R.: Global stability results for switched systems based on weak Lyapunov functions. *IEEE Trans. Autom. Control* **62**(6), 2764–2777 (2017)
9. Wu, X., Tang, Y., Cao, J., Mao, X.: Stability analysis for continuous-time switched systems with stochastic switching signals. *IEEE Trans. Autom. Control* **63**(9), 3083–3090 (2018)
10. Mancillaaguilar, J., Haimovich, H.: Uniform input-to-state stability for switched and time-varying impulsive systems. *IEEE Trans. Autom. Control* (2020)
11. Zhang, K., Jiang, B., Chen, M., Yan, X.: Distributed fault estimation and fault-tolerant control of interconnected systems. *IEEE Trans. Cybern.*, 1–11(2019)
12. Meskin, N., Khorasani, K.: Actuator fault detection and isolation for a network of unmanned vehicles. *IEEE Trans. Autom. Control* **54**(4), 835–840 (2009)

13. Zhang, Z., Jaimoukha, M.: On-line fault detection and isolation for linear discrete-time uncertain systems. *Automatica* **50**(2), 513–518 (2014)
14. Tang, W., Wang, Z., Shen, Y.: Fault detection and isolation for discrete-time descriptor systems based on  $H_-/L_\infty$  observer and zonotopic residual evaluation. *Int. J. Control.*, 1–12(2018)
15. Jiang, B., Staroswiecki, M., Cocquempot, V.: Fault estimation in nonlinear uncertain systems using robust/sliding-mode observers. *IEE Proc. Control Theory Appl.* **151**(1), 29–37 (2004)
16. Gao, C., Duan, G.: Robust adaptive fault estimation for a class of nonlinear systems subject to multiplicative faults. *Circ. Syst. Sig. Process.* **31**(6), 2035–2046 (2012)
17. Zhu, J., Yang, G., Wang, H., Wang, F.: Fault estimation for a class of nonlinear systems based on intermediate estimator. *IEEE Trans. Autom. Control* **61**(9), 2518–2524 (2016)
18. Yang, J., Zhu, F., Wang, X., Bu, X.: Robust sliding-mode observer-based sensor fault estimation, actuator fault detection and isolation for uncertain nonlinear systems. *Int. J. Control.* **13**(5), 1037–1046 (2015)
19. Yang, G., Huang, S.: Fault estimation for a class of non-linear systems via full-column-rank state variable substitution. *IET Control Theory Appl.* **10**(17), 2260–2270 (2016)
20. Wang, G., Yi, C.: Fault estimation for nonlinear systems by an intermediate estimator with stochastic failure. *Nonlinear Dyn.* **89**(2), 1195–1204 (2017)
21. Ng, K., Tan, C., Edwards, C., Kuang, Y.: New results in robust actuator fault reconstruction for linear uncertain systems using sliding mode observers. *Int. J. Robust Nonlinear Control* **17**(14), 1294–1319 (2007)
22. Liu, C., Patton, R., Zhang, K.: Hierarchical structure-based fault estimation and fault-tolerant control for multi-agent systems. *IEEE Trans. Control Netw. Syst.* **6**(2), 586–597 (2019)
23. Li, X., Zhang, W., Wang, Y.: Simultaneous fault estimation for uncertain Markovian jump systems subjected to actuator degradation. *Int. J. Robust Nonlinear Control* **29**(13), 4435–4453 (2019)
24. Du, D., Xu, S., Cocquempot, V.: Actuator fault estimation for discrete-time switched systems with finite-frequency. *Syst. Control Lett.* **108**, 64–70 (2017)
25. Chen, L., Zhao, Y., Fu, S., Liu, M., Qiu, J.: Fault estimation observer design for descriptor switched systems with actuator and sensor failures. *IEEE Trans. Circ. Syst. I: Reg. Pap.* **66**(2), 810–819 (2019)



# Deep Convolutional Neural Network for Real and Fake Face Discrimination

Yuanyuan Li<sup>1</sup>, Jun Meng<sup>1</sup>, Yaqin Luo<sup>1</sup>, Xinghua Huang<sup>2(✉)</sup>, Guanqiu Qi<sup>3</sup>, and Zhiqin Zhu<sup>1</sup>

<sup>1</sup> Chongqing University of Posts and Telecommunications, Chongqing 400065, China  
 [{liyy,zhuzq}@cqupt.edu.cn](mailto:{liyy,zhuzq}@cqupt.edu.cn), [mikuzip01@gmail.com](mailto:mikuzip01@gmail.com), [Luoyq00@outlook.com](mailto:Luoyq00@outlook.com)

<sup>2</sup> Key Laboratory of Complex System Safety and Control, Ministry of Education, Chongqing University, Chongqing 400044, China  
[huangxh1980@126.com](mailto:huangxh1980@126.com)

<sup>3</sup> State University of New York at Buffalo State, Buffalo, NY 14222, USA  
[qig@buffalostate.edu](mailto:qig@buffalostate.edu)

**Abstract.** With the progress of society, and the rapid development of science and technology, information identification security issues have become more important. With the use of technologies such as Generative Adversarial Networks (GAN) to generate the generalization of faces, in order to make the face identification system more secure, it is necessary to detect the fake face. With the maturity of artificial intelligence technology, face identification technology is widely used in various fields of real life, especially in the research of identifying computer generated faces. This paper is aimed at discriminating the generation of face problems. An algorithm based on Deep Convolutional Neural Network (DCNN) for neural network architecture for face image style classification is proposed. And provide experimental basis. The results show that the method can obtain satisfactory results, and the average accuracy is above 99.24%.

**Keywords:** DCNN · Face discrimination · GAN · Face identification

## 1 Introduction

Current face identification systems are more common than ever before. From face identification in smartphones to face identification in large-scale surveillance, the application of face identification systems is ubiquitous. But face identification systems are also easily fooled by fake and unreal faces. With the use of technologies such as Generative Adversarial Networks (GAN) to generate faces more convenient, in order to make the face identification system more secure, we need to detect such forged faces. For face identification, there are usually four steps involved: 1) Face image acquisition and detection; 2) Face image preprocessing; 3) Face image feature extraction; 4) Face image matching and identification. Extract the facial feature data in the image, and match the extracted data with the facial feature data in the database. With the advent of deep learning, face

identification algorithms based on convolutional neural networks (CNN) have also emerged. At present, there are few studies on face style discrimination based on face identification technology.

Generative Adversarial Networks (GAN) was first proposed by Ian Goodfellow in 2014 [1]. With its superior performance, it has quickly become a research hotspot in less than two years, with theoretical algorithms and applications. Rich results. Ian Goodfellow et al. [1] proposed a framework for generating models through confrontational process estimation. GAN is composed of the generation module and the discrimination module. The main function of the generation module is to generate false data through simulated learning of real data; the main role of the discrimination model is whether the given data is real data or false data generated by the generator. The goal of GAN is to train one to generate false data by simulating real data, and make the discriminator unable to distinguish whether the data is real or false. Based on the framework, Radford et al. [2] introduced a deep convolution generation confrontation network (DCGANs), This network has a clear structural constraint, and there is a strong credibility for unsupervised learning. Karras et al. [3] proposed a new GAN training method. By simultaneously increasing the power of the generator and discriminator. Finally, a high quality generated image is obtained. Karras et al. [4] created a generator architecture for the GAN framework. The framework learns to generate advanced attributes and random variables in images. The emergence of GAN provides people with new ideas, it provides a new method and framework for computer vision. Compared with the traditional machine learning algorithm, GAN adopts the idea of confrontation training, and is more powerful in feature learning and representation. The application of this technology in face generation, it also evolved from the early fuzzy phase to the current level of reality.

Face identification technology due to its wide range of applications, in recent years, it has received extensive attention, with peoples high attention to security and the rapid development of the network, the need to identify real users through different authentication mechanisms is growing rapidly. With the generalization of fake faces generated by technologies such as GAN, in order to protect the identity of personnel, it is very important to detect fake faces. Face identification is a traditional research work, with face information as the main feature for identity determination, and the technology has reached a fairly mature stage. Face identification algorithm research mainly focuses on principal component analysis [5] (PCA), local binary mode [6] (LBP) algorithm and deep learning algorithm [7,8]. Sushama et al. [9] proposed a face identification method using explicit rotation local binary mode (DRLBP) and scale invariant feature transform (SIFT) feature extraction. The face features are extracted using SIFT, and face detection is performed through a back propagation network (BPN). Zhao et al. [10] replaced the gray value of the image pixel with the median value of its neighborhood sampling value, and then extracted the feature value of the image through the sub-blocks, using statistical histogram to establish a method for recognizing the face MLBPH feature identity dictionary. Wang et al. [11] proposed a method for

face identification in real-world surveillance video through deep learning. The data set is built by the process of face detection, tracking and graph clustering. A method of face identification in real-world surveillance video by fine-tuning a depth model using a new data set. Zhang et al. [12] proposed a face identification model based on LBP features and CNN. The model processes the LBP image to generate the LBP feature map of the image, and then uses the LBP feature map to train the CNN. In the test identification image, the extracted image LBP feature map is used to enter the CNN classifier for identification. Compared with the classical face identification algorithm, the algorithm based on convolutional neural network [13, 14] (CNN) has the advantages of strong robustness and high identification rate.

In recent years, the use of generated false face images to replace real image faces is common. In order to protect the privacy of data, we need to detect forged faces. This paper cuts into the face style and performing face discriminate based on face identification. It mainly distinguishes between animated false faces and real faces obtained by face generation technology. Based on DCNN, a neural network architecture that can be used for face image style classification is proposed. Use Adam to optimize network for CNN, and achieve accurate identification of the face style. The main contributions of the paper are as follows:

- A deep convolutional neural architecture for facial feature extraction is proposed, which reduces the computational complexity as much as possible while ensuring discrimination accuracy.
- Use face detection technology to process data sets. Guarantee the consistency of the input data of the neural network, make training and testing of neural networks more effective.

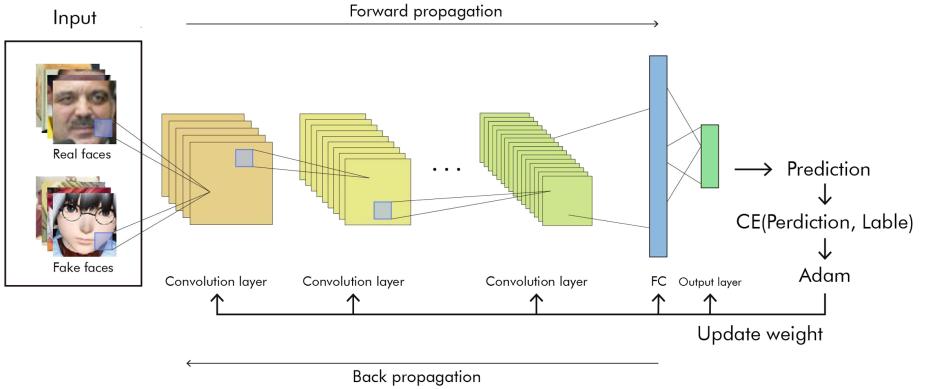
The rest of the paper is organized as follows. Sect. 2 details the algorithm used for face discrimination; Sect. 3 is the experimental part and we summarize the paper in Sect. 4.

## 2 Method

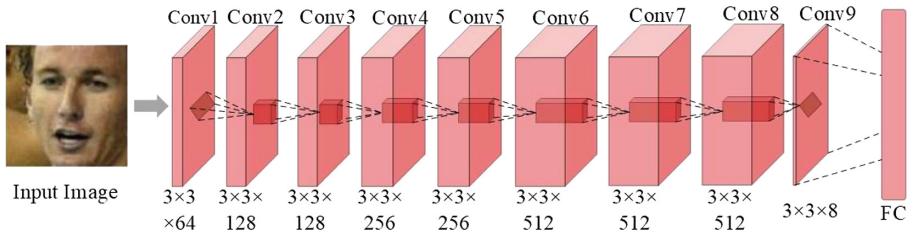
In order to solve the problem, we proposed a CNN network architecture for image style classification. The convolutional neural network is used to process the input image, which can effectively extract the potential information of the image features. The deep convolution structure allows the network to extract abstract image semantic information from the image. In this section, we present a well-designed image style classification network. Then we explained the specific optimization method.

### 2.1 Proposed Fake Face Discrimination Framework

Our network structure is shown in Fig. 1. Our face discrimination network contains a total of 10 layers. Our network accepts three-channel images whose face positions have been cropped, with a resolution of 96\*96. These images are sent to



**Fig. 1.** The face data we input is sent to a fully convolved network for feature extraction, and then the extracted features are flattened into the fully connected layer for classification output. We use CE to calculate the loss of the entire network, and then use the Adam method to optimize the network. The specific architecture of our network is shown in Fig. 2, Our network consists of 9 convolutional layers and one FC layer. The input image of our network is 96\*96\*3.



**Fig. 2.** The specific architecture of our network.

the first layer network for processing. In the first layer of the network, we used 64 convolution kernels of size  $3 \times 3$  to perform preliminary convolution on the image, and then used the modified linear function LeakyReLU to activate the convolution results, and obtained 64 sizes of  $96 \times 96$  rough feature map. We perform these Con-Bn-LeakyReLU operations on these feature maps 8 more times, using the convolution operation multiple times can extract deep-level semantic information in the input image. Among them, 5 convolution kernels with a moving step of 2 are used to replace the MaxPooling structure, and then the output features are down-sampled. We use BN to alleviate the problem of gradient dispersion during reverse conduction. Finally, an 8-channel feature map with a resolution of  $3 \times 3$  is obtained. We flatten these final feature maps and send them into a fully-connected network with  $8 \times 3 \times 3$  neurons. We use this network to classify these features and output the results expressed in One Hot mode. If our network is in the training state, we use cross-entropy to measure the error between the

network output and its true label, and then use the Adam method to perform gradient descent on the entire network, and update the weight of each neuron in the network.

## 2.2 Learning Algorithm

We use the Adam [15] algorithm to optimize our network, and its specific function is expressed as follows:

$$\begin{cases} m_i = \alpha_1 m_{i-1} + (1 - \alpha_1) g_i \\ v_i = \alpha_2 v_{i-1} + (1 - \alpha_2) g_i^2 \\ \hat{m}_i = \frac{m_i}{1 - \alpha_1^i}, \hat{v}_i = \frac{v_i}{1 - \alpha_2^i} \\ W_{i+1} = W_i - \frac{\eta}{\sqrt{\hat{v}_i} + \varepsilon} \hat{m}_i \end{cases} \quad (1)$$

where  $m_i$  and  $v_i$  present first-order momentum and second-order momentum terms, respectively.  $\alpha_1, \alpha_2$  are the power values, and  $\hat{m}_i, \hat{v}_i$  are the correction values.  $W_i$  represents the parameters of the  $i-th$  iteration model. Where  $g_i$  is defined as follows:

$$g_i = \Delta J(W_i) \quad (2)$$

It represents the gradient of the cost function with respect to  $W$  at the  $i-th$  iteration.

We are dealing with a binary classification problem, so we use cross entropy (CE) as the cost function of the whole network, which is defined as follows:

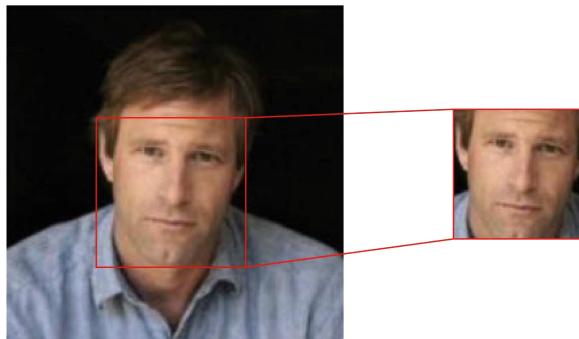
$$J(w) = -\frac{1}{N} \sum_{n=1}^N [y_n \log \hat{y}_n + (1 - y_n) \log (1 - \hat{y}_n)] \quad (3)$$

where  $y$  is ground truth,  $\hat{y}$  is the prediction, and  $N$  is the size of the batch. In our framework,  $\eta = 0.001$ ,  $\alpha_1 = 0.9$ ,  $\alpha_2 = 0.999$ ,  $\varepsilon = 1e-8$ .

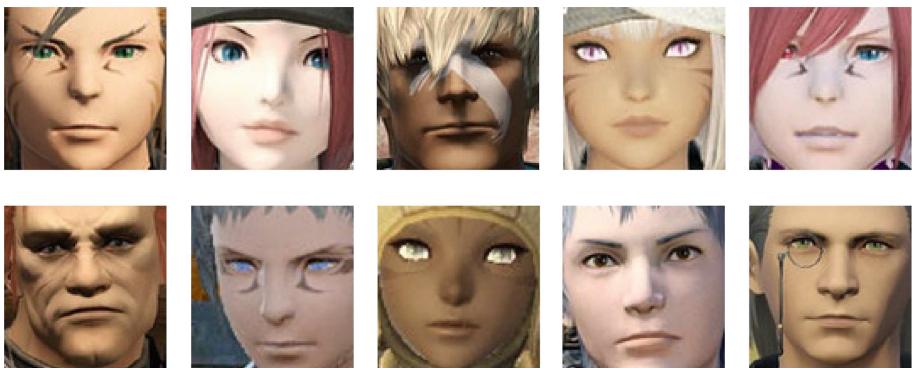
## 3 Experimental

### 3.1 Experimental Image Data Set

In the experimental part, for anime style face images, we use the web crawler to download in batches on the <http://konachan.net/> website. For live-action images, we use the data set from <http://vis-www.cs.umass.edu/lfw/>. In our experiments, we first cut out all the faces in the dataset using the open source faced identification library Dlib (see Fig. 3), and then resize to 96\*96. After the processing is completed, there are a total of 30,000 face images (as shown in Fig. 4 and Fig. 5). We divide the training set, validation set, and test set by a ratio of 90:9:1, we use a batch size of 64 and train the proposed CNN for 15 epochs. All experiments were performed using Pytorch on an Intel I7 6700K CPU and a 56.00GB RAM computer.



**Fig. 3.** Face cropping



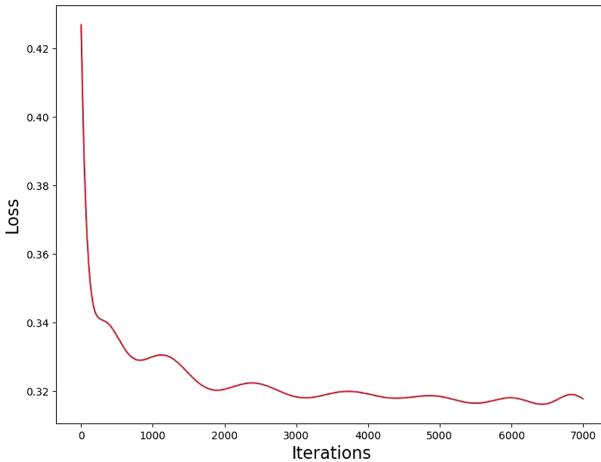
**Fig. 4.** Anime style face image

### 3.2 Experimental Results

In this section, our goal is to identify whether a given face image is real or generated. As shown in Fig. 6, the network gradually converges during the training process, and the loss tends to be stable after about 3000 iterations. On the verification set, we can find that the accuracy is the highest in the fifth epoch, and the performance is degraded after the 15th. This should be due to the over-fitting of the neural network, which leads to the deterioration of the network generalization ability. In order to obtain more convincing results, we evaluated the trained model. For the test set, our AP reached 99.24%.



**Fig. 5.** Live-action images



**Fig. 6.** Loss on the training set

#### 4 Conclusions

With the maturity of artificial intelligence technology, it is becoming more common to generate fake faces. It is important to detect fake faces. Face identification technology has developed rapidly as a safe and reliable biometric technology. Face identification technology is now widely used in various fields of real life, although many previous methods have solved the problem of identifying false and real faces. However, there are few studies on the discrimination between anime-style face images and real faces. In this paper, we propose a neural network architecture based on DCNN that can be used for face image style classification. The Adam method is used to optimize the CNN network, and reduce the computational complexity in the case of high-precision identification. Face detection technology is used to process the data set, which ensures the consistency of the

input data of the neural network, and makes the training and testing of the neural network more effective. And finally realized accurate identification of face style.

**Acknowledgment.** This research was funded by the National Natural Science Foundation of China under Grants 61906026, 61803061 and 51705056; the Common Key Technology Innovation Special of Key Industries of Chongqing Science and Technology Commission under Grant Nos. cstc2017zdcy-zdyfx0067, cstc2017zdcy-zdyfx0055, and cstc2018jszx-cyzd0634; the Artificial Intelligence Technology Innovation Significant Theme Special Project of Chongqing Science and Technology Commission under Grant No. cstc2017rgzn-zdyfx0014 and No. cstc2017rgzn-zdyfx0035.

## References

1. Goodfellow, I., Pougetabadi, J., Mirza, M., et al.: Generative Adversarial Nets. In: Neural Information Processing Systems, pp. 2672–2680 (2014)
2. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint [arXiv:1511.06434](https://arxiv.org/abs/1511.06434) (2015)
3. Karras, T., Aila, T., Laine, S., et al.: Progressive growing of GANs for improved quality, stability, and variation, arXiv preprint [arXiv:1710.10196](https://arxiv.org/abs/1710.10196) (2017)
4. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019). <https://doi.org/10.1109/CVPR.2019.00453>
5. Kang, J., Lin, X., Yang, G.: Research of Multi-scale PCA Algorithm for Face Recognition (2015). <https://doi.org/10.1049/cp.2015.0245>
6. Zhou, X., Sanchez, S.A., Kuijper, A.: 3D face recognition with local binary patterns. In: Sixth International Conference on Intelligent Information Hiding & Multimedia Signal Processing. IEEE Computer Society, pp. 329–332 (2010). <https://doi.org/10.1109/IIHMSP.2010.87>
7. Zhang, X., Peng, M., Chen, T.: Face recognition from near-infrared images with convolutional neural network. In: 2016 8th International Conference on Wireless Communications & Signal Processing (WCSP), pp. 1–5. IEEE (2016). <https://doi.org/10.1109/WCSP.2016.7752592>
8. Wu, Z., Peng, M., and Chen, T.: Thermal face recognition using convolutional neural network. In: International Conference on Optoelectronics and Image Processing (ICOIP), pp. 6–9. IEEE (2016). <https://doi.org/10.1109/OPTIP.2016.7528489>
9. Sushama, M., Rajinikanth, E.: Face recognition using DRLBP and SIFT feature extraction. In: 2018 International Conference on Communication and Signal Processing (ICCSP), pp. 994–999. IEEE (2018). <https://doi.org/10.1109/ICCSP.2018.8524427>
10. Zhao, X.M., Wei, C.B.: A real-time face recognition system based on the improved LBPH algorithm. In: 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), pp. 72–76. IEEE (2017). <https://doi.org/10.1109/SIPROCESS.2017.8124508>
11. Wang, Y., Bao, T., Ding C., et al.: Face recognition in real-world surveillance videos with deep learning method. In: 2017 2nd International Conference on Image, Vision and Computing (ICIVC), pp. 239–243. IEEE (2017) <https://doi.org/10.1109/ICIVC.2017.7984553>

12. Zhang, H., Qu, Z., Yuan, L., et al.: A face recognition method based on LBP feature for CNN. In: 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), pp. 544–547. IEEE (2017). <https://doi.org/10.1109/IAEAC.2017.8054074>
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012). <https://doi.org/10.1145/3065386>
14. Schmidhuber, J.: Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015). <https://doi.org/10.1016/j.neunet.2014.09.003>
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)



# Control Design for One-Sided Lipschitz Nonlinear Systems with Actuator Saturation

Lin Yang, Jun Huang<sup>(✉)</sup>, and Haoran Zhang

School of Mechanical and Electrical Engineering, Soochow University,  
Suzhou 215131, Jiangsu, China  
[cauchyhot@163.com](mailto:cauchyhot@163.com)

**Abstract.** In this article, the saturated controller is designed for a type of nonlinear systems satisfying one-sided Lipschitz constraints. The nonlinearity in the saturation element is transformed into a series of convex hull of linear state feedback. The stability of the closed-loop system is analyzed by the method of Lyapunov function, and the parameters of the saturated controller are derived by the constraints of several linear and bilinear matrix inequalities. At last, we provide an example to demonstrate the availability of the presented method.

**Keywords:** Saturated controller · One-sided Lipschitz · Lyapunov function

## 1 Introduction

Since the definition of one-sided Lipschitz (OSL) was introduced to nonlinear systems, the kind of systems have become one of the research hotspots in the field of control. At present, the study on OSL nonlinear systems is mainly divided into two aspects: observation and stabilization. On the observation of OSL nonlinear systems, [1] gave the sufficient condition for the existence of observer, and proved that the derived conditions by using OSL constraints were less conservative than those derived by Lipschitz constraints. [2] designed the observer for OSL nonlinear systems by linear matrix inequality technique and [3] constructed the reduced-order observer for the systems. [4] improved the results of [2] and gave less conservative conditions using S-procedure method. Recently, the interesting works on observer for OSL systems can be found in [5,6]. While, for the stabilization of OSL systems, the stability of OSL systems was analyzed firstly in [7]. After that, many scholars have made remarkable achievements in the stabilization of OSL systems, such as robust control [8], finite-time  $H_\infty$  control [9,10] and so on.

In addition, actuator saturation is a common phenomenon in practical systems [11–13], and it is often caused by the physical limitations of actuator. It is impossible to linearize the nonlinearity in actuator saturation, which makes the

analysis and design of control systems very difficult. The framework of nonlinear controller design for saturated control systems was put forward in [14]. In other words, the nonlinearity in actuator saturation was treated as a series of convex hull forms of linear control [15, 16]. The saturated controller for linear systems was designed in [17, 18]. Whereas, [19] presented the controller design method for Lipschitz nonlinear systems with actuator saturation. However, there are few works on OSL systems with actuator saturation.

Based on the above discussion, a kind of saturated nonlinear systems satisfying OSL are investigated in the paper. The content of this paper includes the following: Sect. 2 introduces the system description and preliminaries. In Sect. 3, a saturated controller is designed for OSL systems and the closed-loop systems are asymptotically stable under the given conditions, then the maximal estimation of domain of attraction (DA) is also presented. In Sect. 4, a simulation example is provided to verify the effectiveness of the presented controller.

**Notations:**  $R^n$  represents the set of real  $n$ -vectors and  $R^{r \times s}$  represents the set of  $r \times s$  real matrices.  $P > (\geq)0$  or  $P < (\leq)0$  indicates that the matrix  $P$  is positive (semi-positive) definite or negative (semi-negative) definite.  $\|\cdot\|$  is the norm and  $\langle \cdot, \cdot \rangle$  is the inner product.  $Co\{\Xi\}$  means the convex hull of polytope  $\Xi$  and  $sign(u)$  is the symbol function with respect to  $u$ .  $*$  is an ellipsis for the term introduced symmetrically.

## 2 System Description and Preliminaries

Consider the saturated nonlinear system described by

$$\dot{x}(t) = Ax(t) + B\varphi(x(t)) + DN_{sat}(u(t)), \quad (1)$$

where  $x(t) \in R^n$  is the state and  $N_{sat}(u(t)) \in R^s$  is the saturated input.  $\varphi(x(t)) \in R^m$  is the nonlinear function satisfying OSL and quadratic inner-boundedness.  $A \in R^{n \times n}$ ,  $B \in R^{n \times m}$  and  $D \in R^{n \times s}$  are given matrices. When the variable  $t$  is omitted, (1) is simplified as

$$\dot{x} = Ax + B\varphi(x) + DN_{sat}(u). \quad (2)$$

The actuator saturation function  $N_{sat}(u)$  is defined by

$$N_{sat}(u) = [N_{sat}(u_1), N_{sat}(u_2), \dots, N_{sat}(u_s)],$$

where

$$N_{sat}(u_j) = sign(u_j) \min\{1, |u_j|\}.$$

Let  $u = Fx$  be the linear state feedback for the system (2), one can get

$$\dot{x} = Ax + B\varphi(x) + DN_{sat}(Fx). \quad (3)$$

**Definition 1.**  $\varphi(x)$  is an OSL nonlinear function with OSL constant  $\gamma$ , if the following inequality holds for any  $x_a, x_b \in R^n$

$$\langle \varphi(x_a) - \varphi(x_b), x_a - x_b \rangle \leq \gamma \|x_a - x_b\|^2. \quad (4)$$

**Definition 2.**  $\varphi(x)$  is quadratically inner-bounded if there exist constants  $\alpha$  and  $\beta$  such that the following inequality holds for any  $x_a, x_b \in R^n$

$$\|\varphi(x_a) - \varphi(x_b)\|^2 \leq \alpha \|x_a - x_b\|^2 + \beta \langle \varphi(x_a) - \varphi(x_b), x_a - x_b \rangle. \quad (5)$$

In the sequel, the nonlinear function  $\varphi(x)$  is assumed to satisfy OSL and quadratic inner-boundedness.

**Definition 3.** Denote the solution set of the system (3) as  $M(x, t)$ . It is assumed that the solution  $x = 0$  is asymptotically stable, then the original DA is

$$\Phi = \{x : \lim_{t \rightarrow \infty} M(x, t)\} = 0.$$

A set  $N \subset R^n$  is invariant if  $M(x, t) \subset N$  for any  $x \in N$ .

**Lemma 1** [15]. Let  $F, G \in R^{s \times n}$  be the state feedback matrices. If  $x \in \Gamma(G)$  and  $\Gamma(G) = \{x \in R^n : \|Gx\|_\infty \leq 1\}$ , we have

$$N_{sat}(u) \in co\{E_i Fx + E_i^- Gx : i \in [1, 2^s]\},$$

where  $E_i, E_i^- \in E$  and  $E_i^- = I - E_i$ .  $E$  is a group of diagonal matrices with diagonal elements 0 or 1, and contains  $2^s$  elements, which are labeled as  $E_i$  for  $i \in [1, 2^s]$ .

**Lemma 2** [20]. Denote that quadratic Lyapunov function  $V(x) = x^T Px$ , where the matrix  $P \in R^{n \times n} > 0$ . The  $\lambda$ -level set of  $V(x)$  is defined by  $L_V(\lambda) = \{x : x \in \varepsilon(P, \lambda)\}$ , where  $\varepsilon(P, \lambda) = \{x : x^T Px \leq \lambda\}$ . If  $V(x) > 0$  and  $\dot{V}(x) < 0$  for any  $x \in \varepsilon(P, \lambda)$ , the system (2) is asymptotically stable and the set  $L_V(\lambda)$  is an invariant set inside DA.

**Lemma 3** [21]. Let  $G \in R^{s \times n}$ ,  $Q \in R^{n \times n} > 0$  be constant matrices, then  $\varepsilon(Q^{-1}, \lambda) \subset \Gamma(G)$  can be represented by

$$\begin{bmatrix} 1 & \lambda n_i \\ \lambda n_i^T & \lambda Q \end{bmatrix} \geq 0,$$

where  $n_i$  is the  $i$ -th row of  $GQ$ .

### 3 Nonlinear Feedback for Robust Stabilization

In this section, we would like to design the feedback law to make the closed-loop system asymptotically stable and maximize the estimation of DA.

**Theorem 1.** Let the nonlinear function  $\varphi(x)$  satisfy the conditions (4)–(5). If there exist matrices  $Q \in R^{n \times n} > 0$ ,  $U \in R^{s \times n}$ ,  $W \in R^{s \times n}$  and  $\lambda > 0$  such that

$$\begin{bmatrix} 1 & \lambda n_i \\ \lambda n_i^T & \lambda Q \end{bmatrix} \geq 0, \quad (6)$$

where  $n_i$  is the  $i$ -th row of  $GQ$ ,

and

$$\Xi = \begin{bmatrix} \Xi_{11} & B + \frac{\beta-1}{2}Q \\ * & -I \end{bmatrix} < 0, \quad (7)$$

where

$$\Xi_{11} = AQ + QA^T + D(E_i U + E_i^- W) + (E_i U + E_i^- W)^T D^T + (\alpha + \gamma)Q^2,$$

with

$$U = FQ, \quad W = GQ.$$

Then, the system (2) is asymptotically stable and the invariant set  $L_V$  is inside DA under the following state feedback law

$$N_{sat}(u) \in co\{E_i Fx + E_i^- Gx : i \in [1, 2^s]\}. \quad (8)$$

**Proof.** Multiplying (7) sides by left and right with the matrix  $\begin{bmatrix} P & 0 \\ * & I \end{bmatrix}$ , it yields

$$\Pi = \begin{bmatrix} \Pi_{11} & PB + \frac{\beta-1}{2}I \\ * & -I \end{bmatrix} < 0, \quad (9)$$

where

$$\Pi_{11} = PA + A^T P + PD(E_i F + E_i^- G) + (E_i F + E_i^- G)^T D^T P + (\alpha + \gamma)I. \quad (10)$$

Let  $V(x) = x^T Px$  be the quadratic Lyapunov candidate function, where  $P = Q^{-1}$ , then

$$\dot{V}(x) = 2x^T P \dot{x}. \quad (11)$$

In view of Lemma 1, the system (2) can be described by the following convex hull

$$\dot{x} \in co\{Ax + B\varphi(x) + DE_i Fx + DE_i^- Gx : i \in [1, 2^s]\}. \quad (12)$$

Substituting (12) into (11) yields

$$\dot{V}(x) \leq \max_{i \in [1, 2^s]} 2x^T P[Ax + B\varphi(x) + DE_i Fx + DE_i^- Gx]. \quad (13)$$

By the conditions (4) and (5), the following expressions hold

$$\gamma x^T x - x^T \varphi(x) \geq 0, \quad (14)$$

$$\alpha x^T x + \beta x^T \varphi(x) - \varphi^T(x) \varphi(x) \geq 0. \quad (15)$$

By using S-procedure, it can be deduced from (14) and (15) that

$$\begin{aligned} \dot{V}(x) &\leq \max_{i \in [1, 2^s]} x^T [PA + A^T P + PD(E_i F + E_i^- G) + (E_i F + E_i^- G)^T D^T P + (\alpha + \gamma)I] x \\ &\quad + x^T (PB + \frac{\beta - 1}{2} I) \varphi(x) + \varphi^T(x) (PB + \frac{\beta - 1}{2} I)^T x - \varphi^T(x) \varphi(x). \end{aligned} \quad (16)$$

Let  $\zeta = [x^T \varphi^T(x)]^T$ , and the inequality (16) becomes

$$\dot{V}(x) \leq \max_{i \in [1, 2^s]} \zeta^T \Pi \zeta, \quad (17)$$

where the matrix  $\Pi < 0$  is defined in (9). Combining (9) with (17), one can get  $\dot{V}(x) < 0$ . According to Lemma 2, the system (2) satisfies the condition of synchronous stability and the invariant set  $L_V(\lambda)$  is inside DA under the state feedback (8).

*Remark 1.* It is obvious that (7) is a bilinear inequality since it contains the bilinear term  $(\rho + \gamma)Q^2$ , and the path-following algorithm [22] can be employed to provide the feasible solutions for it.

In the sequel, the maximal estimation of  $L_V(\lambda)$  will be given. According to the definition of  $L_V(\lambda)$ , it can be seen that the size of the invariant set  $L_V(\lambda)$  is related to  $\varepsilon(P, \lambda)$ . For this purpose, we adopt some results in [16]. Let the set  $X_0 \subset R^n$  be an ellipsoid, i.e.,  $X_0 = \{x : x^T Rx \leq 1\}$ , where  $R > 0$  is a given matrix. Meanwhile, let  $X_0$  also be a convex set and it contains origin. Then, the size of a set  $M \subset R^n$  can be defined by  $X_0$  as

$$\sigma(M) = \sup\{\sigma > 0 : \sigma X_0 \subset M\}. \quad (18)$$

Obviously,  $X_0 \subset M$  when  $\sigma > 1$ . Thus, the maximal DA can be estimated by solving the following optimal problem

$$\begin{aligned} &\sup_{Q > 0, F, G} \sigma, \\ (i) \quad &\sigma X_0 \subset \varepsilon(P, \lambda), \\ (ii) \quad &\Xi < 0, \\ (iii) \quad &\begin{bmatrix} 1 & \lambda n_i \\ \lambda n_i^T & \lambda Q \end{bmatrix} \geq 0. \end{aligned} \quad (19)$$

By using the results in [23], (i) is equivalent to

$$\begin{bmatrix} \delta R & I \\ I & \lambda Q \end{bmatrix} \geq 0,$$

where  $\delta = \sigma^{-2}$ . Thus, (19) can be translated to

$$\begin{aligned} & \inf_{Q>0,F,G} \delta, \\ (i) \quad & \begin{bmatrix} \delta R & I \\ I & \lambda Q \end{bmatrix} \geq 0, \\ (ii) \quad & \Xi < 0, \\ (iii) \quad & \begin{bmatrix} 1 & \lambda n_i \\ \lambda n_i^T & \lambda Q \end{bmatrix} \geq 0. \end{aligned} \tag{20}$$

## 4 Simulation

Consider the system (2) with

$$A = \begin{bmatrix} 2 & 29 \\ -4 & 0.7 \end{bmatrix}, \quad B = \begin{bmatrix} -0.52 & 0.1 \\ 0.2 & 0.1 \end{bmatrix},$$

$$D = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \varphi(x) = \begin{bmatrix} \sin(x_1) \\ 0 \end{bmatrix}.$$

From the expression of  $\varphi(x)$ , it is easy to determine that  $\alpha = 1$ ,  $\beta = 0$  and  $\gamma = 1$ . Here we choose

$$R = \begin{bmatrix} 2.065 & -0.326 \\ -0.326 & 2.065 \end{bmatrix}, \quad \lambda = 0.5.$$

Then, by solving optimization problem (20), we can obtain that

$$Q = \begin{bmatrix} 0.9747 & -0.0601 \\ -0.0601 & 0.2141 \end{bmatrix}, \quad U = [-1.4032 \ -0.7489],$$

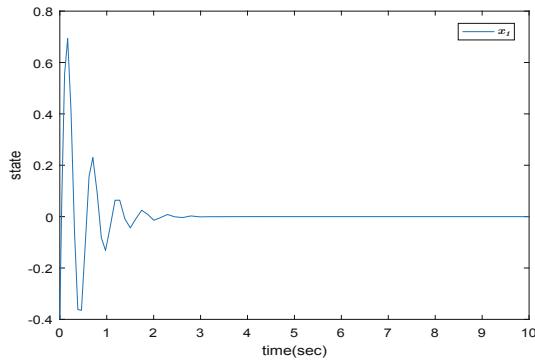
$$W = [-0.8383 \ -0.4664], \quad \delta^* = 4.9.$$

Since  $P = Q^{-1}$ ,  $U = FQ$  and  $W = GQ$ ,  $P$ ,  $F$  and  $G$  can be obtained

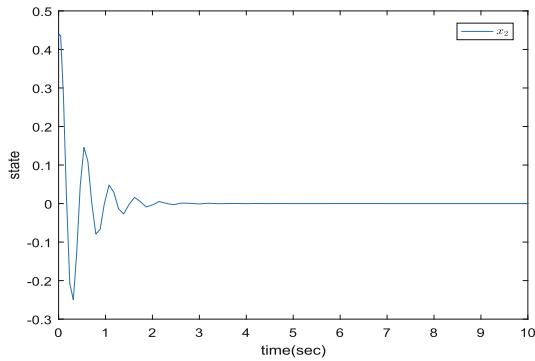
$$P = \begin{bmatrix} 1.0440 & 0.2931 \\ 0.2931 & 4.7530 \end{bmatrix}, \quad F = [-1.6845 \ -3.9713],$$

$$G = [-1.012 \ -2.463].$$

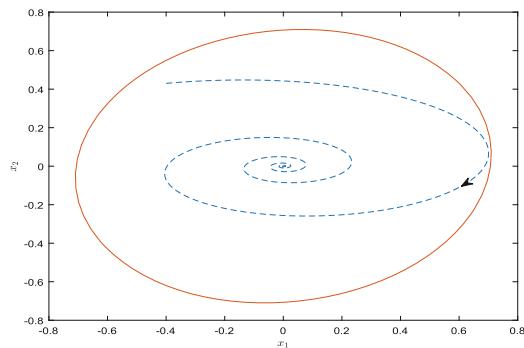
Now we choose the initial state  $x_0 = [-0.4 \ 0.43]^T \in \varepsilon(P, \lambda)$  and complete the simulation by Matlab. Figure1-2 display the state trajectory of the saturated nonlinear system. The simulation results show that the state trajectories of  $x_1$  and  $x_2$  are stable gradually after 3 s. According to the optimization solution  $\delta^* = 4.9$ , we can get  $\alpha^* = 0.4518$ . Next, we simulate the area  $\alpha^* X_0 = \alpha^* \{x | x^T Rx \leq 1\}$  to obtain the DA for the system (2) under the feedback law (8). Figure3 shows that the state trajectory remain in the maximal estimation of DA under the initial condition  $x_0$ .



**Fig. 1.** Response of the system state  $x_1$  under the state feedback (8).



**Fig. 2.** Response of the system state  $x_2$  under the state feedback (8).



**Fig. 3.** The estimation of DA and state responses under the initial state  $x_0 \in \varepsilon(P, \lambda)$  for  $\lambda = 0.5$ .

## 5 Conclusion

In this article, we design the controller for a class of saturated nonlinear systems satisfying OSL. Firstly, the nonlinearity in the saturation element is treated as a series of convex hull of linear state feedback. Then, the existence conditions of the saturated controller are obtained by the constraints of linear and bilinear matrix inequalities. Specially, the maximal estimation of DA is also given. The validation of the presented method is verified by an example.

## References

1. Hu, G.: Observers for one-sided Lipschitz nonlinear systems. *IMA J. Math. Control Inf.* **23**(4), 395–401 (2006)
2. Abbaszadeh, M., Marquez, H.: Nonlinear observer design for one-sided Lipschitz systems. In: Proceedings of the American Control Conference (2009)
3. Xu, M., Hu, G., Zhao, Y.: Reduced-order observer design for one-sided Lipschitz nonlinear systems. *IMA J. Math. Control Inf.* **26**(3), 299–317 (2009)
4. Zhang, W., Su, H., Zhu, F., Yue, D.: A note on observers for discrete-time Lipschitz nonlinear systems. *IEEE Trans. Circ. Syst. II Express Briefs* **59**(2), 123–127 (2012)
5. Zhang, W., Su, H., Zhu, F., Bhattacharyya, S.: Improved exponential observer design for one-sided Lipschitz nonlinear systems. *Int. J. Robust Nonlinear Control* **26**(18), 3958–3973 (2016)
6. Yang, Y., Lin, C., Chen, B., Zhao, X.:  $H_\infty$  observer design for uncertain one-sided Lipschitz nonlinear systems with time-varying delay. *Appl. Math. Comput.* **375**(11), 1250–1266 (2020)
7. Donchev, T., Farkhi, E.: Stability and Euler approximation of one-sided Lipschitz differential inclusions. *SIAM J. Control Optim.* **36**(2), 780–796 (1998)
8. Sad, W., Sellami, A., Garcia, G.: Robust stabilization of one-sided Lipschitz nonlinear systems via adaptive sliding mode control. *J. Vibr. Control* **26**(7), 399–412 (2020)
9. Song, J., He, S.: Finite-time  $H_\infty$  control for quasi-one-sided Lipschitz nonlinear systems. *Neurocomputing* **149**, 1433–1439 (2015)
10. Song, J., He, S.: Robust finite-time  $H_\infty$  control for one-sided Lipschitz nonlinear systems via state feedback and output feedback. *Neurocomputing* **352**(8), 3250–3266 (2015)
11. Gao, W., Selmic, P.: Neural network control of a class of nonlinear systems with actuator saturation. *IEEE Trans. Neural Networks* **17**(1), 147–156 (2006)
12. Saifia, D., Chadli, M., Labiod, S.:  $H_\infty$  control of multiple model subject to actuator saturation: application to quarter car suspension system. *Analog Integr. Cir. Sig. Process.* **69**(1), 81–90 (2011)
13. Maddela, C., Subudhi, B.: Delay-dependent supplementary damping controller of TCSC for interconnected power system with time-delays and actuator saturation. *Electr. Pow. Syst. Res.* **164**, 39–46 (2018)
14. Liu, D., Michel, A.: *Dynamical Systems With Saturation Nonlinearities: Analysis and Design*. Springer, Heidelberg (1994)
15. Hu, T., Lin, Z.: *Control Systems with Actuator Saturation: Analysis and Design*, vol. 38, no. 2, pp. 351–359. Springer, New York (2002)
16. Hu, T., Lin, Z., Chen, B.: An analysis and design method for linear systems subject to actuator saturation and disturbance. *Automatica* **38**(2), 351–359 (2002)

17. Lu, L.: Output regulation of a class of switched linear systems with saturated continuous feedback. In: Proceedings of the 30th Chinese Control Conference (2011)
18. Guan, W., Yang, G.: Adaptive fault-tolerant control of linear systems with actuator saturation and  $L_2$ -disturbances. *J. Control Theor. Appl.* **7**(2), 119–126 (2009)
19. Rehan, M., Tufail, M., Ahn, C., Chadli, M.: Stabilization of locally Lipschitz nonlinear systems under input saturation and quantisation. *IET Control Theor. Appl.* **11**(9), 1459–1466 (2017)
20. Shevitz, D., Paden, B.: Lyapunov stability theory of nonsmooth systems. *IEEE Trans. Autom. Control* **39**(9), 1910–1914 (1944)
21. Hu, T., Lin, Z.: Composite quadratic Lyapunov functions for constrained control systems. *IEEE Trans. Autom. Control* **48**(3), 440–450 (2003)
22. Hassibi, A., How, J., Boyd, S.: A path-following method for solving BMI problems in control. In: Proceedings of the 1999 American Control Conference (1999)
23. Boyd, S., Ghaoui, L., Feron, E., Balakrishnan, V.: Linear matrix inequalities in system and control theory. *SIAM J. Control Optim.* **15** (1994)



# Stochastic Stability of Itô Stochastic Systems with Semi-Markov Jump

Min Zhang, Jun Huang<sup>(✉)</sup>, and Haoran Zhang

School of Mechanical and Electrical Engineering, Soochow University, Suzhou 215131, China  
cauchyhot@163.com

**Abstract.** The stability analysis of Itô stochastic systems with semi-Markov jump elements is investigated in this paper. Based on the boundness of transition rates which vary with respect to sojourn time in the semi-Markov process, the sufficient condition is presented in the pattern of linear matrix inequalities. The validity of conclusions is proved by an appropriate numerical example.

**Keywords:** Itô stochastic system · Semi-Markov process · Stochastic stability analysis

## 1 Introduction

In many actual control engineering systems, dynamic systems are not deterministic systems. Some objective or subjective factors can cause the system to change randomly, such as component failure, unexpected environmental changes, sudden noise interference, etc. In order to facilitate the study of this class of uncertain systems, Markov jump systems were employed to describe them. Researchers have begun to study Markov jump systems since the 1960s, including systems with determined transition rates (TRs) [1–3] and systems with uncertain transition rates [4,5]. In order to expand the application scope, the concept of the semi-Markov process (SMP) was put forward. It can represent more general stochastic jump systems, i.e. semi-Markov jump systems (SMJSs). There are also many works on SMJSs. The stability analysis of SMJSs whose sojourn time (ST) follows phase-type distribution was discussed in [6]. After that, [7–9] presented the stability analysis method of SMJSs when ST obeys to Weibull distribution. Recently, the sufficient conditions of stability for SMJSs without constraints were derived in [10]. It extends the scope of application, but the sufficient conditions in [10] cannot be expressed by linear matrix inequalities (LMIs).

On the other side, stochastic disturbance often can not be ignored in many practical systems. When the noises are related to Brownian motion, stochastic systems can be described by Itô stochastic different equations. A lot of works on the stability analysis of Itô stochastic systems with Markov jump have been done for many years, such as mean-square stability [11],  $p$ th moment stability [12], exponential stability in mean square [13], and finite-time stability [14], etc. However, there are few works on the stability of Itô stochastic systems with semi-Markov jump (SMJ).

Encouraged by the above investigations, this paper concentrates on the stochastic stability analysis of Itô stochastic systems with SMJ. The main contribution of this

paper is extending the application scope of Itô stochastic systems. Itô stochastic systems with Markov jump is promoted to Itô stochastic systems with semi-Markov jump. The rest of this paper includes: Sect. 2 lists problem description and preparations in detail and the sufficient conditions of stochastic stability are represented in Sect. 3. In Sect. 4, one simulation example is used to verify the validity of the proposed conclusions.

### Notations

- $Pr\{\alpha\}$ : the probability of the stochastic event  $\alpha$ ;
- $\mathcal{E}\{\mu\}$ : the mathematical expectation of the stochastic variable  $\mu$ ;
- $tr(A)$ : the sum of diagonal elements of matrix  $A$ ;
- $\lambda_{max}(\cdot)$ : the maximum eigenvalue of the argument;
- $*$ : an ellipsis for the term introduced symmetrically.

## 2 Problem Description and Preparations

Let  $\gamma(t) = \{1, 2, \dots, N\}$  ( $t \geq 0$ ) be a right-continuous SMP, and  $\gamma(t)$  is governed by

$$\mathcal{P}\{\gamma(t + \Delta) = \beta | \gamma(t) = \alpha\} = \begin{cases} \lambda_{\alpha\beta}(\hbar)\Delta + o(\Delta), & \alpha \neq \beta, \\ 1 + \lambda_{\alpha\alpha}(\hbar)\Delta + o(\Delta), & \text{others}, \end{cases} \quad (1)$$

where the ST  $\hbar > 0$ , and  $\lambda_{\alpha\beta}(\hbar)$  is shorthand for the TR when SMP jumps from mode  $\alpha$  to different mode  $\beta$ , but  $\lambda_{\alpha\alpha}(\hbar) = -\sum_{\beta=1, \beta \neq \alpha}^N \lambda_{\alpha\beta}(\hbar)$ . It should also be noted that  $\lim_{\Delta \rightarrow 0} \frac{o(\Delta)}{\Delta} = 0$ .

The research object of this paper is the linear Itô stochastic system with SMJ as shown in (2),

$$\begin{cases} dx(t) = A(\gamma(t))x(t)dt + E(\gamma(t))x(t)dw, \\ y(t) = C(\gamma(t))x(t), \\ x_0 = x(0), y_0 = \gamma(0), \end{cases} \quad (2)$$

where state  $x(t) \in \mathbb{R}^n$ , and measurable output  $y(t) \in \mathbb{R}^r$ .  $\gamma(t)$  is a continuous-time SMP defined above.  $w(t)$  is a 1-dimensional Brownian motion, and the SMP  $\gamma(t)$  is independent of  $w(t)$ . For any mode  $\gamma(t) = \alpha$ ,  $\alpha \in S$ ,  $A(\alpha)$ ,  $E(\alpha)$  and  $C(\alpha)$  are given matrices with compatible dimensions. For simplicity,  $A(\alpha)$ ,  $E(\alpha)$  and  $C(\alpha)$  are denoted by  $A_\alpha$ ,  $E_\alpha$  and  $C_\alpha$  respectively. In addition, if necessary, we will omit the variable  $t$  in the following expression.

*Remark 1.* The ST  $\hbar$  of SMP conforms to general distribution, i.e. Weibull distribution, which is different from the Markov process (MP). Defining the TR from the perspective of probability distribution, we can get

$$\lambda(\hbar) = \frac{\phi(\hbar)}{1 - \Phi(\hbar)},$$

where  $\phi(\hbar)$  is the probability density function (PDF) and  $\Phi(\hbar)$  is the cumulative distribution function (CDF).

**Proposition 1.** According to property of SMP, we can get the following propositions,

$$\begin{aligned}\lim_{\Delta \rightarrow 0} \frac{\Phi_\alpha(\hbar + \Delta) - \Phi_\alpha(\hbar)}{(1 - \Phi_\alpha(\hbar))\Delta} &= \lambda_\alpha(\hbar), \\ \lim_{\Delta \rightarrow 0} \frac{\Phi_\alpha(\hbar + \Delta) - \Phi_\alpha(\hbar)}{1 - \Phi_\alpha(\hbar)} &= 0, \\ \lim_{\Delta \rightarrow 0} \frac{1 - \Phi_\alpha(\hbar + \Delta)}{1 - \Phi_\alpha(\hbar)} &= 1,\end{aligned}$$

where  $\Phi_\alpha(\hbar)$  is the CDF of the ST  $\hbar$  when SMP remains in mode  $\alpha$ , and  $\lambda_\alpha(\hbar)$  is the TR of system jumping from  $\alpha$ .

**Definition 1** (see [4]). System (2) is stochastically stable (SS) if inequality (3) holds for all  $x(0) = x_0 \in \Re^n$ , and  $\gamma(0) = \gamma_0 \in S$ :

$$\lim_{t \rightarrow \infty} \mathcal{E} \left[ \int_0^t x^T(s) x(s) ds | (x_0, \gamma_0) \right] < \infty. \quad (3)$$

**Lemma 1** (see [15]). Given matrices of appropriate dimensions  $Q, W, F$ , where  $Q = Q^T < 0$ ,  $F = F^T > 0$ , then  $Q + WF^{-1}W^T < 0$ , if and only if

$$\begin{bmatrix} Q & W \\ * & -F \end{bmatrix} < 0.$$

### 3 Stochastic Stability Analysis

In this section, the conditions which ensure the stochastic stability of the linear Itô stochastic system with SMJ (2) are given.

**Theorem 1.** If there exist matrix  $P_\alpha > 0 \in \Re^{n \times n}$  such that for any  $\alpha, \beta \in S, \hbar > 0$ ,

$$\begin{bmatrix} P_\alpha A_\alpha + A_\alpha^T P_\alpha + \sum_{\beta=1}^N \lambda_{\alpha\beta}(\hbar) P_\beta & E_\alpha^T P_\alpha \\ * & -P_\alpha \end{bmatrix} < 0, \quad (4)$$

then system (2) is SS.

*Proof.* Consider the Lyapunov function  $\mathcal{V}(x(t), \gamma(t)) = x^T(t)P(\gamma(t))x(t)$ , and  $\mathcal{L}\mathcal{V}(x(t), \gamma(t))$  is the infinitesimal generator. It is assumed that the SMP keep  $\gamma(t) = \alpha$  at the time  $t$ , after  $\Delta$  the mode may jump to  $\beta$  or remain in  $\alpha$ . We have

$$\begin{aligned}\mathcal{L}\mathcal{V}(x(t), \gamma(t)) &= \lim_{\Delta \rightarrow 0} \frac{\mathcal{E}[\mathcal{V}(x(t+\Delta), \gamma(t+\Delta))|x(t), \gamma(t)] - \mathcal{V}(x(t), \gamma(t))}{\Delta} \\ &= \lim_{\Delta \rightarrow 0} \frac{1}{\Delta} \left\{ \sum_{\beta \neq \alpha}^N \mathcal{P}\{\gamma(t+\Delta) = \beta | \gamma(t) = \alpha\} \mathcal{V}(x(t+\Delta), \beta) \right. \\ &\quad \left. + \mathcal{P}\{\gamma(t+\Delta) = \alpha | \gamma(t) = \alpha\} \mathcal{V}(x(t+\Delta), \alpha) - \mathcal{V}(x(t), \alpha) \right\} \\ &= \lim_{\Delta \rightarrow 0} \frac{1}{\Delta} \left\{ \sum_{\beta \neq \alpha}^N \frac{q_{\alpha\beta}(\Phi_\alpha(\hbar + \Delta) - \Phi_\alpha(\hbar))}{1 - \Phi_\alpha(\hbar)} \mathcal{V}(x(t+\Delta), \beta) \right. \\ &\quad \left. + \frac{1 - \Phi_\alpha(\hbar + \Delta)}{1 - \Phi_\alpha(\hbar)} \mathcal{V}(x(t+\Delta), \alpha) - \mathcal{V}(x(t), \alpha) \right\},\end{aligned} \quad (5)$$

where  $q_{\alpha\beta}$  is the probability strength when system jumps from mode  $\alpha$  to  $\beta$ . Applying Itô formula [16] to  $\mathcal{V}(x(t + \Delta), \gamma(t))$ , one can get

$$\begin{aligned} & \mathcal{V}(x(t + \Delta), \beta) \\ &= \mathcal{V}(x(t), \beta) + \{\mathcal{V}_x(x(t), \beta)A_\alpha x(t) + \frac{1}{2}\text{tr}[(E_\alpha x(t))^T \mathcal{V}_{xx}(x(t), \beta)(E_\alpha x(t))]\} \cdot \Delta, \\ & \mathcal{V}(x(t + \Delta), \alpha) \\ &= \mathcal{V}(x(t), \alpha) + \{\mathcal{V}_x(x(t), \alpha)A_\alpha x(t) + \frac{1}{2}\text{tr}[(E_\alpha x(t))^T \mathcal{V}_{xx}(x(t), \alpha)(E_\alpha x(t))]\} \cdot \Delta. \end{aligned}$$

Substituting  $\mathcal{V}(x(t + \Delta), \beta)$  and  $\mathcal{V}(x(t + \Delta), \alpha)$  into (5), together with Proposition 1, it yields to

$$\begin{aligned} & \mathcal{L}\mathcal{V}(x(t), \gamma(t)) \\ &= x^T(t)P_\alpha A_\alpha x(t) + x^T(t)A_\alpha^T P_\alpha x(t) + x^T(t)E_\alpha^T P_\alpha E_\alpha x(t) + \sum_{\beta=1}^N \lambda_{\alpha\beta}(\hbar)x^T(t)P_\beta x(t) \\ &= x^T(t)[P_\alpha A_\alpha + A_\alpha^T P_\alpha + E_\alpha^T P_\alpha E_\alpha + \sum_{\beta=1}^N \lambda_{\alpha\beta}(\hbar)P_\beta]x(t) \\ &= x^T(t)Q_\alpha(\hbar)x(t) \\ &\leq \max_{\alpha \in S}\{\lambda_{\max}(Q_\alpha(\hbar))\}x^T(t)x(t). \end{aligned}$$

Based on the Dynkin's formula,

$$\begin{aligned} & \mathcal{E}[\mathcal{V}(x(t), \gamma(t))] - \mathcal{V}(x_0, \gamma_0) \\ &= \mathcal{E}\left[\int_0^t \mathcal{L}\mathcal{V}(x(s), \gamma(s))ds\right](x_0, \gamma_0) \\ &\leq \max_{\alpha \in S}\{\lambda_{\max}(Q_\alpha(\hbar))\}\mathcal{E}\left[\int_0^t x^T(s)x(s)ds\right](x_0, \gamma_0), \end{aligned}$$

then,

$$\begin{aligned} & -\max_{\alpha \in S}\{\lambda_{\max}(Q_\alpha(\hbar))\}\mathcal{E}\left[\int_0^t x^T(s)x(s)ds\right](x_0, \gamma_0) \\ &\leq \mathcal{V}(x_0, \gamma_0) - \mathcal{E}[\mathcal{V}(x(t), \gamma(t))] \\ &\leq \mathcal{V}(x_0, \gamma_0). \end{aligned}$$

According to Lemma 1, (4) is equivalent to

$$Q_\alpha(\hbar) = P_\alpha A_\alpha + A_\alpha^T P_\alpha + E_\alpha^T P_\alpha E_\alpha + \sum_{\beta=1}^N \lambda_{\alpha\beta}(\hbar)P_\beta < 0.$$

Furthermore,  $Q_\alpha(\hbar) < 0$  indicates  $\max_{\alpha \in S}\{\lambda_{\max}(Q_\alpha(\hbar))\} < 0$ , so

$$\mathcal{E}\left[\int_0^t x^T(s)x(s)ds\right](x_0, \gamma_0) \leq -\frac{\mathcal{V}(x_0, \gamma_0)}{\max_{\alpha \in S}\{\lambda_{\max}(Q_\alpha(\hbar))\}} < \infty.$$

According to Definition 1, it is obvious that system (2) is SS.

*Remark 2.* The TR  $\lambda_{\alpha\beta}(\hbar)$  is time-varying with regard to the ST  $\hbar$ , which makes it difficult to solve inequality (4). We can take advantage of the boundary of the TR to solve this problem.

**Theorem 2.** If there exist matrix  $P_\alpha > 0 \in \Re^{n \times n}$  such that for any  $\alpha, \beta \in S, \hbar > 0$ ,

$$\begin{aligned} & \begin{bmatrix} P_\alpha A_\alpha + A_\alpha^T P_\alpha + \sum_{\beta=1}^N \bar{\lambda}_{\alpha\beta} P_\beta & E_\alpha^T P_\alpha \\ * & -P_\alpha \end{bmatrix} < 0, \\ & \begin{bmatrix} P_\alpha A_\alpha + A_\alpha^T P_\alpha + \sum_{\beta=1}^N \underline{\lambda}_{\alpha\beta} P_\beta & E_\alpha^T P_\alpha \\ * & -P_\alpha \end{bmatrix} < 0, \end{aligned} \quad (6)$$

where  $\bar{\lambda}_{\alpha\beta}, \underline{\lambda}_{\alpha\beta}$  are the highest and lowest value of TR  $\lambda_{\alpha\beta}(\hbar)$ . Then system (2) is SS.

*Proof.* According to the linear relationship  $\lambda_{\alpha\beta}(\hbar) = k\underline{\lambda}_{\alpha\beta} + (1-k)\bar{\lambda}_{\alpha\beta}$ , where  $0 \leq k \leq 1$ , one can get

$$\begin{aligned} & P_\alpha A_\alpha + A_\alpha^T P_\alpha + k \sum_{\beta=1}^N \underline{\lambda}_{\alpha\beta} P_\beta + (1-k) \sum_{\beta=1}^N \bar{\lambda}_{\alpha\beta} P_\beta \\ & = P_\alpha A_\alpha + A_\alpha^T P_\alpha + \sum_{\beta=1}^N [k\underline{\lambda}_{\alpha\beta} + (1-k)\bar{\lambda}_{\alpha\beta}] P_\beta \\ & = P_\alpha A_\alpha + A_\alpha^T P_\alpha + \sum_{\beta=1}^N \lambda_{\alpha\beta}(\hbar) P_\beta. \end{aligned}$$

In other words, if (6) holds then (4) holds. Hence system (2) is SS.

## 4 Simulation

In this part, the validity of the sufficient stochastic stability condition proposed in Sect. 3 is verified by a simulation example. Consider system (2) with two modes

$$A_1 = \begin{bmatrix} 0.01 & -1.916 \\ 3.12 & -0.832 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -0.88 & -0.427 \\ -0.079 & -0.388 \end{bmatrix},$$

$$E_1 = \begin{bmatrix} 0.025 & -0.431 \\ 0.425 & -0.005 \end{bmatrix}, \quad E_2 = \begin{bmatrix} -0.325 & 0.64 \\ -0.65 & -0.24 \end{bmatrix}.$$

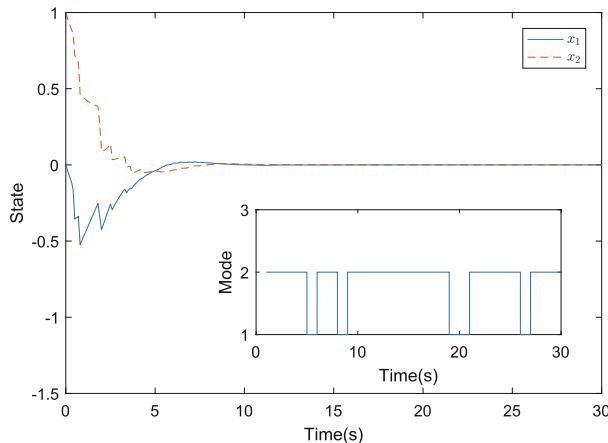
The bounds of time-varying TR  $\lambda_{\alpha\beta}(\hbar)$  are chosen as follows:

$$\underline{\lambda}_{\alpha\beta} = \begin{bmatrix} -0.4 & 0.4 \\ 0.1 & -0.1 \end{bmatrix}, \quad \bar{\lambda}_{\alpha\beta} = \begin{bmatrix} -1 & 1 \\ 0.143 & -0.143 \end{bmatrix}.$$

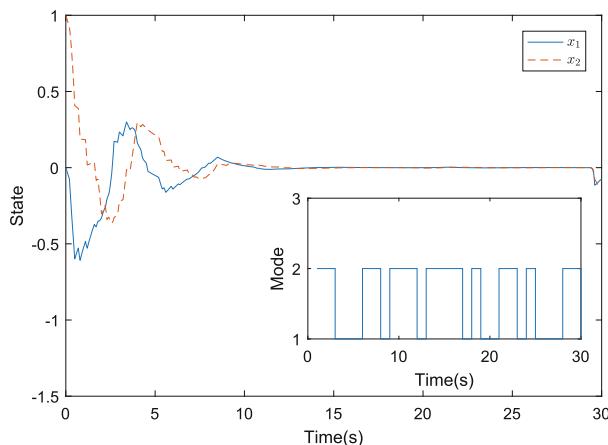
According to Theorem 2, one can obtain

$$P_1 = \begin{bmatrix} 100.8712 & -4.7737 \\ -4.7737 & 64.7559 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 46.6102 & -6.0009 \\ -6.0009 & 75.0531 \end{bmatrix}.$$

Subgraphs in Fig. 1 and Fig. 2 show the stochastic jump signal  $\gamma(t)$ , then the two modes of the system will change stochastically with the signal  $\gamma(t)$ . The trajectories of state variables  $x_1$  and  $x_2$  are depicted in Fig. 1 and Fig. 2, where initial condition is  $x_0 = [0 \ 1]$ . The trajectories are different with different jump signals, which can be seen from Fig. 1 and Fig. 2. But the trajectories of  $x_1$  and  $x_2$  are both SS.



**Fig. 1.** Simulation of jump signal  $\gamma(t)$  and state trajectories of  $x_1$  and  $x_2$



**Fig. 2.** Simulation of jump signal  $\gamma(t)$  and state response of  $x_1$  and  $x_2$

## 5 Conclusion

This paper focuses on the stochastic stability of Itô stochastic systems with SMJ. The stochastic process of jump signal is extended to SMP from MP, which expands the application scope of Itô stochastic systems. The bounds of TRs are employed to derive the sufficient conditions which are formulated by LMIs. Last but not least, an appropriate simulation example is applied to verify the validity of the conclusions proposed in Sect. 3.

**Acknowledgement.** This work is supported by National Natural Science Foundation of China (61403267), China Postdoctoral Science Foundation (2017M611903).

## References

- Shi, Y., Yu, B.: Output feedback stabilization of networked control systems with random delays modeled by Markov chains. *IEEE Trans. Autom. Control* **54**(7), 1668–1674 (2009)
- Zhu, Z., Huang, J.: State estimation for one-sided Lipschitz system with Markovian jump parameters. In: Proceeding of 2019 Chinese Intelligent Systems Conference (2019)
- Huang, J., Wang, P.: Observer design for the Lur'e differential inclusion system with Markovian jumping parameters. *Int. J. Syst. Sci.* **44**(12), 2338–2348 (2013)
- Kim, S.: Control synthesis of Markovian jump fuzzy systems based on a relaxation scheme for incomplete transition probability descriptions. *Nonlinear Dyn.* **78**(1), 691–701 (2014)
- Shi, P., Yin, Y., Liu, F.: Robust control on saturated Markov jump systems with missing information. *Inf. Sci.* **265**, 123–138 (2014)
- Hou, Z., Luo, J., Shi, P.: Stochastic stability of Itô different equations with semi-Markovian jump parameters. *IEEE Trans. Autom. Control* **51**(8), 1383–1387 (2006)
- Huang, J., Shi, Y.: Stochastic stability of semi-Markov jump linear systems: an LMI approach. In: Decision Control European Control Conference (2011)
- Huang, J., Shi, Y.: Stochastic stability and robust stabilization of semi-Markov jump linear systems. *Int. J. Robust Nonlinear Control* **23**(18), 2028–2043 (2013)
- Kim, S.: Stochastic stability and stabilization conditions of semi-Markovian jump systems with mode transition-dependent sojourn-time distributions. *Inf. Sci.* **385**, 314–324 (2017)
- Wang, B., Zhu, Q.: Stability analysis of semi-Markov switched stochastic systems. *Automatica* **94**, 72–80 (2018)
- Huang, L., Mao, X.: Stability of singular stochastic systems with Markovian switching. *IEEE Trans. Autom. Control* **56**(2), 424–429 (2011)
- Peng, S., Zhang, Y.: Some new criteria on  $p$ th moment stability of stochastic functional differential equations with Markovian switching. *IEEE Trans. Autom. Control* **55**(12), 2886–2890 (2010)
- Yuan, C., Lygeros, J.: Stabilization of a class of stochastic differential equations with Markovian switching. *Syst. Control Lett.* **54**, 819–833 (2005)
- Yan, Z., Zhang, W.: Finite-time stability and stabilization of Itô stochastic systems with Markovian switching: mode-dependent parameter approach. *IEEE Trans. Autom. Control* **60**(9), 2428–2433 (2015)
- Yu, W., Chen, G., Cao, M.: Second-order consensus for multiagent systems with directed topologies and nonlinear dynamics. *IEEE Trans. Syst. Man Cybern.* **40**(3), 881–891 (2010)
- Mao, X.: *Stochastic Differential Equations and Application*, 2nd edn., Chichester, UK (2006)



# Agility Detector Designed for Automobile Detection

Shuang Liu, Xizhong Shen<sup>(✉)</sup>, and Rongfan Leo

School of Shanghai Institute of Technology, Shanghai 200000, China  
xzshen@yeah.net

**Abstract.** For self-driving automobile, a prompt object detector under limited computation budget is highly desired yet indispensable for safety check. In this paper, taking restrained computation resource into consideration, we proposed a highly efficient detection backbone composed of transform shuffle units and a novel lite detection head made of depthwise convolutions to achieve real time performance for automobile detection. A set of experiments were conducted to demonstrate the customization ability of the proposed detection model and the well balanced one in terms of speed, accuracy and memory size along with several guiding rules in lite detection neural network design.

**Keywords:** Automobile detection · Detection backbone · Transform shuffle unit · Neural network

## 1 Introduction

Recent years Deep Neural Network (DNN) based solutions have become ubiquitous in computer vision, and Convolutional Neural Network (CNN) based methods have dominated the object detection field since R-CNN [1] ushered in the neural detection paradigm era in 2015. For self-driving cars, a prompt object detector under limited computation budget is highly desired yet indispensable for safety check. Specifically for driving scenario, to tackle the barriers for deploying neural networks on embedded systems, we proposed a deliberately refined lite model for automobile detection, which is called lite refined automobile detector. It composes of three functional modules: rapid feature extracting module, feature fusion module and refined anchor discretizing module. To bridge the gap between general classification neural networks and detection backbone, we proposed transform shuffle unit (TSU) as basic feature extractor and corresponding lite detection head as predictor to perform object classification and localization.

TSU based backbone joint with the lite detection head achieved well balanced performance in terms of speed, accuracy and memory size, which is in favor of real time detection on embedded systems. Trained and tested on Utacity driving dataset, the model we suggest achieved relatively high detection speed at 73 FPS as well as acceptable average precision (AP) of 0.511 for automobile yet requires less than 10 Mb storage memory. To push it further, our smallest model has AP of 0.478 with 4.24 Mb and fastest model has AP of 0.494 at 80 FPS.

## 2 Related Work

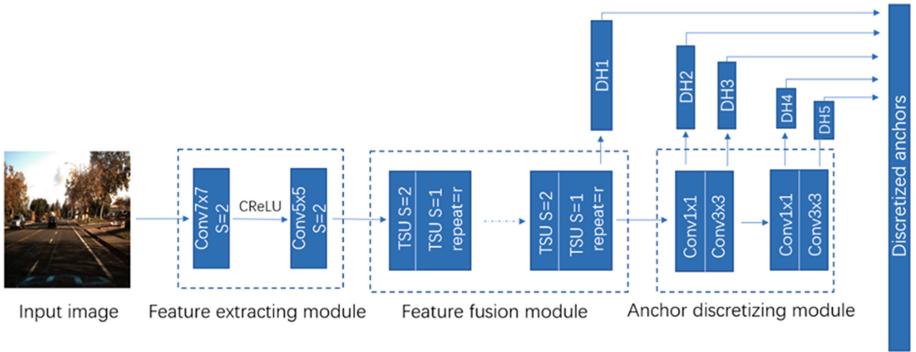
AlexNet popularized deep convolutional neural networks by winning the ImageNet Challenge: ILSVRC 2012. Vgg-net family had achieved an important milestone in ImageNet classification, which utilized the formalized design of deep neural networks. Another stride was made by GooLeNet family in which inception module was employed to enhance the represent ability and enrich receptive field of convolutional layers. The next architectural design was the ResNet family [2], which combined convolution layers with skip connections to ameliorate degradation in deep neural networks resulting in a model with 152 layers and competitive performance in multiple learning tasks.

As one can see, building deeper and more complex networks to achieve better accuracy has always been a trend in neural network design. Another trend in developing neural models is toward efficiency and lower latency where memory and computation budget are the main concerns. The advance of neural networks promotes the progress of object detection. As a two-stage detector, R-CNN pioneered CNN method in the second-stage classifier, which has greatly improved the accuracy and popularized CNN in the detection of modern objects. Over the years, R-CNN has improved in both speed and accuracy through using learned object proposals [3]. Numerous extensions to this framework have been proposed, e.g. [4, 5]. The representative one-stage detectors are YOLO, SSD and RetinaNet that have been tuned for speed. Focusing on an extreme speed/precision tradeoff, YOLO uses a single straightforward CNN network to regress the object location and simultaneously classify its category. SSD proposed a novel multi-scale convolutional prediction layers to fit various size of objects in an image. RetinaNet shares many similarities with previous dense detectors, especially anchors concept introduced by RPN and the use of the feature pyramid in SSD and FPN [4]. Notably, the author argued that single-stage detectors can catch up with or exceed the accuracy of two-stage detectors at a faster speed with focal loss. Another emerging trend for object detection is anchor free while key points are employed for detection. Among them there are CornetNet [6] and CenterNet [7] that demonstrated competitive performance lately.

Fancy as those detectors are, time-consuming is their common character and not friendly to mobile or embedded systems with limited computing resource. Towards practical applications, more efficient and compact detector as SqueezeDet [8], Tiny SSD [9], YOLO-LITE [10] were introduced, along with specific object detectors like FaceBoxes. Our work also follows this track for automobile detection used for embedded systems in timely manner. Extensive experiments designed to bridge the gap between general object detector and automobile detector for driving scenarios leaded to our main contributions that are 1) a well speed and accuracy trade-off detect neural backbone, 2) a novel efficient detect head and 3) an uniformed detection structure that can be customized for certain embedding conditions. Namely, we call our lite refined automobile detector LRADet.

### 3 Lite Refined Automobile Detector

This section presents the details in our model designed for automobile detection for driving scenarios which balances well between accuracy, speed and memory footprint. LRADet achieves relatively high detection speed at 73FPS as well as acceptable average precision of 0.513 for automobile detection yet requires less than 10 Mb storage memory. As demonstrated in Fig. 1, LRADet consists of three functional modules: rapid feature extracting module, feature fusion module and refined anchor discretizing module. Those modules are all made of convolutional layers which indicates our model is fully convolutional even without pooling layers for faster inference speed.



**Fig. 1.** Structure of LRADet

#### 3.1 Rapid Feature Extracting Module

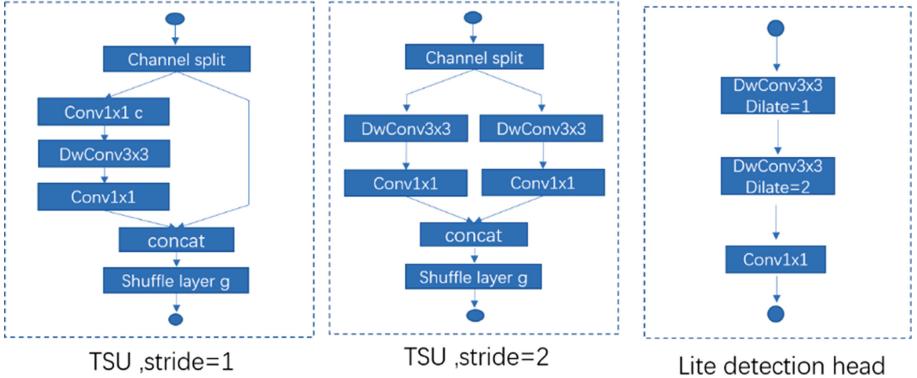
Since at the input stage large spatial size usually consumes more computation, we shrink the spatial size of inputs rapidly by utilizing larger kernel and employ C.ReLU to further speedup. As common guidance in convolutional network design, we also adopt the variational channel strategy in, by assigning a smaller number of channels in shallow layers and increasing the number linearly when the feature extractors go deeper. This will also help saving computing cost.

In the early stage, greater receiving field and wider feature vector are critical to improve the localization capabilities, we adopt  $7 \times 7$ ,  $5 \times 5$  kernels for layer conv1 and conv2 with stride 2 and same padding to downsample the input 4 times and the number of kernels are 32, 64 respectively as illustrated in Fig. 1.

We utilize the C.ReLU activation function to speedup the inference. In CNN, the lower layers filters are paired. To this end, before applying ReLU, C.ReLU can simply connect the negative outputs, doubling the number of output channels. In other word it twice the number of channels by concatenating operation. Employing C.ReLU can increase speed with negligible compromise in accuracy.

### 3.2 Feature Fusion Module

In order to achieve an excellent balance between representation ability and computing cost, leading networks such as Xception and ResNeXt [11] have introduced depthwise separable convolutions (also known as group convolution) into their building blocks. Depthwise separable convolutions can save lots of weight parameters and float point operations which is in favor of faster speed and smaller memory footprint but, especially for lite or tiny networks, expensive point-by-point convolutions, followed by depthwise convolutions, leads to a limited number of channels to meet the complexity or resource constraints, which might seriously compromise accuracy. Taking that into consideration, we examined typical efficient feature extractors including inception module, ResNeXt module [11] and shuffle unit along with a set of configurations respectively. Our resulting feature fusion module is based on shuffle unit with a mobile-Net like structure to enhance representation capacity and performance, which we refer to as transform shuffle unit (TSU).



**Fig. 2.** Tructures of the proposed TSU and lite detection head

As demonstrated in Fig. 2, the TSU with stride 1 has two branches that divide the input channels then each branch has the same spacial size and half channels as to the input feature map. The left branch operates transform convolution by two  $1 \times 1$  convolution and one depthwise convolution with a hyper-parameter  $c$  to control its capacity and the right branch just preserves its information. These branches are transformed to the same size as inputs by concatenating operation and then the channel shuffle [8] is performed for information fusion with a hyper-parameter  $g$  to control its divergence. The TSU with stride 2 also has two branches and each branch has a  $3 \times 3$  convolution with stride 2 to downsample the input and then they will be concatenated before shuffle operation with the hyper-parameter  $g$ . In that fusion module TSU can be repeat several time that renders another hyper-parameter  $r$ . with these tree hyper-parameters  $cgs$ , the model can be customized to suit different detection tasks.

### 3.3 Refined Anchor Discretizing Module

As illuminated in Fig. 2, based on separable depthwise convolution, we designed a lite detection unit called lite detection head which composes of two stacked depthwise convolution with different dilation rate and one  $1 \times 1$  convolution to collect the information. In addition, classification and localization can also utilize  $1 \times 1$  convolution directly operating on corresponding layers in the backbone. To match the diverse scales of objects, feature maps of various solution are involved to construct feature pyramid which discretizes prior anchors for localization as shown in Fig. 1. Similar to SSD, we set linearly increasing resolution scales for prior anchor boxes at different anchor points in each relative pyramidal feature map and for our automobile detecting scenario the aspect ratios are primarily set to 1:1, 1:3 and 3:1 as for width versus height.

### 3.4 Training and Inference

Automobile detecting is mainly performed on driving scenario which has unique background information of roads or streets and somehow the semantic information on screen plays an indispensable role in detection accuracy and efficiency. Our training data set is the subset of an open source self-driving data set which contains 18000 training images and 4241 validation images.

Before training, data augmentation pipeline was established to preprocess each image. The procedure follows: scale transformation, random flipping, random cropping, color brightness, color distortion and ground truth boxes filtering for the sake of model robustness and generalization.

In training, focal loss was employ for classification loss and smooth L1 loss [1] was assigned for localization loss. There is a significant imbalance between positive and negative examples in the dense prediction. Training for speed optimization and stable, we need to identify anchors corresponding to a ground truth labels so the match strategy was: first, matching each ground truth bounding box to an anchor with the best Jaccard overlap and then preserving anchors to that ground truth label with Jaccard overlap above a threshold; at last, sorting them by Jaccard threshold, the top ones are selected so that the ratio of positives and negatives is 3:1 at most. however focal loss for classification can be free of this matching stage since it is designed to tackle that imbalance.

After training, in inference non-maximum suppression was to be executed to produce the predicting results.

## 4 Experiment

In detection field there are considerable literatures studying on general detection based on VOC or COCO data sets among them the influential ones including Faster RCNN, YOLO, SSD and their numerous variations. In our work we focus on driving scenarios and specifically for car detection used for resource limited devices. Towards a real-world lite detection application, our model drills down to its trade-off between speed, accuracy and memory size.

#### 4.1 Towards Better Accuracy

There seems a consensus in neural networks that deeper structure leads to better accuracy. Based on the framework as demonstrated in Fig. 1, where T2 denotes TSU with stride 2 and T1 stride 1. Parameter r means repeating r times, the corresponding structures of TSU are shown in Fig. 2.

We constructed four backbone networks varied in the number of TSU and the hyper-parameter g, c was set to 2 and 1 respectively. The detailed configuration is shown in Table 1. To focus on the efficiency of the backbones we only utilized plain  $3 \times 3$  convolutions and four pyramidal prediction layers as their detection head, which will be demonstrated in Subsect. 4.2. In the training process, the optimizer was SGD with the learning rate of 0.01 and 0.0005, and the batch size is 32. The maximum number of iterations was 100k. In inference, these models produced a large number of predicted boxes for each input image. First, we filtered out most of the boxes with the confidence threshold of 0.01, then NMS was executed with Jaccard overlap of 0.45, and reserved up to the first 200 boxes. Soft-NMS was also used as an auxiliary. We measured the speed on GTX2070s and cuDNN v7.6 with a test set containing 4241 images at batch size of 8. The metric was average precision (AP) of automobiles and the results were shown in Table 2, the T1 and T2 denote TSU with stride 1 and stride 2 respectively, r means unit repeat. T1(r=3) means repeating TSU (with stride 1) 3 times. xn denotes repeating the block T2-T1(r=1) n times, and in that case one for TSU block1 and one for TSU block2 corresponding to Table 1.

**Table 1.** The detailed configuration of backbone networks

Layer/Block	Input	Kernel/Stride	Output
Conv1	(300, 480, 3)	$k = (7, 7)/s = 2$	(150, 240, 32)
Conv2	(150, 240, 32)	$k = (5, 5)/s = 2$	(75, 120, 64)
TSU block 1	(75, 120, 64)	T2 = 1, T1 = r	(37, 60, 128)
TSU block 2	(37, 60, 128)	T2 = 1, T1 = r	(18, 30, 256)
Conv5	(18, 30, 256)	$k = (1, 1)/s = 1$	(18, 30, 128)
Conv6	(18, 30, 128)	$k = (3, 3)/s = 2$	(9, 15, 256)
Conv7	(9, 15, 256)	$k = (1, 1)/s = 1$	(9, 15, 128)
Conv8	(9, 15, 128)	$k = (3, 3)/s = 2$	(4, 7, 256)

It is clear that model with more weights produce better results with correspondingly slower speed and larger storage memory. The controlled experiments reflected the potential presentation capacity of the introduced model. We chose the model with T2-T1(r=3) as our backbone to perform the rest experiments for its relatively higher AP and inference speed.

**Table 2.** The average precision of automobiles

TSU block	Layers	AP	FLOPs (B)	FPS	Size (M)
T2-T1( $r=1$ ) x2	10	0.424	0.774	82	9.8
T2-T1( $r=2$ ) x2	12	0.494	0.812	79	10.2
T2-T1( $r=3$ ) x2	14	0.530	0.845	77	11.3
T2-T1( $r=5$ ) x2	18	0.551	0.925	65	14.6

## 4.2 Towards Faster Speed

More prediction layers and corresponding prior anchor boxes will definitely have longer latency. Followed the same training and inference setting and based on the backbone selected above, we designed three types of detection heads and each differed in the number of prediction layers that had different spatial sizes. As shown in Fig. 1, we assembled DH3-5 as a detection head while DH2-5 and DH1-5 the other two heads, namely we call them H1, H2 and H3 respectively.

Since our model was designed for specific object detection, we aimed to develop a suitable strategy to discretize prior anchors for higher efficiency. For explicit comparison, we still used plain  $3 \times 3$  convolutions as the feature extractor. As illustrated in Table 4 To further exploit the proposed structure we conducted another set of experiments by switching the last three plain convolutional layers to the TSUs and remained the same detection heading setting. Table 3 shows the resulting performance and it turned out that too many TSUs degenerated the presentation capability of the model or it revealed that shuffle operation in the later stag of a compacted model is not an option since it may introduces degeneration during training and slower convergence and inference speed, which demonstrates a principle in detection architecture design that for higher efficiency different stages require of suitable functional modules.

To this end, we chose the DH2 on backbone (S1-B3/S2-B3) as our basic detection structure for its well balance in speed, accuracy and size.

**Table 3.** Detection results of different detector heads based on the selected backbone

Structure	AP	FLOPs (B)	FPS	Size (M)
H1	0.494	0.850	80	10.1
H2	0.520	0.845	77	11.3
H3	0.530	0.841	66	12.8

## 4.3 Toward Smaller Size

The feature extractors we used for detection head were plain  $3 \times 3$  convolutions which contributed untrivial part of weights in the model but also resulted in bigger memory footprint. So we employed separable convolutions and lite detection

**Table 4.** Detection results of different detection heads on unified TSU backbones

Structure	AP	FLOPs (B)	FPS	Size (M)
DH1-TSU	0.43	0.848	56	11.7
DH2-TSU	0.40	0.852	46	13.8
DH3-TSU	0.48	0.857	40	14.2

head to shrink model size. In Table 5, the controlled experiments based on the selected backbone showed that plain convolution had the best accuracy because of more trainable weights which also resulted in larger memory size. The separable convolution decreased the number of weights greatly but also led to a sharp drop in its accuracy then lite detection head balanced well between them in terms of accuracy, speed and size.

**Table 5.** The comparison of different components in detection heads

Structure	AP	FLOPs (B)	FPS	Size (M)
$3 \times 3conv$	0.520	0.845	77	11.3
$3 \times sep - conv$	0.501	0.840	66	7.6
Lite-head	0.511	0.842	73	9.5

## 5 Conclusion

For automobile detection in driving scenario, we proposed a lite detection framework designed for source restrained applications which consists of fast extraction module, transform shuffle units and lite detection head. Through extensive experiments, we developed a well balanced automobile detector with the speed of 73FPS, mAP of 0.511 on Utacity self-driving data set and a size of 9.5M. In addition, several general guiding rules in detection neural network design can be drawn as follows: 1) uniformed structure has faster inference speed; 2) in the detection backbone different stages require of suitable functional modules for higher efficiency and 3) for specific object detection, prior anchor should be crafted according to its scale and aspect ratios.

## References

1. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)

2. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
3. Pinheiro, P., Collobert, R., Dollár, P.: Learning to segment object candidates. In: Advances in Neural Information Processing Systems, pp. 1990–1998 (2015)
4. Lin, T., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
5. Shrivastava, A., Sukthankar, R., Malik, J., et al.: Beyond skip connections: top-down modulation for object detection. arXiv preprint [arXiv:1612.06851](https://arxiv.org/abs/1612.06851) (2016)
6. Law, H., Deng, J.: CornerNet: detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750 (2018)
7. Duan, K., Bai, S., Xie, L., et al.: CenterNet: keypoint triplets for object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6569–6578 (2019)
8. Wu, B., Iandola, F., Jin, P., et al.: SqueezeDet: unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 129–137 (2017)
9. Wong, A., Shafiee, M.J., Li, F., et al.: Tiny SSD: a tiny single-shot detection deep convolutional neural network for real-time embedded object detection. In: 2018 15th Conference on Computer and Robot Vision (CRV), pp. 95–101. IEEE (2018)
10. Pedoeem, J., Huang, R.: YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. arXiv preprint [arXiv:1811.05588](https://arxiv.org/abs/1811.05588) (2018)
11. Sharma, A., Muttoo, S.K.: Spatial image steganalysis based on ResNeXt. In: 2018 IEEE 18th International Conference on Communication Technology (ICCT), pp. 1213–1216. IEEE (2018)



# Review of Model Predictive Control Methods for Time-Delay Systems

Lei Liu<sup>(✉)</sup>, Yi He, and Cunwu Han

North China University of Technology,  
Beijing 100144, China

liulei\_sophia@163.com, heyi\_reflection@163.com, cwhan@ncut.edu.cn

**Abstract.** Model Predictive Control (MPC) has accomplished several important achievements after decades of growth. Through the perspective of model predictive control for time-delay systems, the system output may deteriorate when there is time-delay or even the system will be unstable. This paper presents the performance of model predictive control methods for time-delay systems with state delays and control delay. The characteristics of the methods used to deal with various types of time-delay systems are illustrated and the benefits, drawbacks and future directions for research are highlighted. Finally, an example is given to test that model predictive control is one of the most successful methods of tackling system control problems.

**Keywords:** Time-delay systems · Model predictive control · Algorithm implementation · Simulation

## 1 Introduction

Since the 1970s, people have been searching for different methods that can achieve high quality control efficiency, formally Model Predictive Control (MPC) was built against this context of demand. It was originally suggested by Richlet and Cutler [1]. It was successful in dealing with practical control issues in the field of industrial application, and rapidly developed in power, aviation, oil, and other industries. It had corresponding theoretical research thereafter.

The predictive control algorithm's principle is to use the process model to predict the system's future dynamic behavior under some control, and then roll to solve the optimal control function and implement the current control according to the performance index and constraint conditions given. The prediction of the future dynamic behavior is detected and corrected in real time at each rolling stage, which is summarized as three features: predictive model, rolling optimization, and inverse feedback correction [2], including model algorithmic control, dynamic matrix control and generalized predictive control [3].

Control system's time-delay problem is divided into control delay and state delay. The control delay and state delay in the phase under control would increase the control system's control difficulty. In industrial process time delay is very

common. The disruption cannot be observed in time due to the presence of time delay, resulting in a extreme overshoot of the regulated variables. The control effect reflected on the plant's production sometimes lags behind for some time, which seriously affects control system performance. Model Predictive Control can predict future performance based on past and present data. It has high adaptability characteristics, good robustness and fast adjustment period [4]. This can effectively enhance the control efficiency of time-delay systems, thus currently forming a hot research path for model predictive control. In this paper, model predictive control methods are summarized and commented on for time-delay systems with constraints, nonlinearity, uncertainty etc.

## 2 Model Predictive Control for Time-Delay Systems

### 2.1 Model Predictive Control for Time Delay Systems with Constraints

Time-delay system is an important industrial process control class. In practice, input quantity constraints are frequently present due to safety requirements. Therefore, Paper [5] proposes a model approach for predictive control of time-delay systems with input restrictions. It adopts dual-mode control structure and constructs offline time-delay system terminal invariant ellipse set based on L-K functional, which ensures the stability of the closed-loop system while reducing the calculation quantity. In industrial processes, unpredictable time-delay systems are widespread, and input saturation constraints often apply. Therefore, This paper [6] proposes a robust model predictive control method for a class of uncertain time-lapse systems. When the predicted state exceeds the invariant set, the dual-mode control structure is used to ensure feasibility of the input. The robust elliptic invariant set of unsure time-delay system is constructed offline based on functional L-K. To ensure the stability of the closed-loop method, the design of the elliptic set is transformed into a convex problem of linear matrix inequality optimisation.

In the design of feedback control laws the influence of time-delay state is often ignored for time-delay systems. A model predictive controller with memory state feedback for a class of discrete-time linear uncertain time-delay systems with input constraints is proposed in this paper [7]. To solve the problem of time-delay systems' uncertainty and input constraints, the upper bound of robust output function is first described, and the LMI is convex optimal. The necessary conditions for the system's stability are given within the transformation framework.

### 2.2 Model Predictive Control for Discrete Time-Delay Systems

The design of a robust predictive output feedback controller is studied in [8] for a class of unpredictable discrete time-delay with input-output constraints to make the closed-loop system asymptotically stable and to reduce the rolling

time-domain performance index online. The sufficient conditions for the existence of the output feedback controller and its construction method are given on the basis of the rolling optimisation principle of predictive control. The power is obtained by applying the concept of linearisation of conic compensation. The solver's iterative method. Factors such as time-varying, interference, and uncertainty make calculation of the system optimization issue more complicated in the actual industrial process. In [9], the system constraints and the specifications of the objective function are considered for a class of discrete systems with known time delay, A kind of generalized neural projection network is used to optimize the system model online, and the MPC problem is transformed into a restricted quadratic programming problem to be solved. The paper [10, 11] aims at the MPC problem of a discrete time-delay system and incorporates the neural dynamic optimization approach, which can increase the online calculation speed, which can also be used in other nonlinear convex optimization problems, which enriches the use of MPC in more fields.

### 2.3 Model Predictive Control for Multi Model Time-Delay Systems

The paper [12] integrates the generalized predictive control [13] in predictive control with the multi-model, obtains the global optimal weight solution using the genetic algorithm [14] and proposes a multi-model predictive control scheme [15] based on the genetic algorithm, based on the characteristics of the time-delay system. Comparing the updated control scheme with the application of PID control in the time-delay system, the algorithm not only realizes the enhanced time-delay system control precision, stability, rapid speed and high robustness.

### 2.4 Model Predictive Control for Nonlinear Time-Delay Systems

The nonlinearity of the [24] method is universal within the functional industrial system. The difficulties in nonlinear MPC research are mainly expressed in many aspects through the time-delay phenomenon, such as model selection, energy function solution and nonlinear algorithm [16]. In this paper [17], aiming at the two major problems of the industrial control system, the nonlinear and time-delay problems are solved by integrating the standard dynamic matrix control algorithm into the predictive control. The instability caused by model mismatch, distortion and disruption can be compensated in time, and the T-S fuzzy model can explain the nonlinear system well. Fuzzy control and predictive control are combined organically to solve the nonlinear time-delay system problem.

In [18], a Lyapunov function is developed for a class of unknown nonlinear discrete-time systems with multiple state delays and nonlinear disturbances, and a state feedback controller is constructed using the min max problem of optimization in the infinite time domain. Based on [20, 21], the paper [19] constructs an improved quadratic Lyapunov functional for a class of unknown nonlinear discrete-time systems with multiple states and input delays by making full use of the upper and lower bound knowledge of the interval, in order to which the system's conservatism. In [22], the control input is solved by rolling optimisation

for a class of uncertain systems with nonlinear disturbances and multiple states and simultaneous input delays. The problem of optimization of the infinite time domain quadratic performance index is converted into a linear objective minimization problem with LMI constraints, and the performance index is optimized online and in real time, rendering the machine state smooth and stabil. In [23], a combination of model predictive control and fuzzy theory is used for nonlinear discrete systems with state delay characteristics, based on the TS fuzzy model, and the parallel distribution compensation principle is used to solve the corresponding state under standard bounded conditions. Law on managing feedback.

## 2.5 Model Predictive Control for Uncertain Time-Delay Systems

The system's control performance and robustness are limited due to the existence of time-delay in many industrial systems. The total robust MPC Method of multi-objective, uncertain time delay framework is proposed in paper [25, 26]. In paper [27], the MPC strategy of feedback is applied with polyhedral uncertainty on the time-delay system. In paper [28], the robust MPC is proposed based on the control invariant set, and the unique state feedback control is designed to handle the system with knots. However, the degree of freedom given by the state feedback method is limited, leading to system conservatism and unsatisfactory control efficiency. A robust predictive control algorithm for the time-delay system with structural uncertainty is conceived in [29]. Based on the control invariant set method, the algorithm adopts dual-mode control and closed-loop control strategy to increase control design freedom, thereby expanding the system's initial feasible region and ensuring better performance in control.

Currently, some achievements have been made in researching robust predictive control for uncertain [30–34] time-delay systems, and article [35] proposes a time-delay compensation method for uncertain time-delay systems represented by a class of multicellular models. A robust predictive control system based on a lag. The difficult min-max optimization problem is transformed into a convex optimization problem with LMI constraints, and a robust model predictive controller is designed that can guarantee the stability of the closed loop system, but in the real industrial production and engineering control field If the time lag is overlooked The effect of the built controller on the system makes it extremely conservative. Thus, [36] proposed a robust LMI-based performance feedback method for a class of unknown discrete time-delay systems with multi-cell structure. The control algorithm reduces the algorithm's online calculation by calculating the offline performance input gain matrix in the stable, predictive controller. Using output feedback breaks by restricting the previous algorithm that the state of the device must be measurable, rendering the algorithm a lower conservator.

## 3 Existing Problems and Research Directions

1. Most current work considers the state feedback in controller design, but the state of the system is often difficult to obtain in the actual process

implementation, so the next step may be the time-delay system's predictive control approach for the output feedback model.

2. The model of nonlinear time-delay method predictive control theory isn't fine. To further that the effect of time-delay on the system, the controller's parameter selection method should be analyzed in detail to ensure the controller's efficiency and stability.
3. In dealing with the issue of time-delay, the time-varying time-delay can be analyzed in the future when predictive regulation of the system is done on the basis of time-invariance, so as to analyze the system's time-delay situation more comprehensively.
4. It promotes the combination of neural network, fuzzy logic and other theories to encourage the production of model predictive control for time-delay systems, with the growing demand for practical application.

## 4 Algorithm Implementation and Simulation Analysis

Dynamic Matrix Control (DMC) is a sort of predictive control based on the response model of the unit phase. Taking as an example the dynamic matrix control of the siso time-delay system, the predictive equation of the time-delay system unit step response model and the expression of the law on dynamic matrix control are as follows:

$$\begin{cases} Y_p(k+1|k) = L\hat{Y}(k) + S_u \Delta U(k) + S_d \Delta d(k) \\ \Delta u(k) = K_{mpc}(R(k+1) - L\hat{Y}(k) - S_d \Delta d(k)) \end{cases} \quad (1)$$

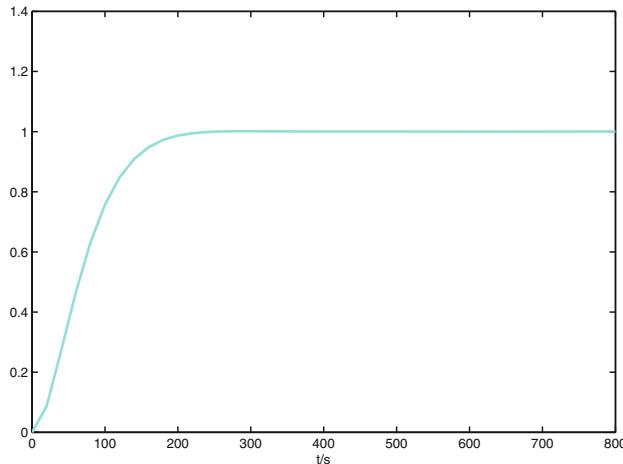
where  $S_u$  and  $S_d$  are the time-delay response coefficient matrix of the control input and observable disturbance input unit,  $K_{mpc}R(k+1)$  is a feedforward control of the future value of the reference input,  $-K_{mpc}S_d\Delta d(k)$  is a feedforward control of the input of interference,  $-K_{mpc}L\hat{Y}(k)$  is a feedback part based on the estimated state.

### 4.1 Algorithm Implementation

Suppose an industrial object's transfer function is shown below, and the dynamic characteristics after using DMC are shown in the figure. The sampling interval in the simulation is  $T = 20\text{s}$  and the optimized time domain is  $P = 10$ ,  $M = 2$  is the monitoring time domain and  $N = 20$  is the simulation time domain.

$$G_p(s) = \frac{e^{-10s}}{90s + 1} \quad (2)$$

The system's phase response curve using DMC control in the figure shows that the device that uses DMC control has quick response speed, little adjustment time and no over-shoot (Fig. 1).

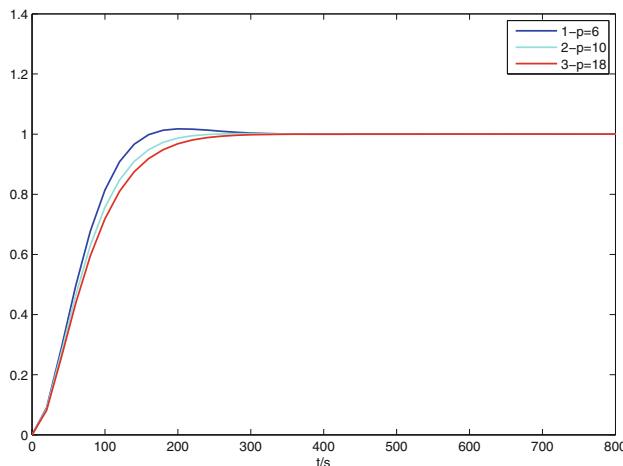


**Fig. 1.** Dynamic response curve of time-delay system

#### 4.2 Effect of P on System Dynamic Performance

The optimized time domain  $P$  describes how many steps from  $K$  to the output in the future approximate the predicted value. When the sampling duration is  $T = 20\text{ s}$ , the control time domain is  $M = 2$ , the simulation time domain is  $N = 20$  and the optimized time domain  $P$  is 6, 10 and 18 respectively (Fig. 2).

The curves 1, 2 and 3 in graph (2) respectively represent the step response curves at  $P = 6$ ,  $P = 10$ , and  $P = 18$ . From the figure it can be seen that the bigger the  $P$ , the greater the system's stability and its speed. The smaller the

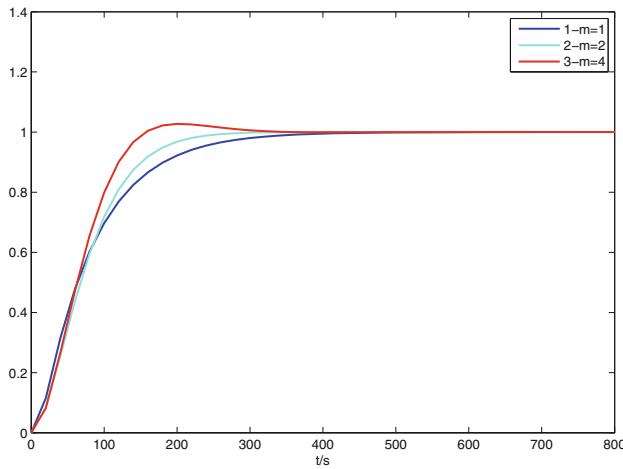


**Fig. 2.** Impact of different  $P$  on system performance

$P$ , the faster the method is and the less robust it is. The  $P$  option must take account of both stability and tempo.

#### 4.3 Effect of M on System Dynamic Performance

The control time domain  $M$  represents the number of changes to be calculated in the future control amount. When the sampling duration is  $T = 20\text{ s}$ , the optimized time domain is  $P = 18$ , and the simulation time. The  $N = 20$  domain is. The control time domain  $M$  is taken, as shown in Fig. 3, at 2, 4 and 6 respectively.



**Fig. 3.** Impact of different  $M$  on system performance

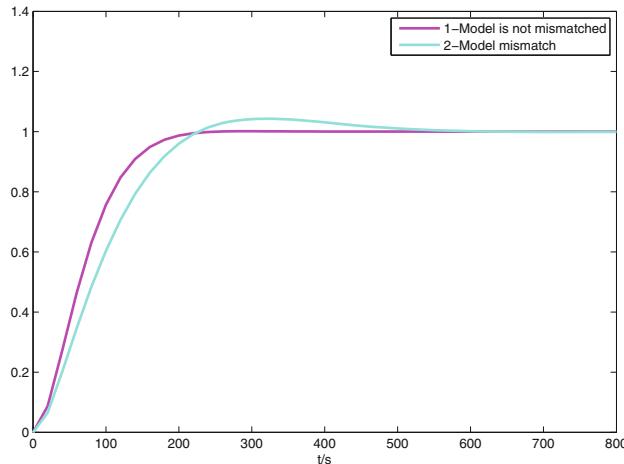
In the figure, the curves 1, 2, and 3 are the curves of response when  $M = 1$ ,  $M = 2$ , and  $M = 4$ , respectively. When  $M$  increases, the speed of the system increases, and the stability decreases. When  $M$  decreases, the speed of the system decreases, and the stability increases. Before  $P$  is modified, and  $P$  is greater than or equal to  $M$ .

#### 4.4 Response Curve When the Model Is Mismatched

When the prediction model is mismatched, that is, when  $G_m(s) \neq G_p(s)$ , the response curve is shown in Fig. 4.

$$G_m(s) = \frac{2e^{-6s}}{40s + 1} \quad (3)$$

In the figure, if the model is not misaligned, curve 1 is the step response curve, and curve 2 is the step response curve when the model is inconsistent. As can be seen from the figure, the speed of the DMC control decreases when the



**Fig. 4.** Model unmatched and mismatched response curves

model is not matched, but the control effect is strong, and the speed of entering the stable state is high. Hence, DMC has better robustness to model mismatch when forming closed-loop control.

## 5 Conclusion

A hot topic in the field of predictive control is model predictive control of the time-delay system. More in-depth work is required to develop a more accurate model and bring forward a more accurate optimisation algorithm. In this paper, model predictive control methods for time-delay systems are evaluated in recent years, the benefits, concepts and drawbacks of the current methods are analyzed, and future directions for study are discussed. Ultimately, simulation of the dynamic matrix control for time-delay systems shows the advantages of model predictive control. The model predictive control theory and technology will develop rapidly, and the prospect for application will be brighter.

**Acknowledgements.** This work is partially supported by the Youth Foundation of Beijing Nature Science Grant (4202022, 4154068), the Youth Talent Cultivation Program of Beijing, the National Natural Science Foundation of China (61473002, 61573024). North China University of Technology Yuyou Talent Support Program.

## References

1. Richalet, J., Rault, A.: Model predictive heuristic control application to industrial process. *Automatic* **14**(1), 413–428 (1978)
2. Xi, Y.G.: Predictive Control. National Defense Industry Press, Beijing (1991)
3. Chen, H.: Model Predictive Control. Science Press, Beijing (2013)

4. Su, D.Q.: Predictive Control System and Its Application. Mechanical Industry Press, Beijing (1996)
5. Zhang, J., Pei, R., Chen, T.S.: Model predictive control for input constrained time-delay systems. *J. Syst. Simul.* **15**(7), 1051–1053 (2003)
6. Zhang, J., Xie, R.H., Ji, B.Y., Wang, B.: Robust model predictive control for input constrained time-delay systems. *Electric Mach. Control* **43**(4), 362–365 (2004)
7. Qin, W.W., Liu, G., Zheng, Z.Q.: Model predictive controller for input constrained time delay systems with memory state feedback. *J. South China Univ. Technol. (Nat. Sci. Ed.)* **40**(6), 63–69 (2012)
8. Chen, Q.X., Yu, L.: Robust model predictive control for uncertain discrete time-delay systems via dynamic output feedback. *Control Theory Appl.* **24**(3), 401–406 (2007)
9. Liang, X., Cui, B.T., Lou, X.Y.: Model predictive control based on generalized projection neural network optimization. *Appl. Res. Comput.* **33**(6), 1666–1669, 1675 (2016)
10. Zhao, L.P., Lou, X.Y.: Predictive control of discrete time-delay system based on neural dynamic optimization. *Technol. Innov. Appl.* **8**(13), 25–27, 29 (2018)
11. Peng, Y.G., Wei, W., Wang, J.: Model predictive control of time delayed restraint system based on neurodynamical optimization. *Chin. J. Sci. Instr.* **34**(5), 961–966 (2013)
12. Liu, G.Y., Shan, C.W.: Multiple models predictive control based on genetic algorithm. *J. Beihua Univ. (Nat. Sci.)* **15**(4), 557–560 (2014)
13. Zhang, Y.G., Shen, J., Li, Y.G.: Multi-model switching based GPC and its application to superheated steam temperature systems. *East China Electric Power* **37**(1), 164–168 (2009)
14. Qin, G.J., Ren, Q.C.: PID control and simulation based on genetic algorithm optimization. *Sci. Technol. West China* **10**(11), 9–10, 12–13 (2011)
15. Liao, E.H., Zhong, C.: Hybrid genetic algorithm using improved mutation operator. *Inf. Technol.* **36**(1), 123–125 (2012)
16. Xu, S.H., Sun, Q.X., Gu, W.J., Jiang, W.Z.: A survey of nonlinear predictive control model methods. *J. Naval Aeronaut. Eng. Inst.* **22**(6), 633–636 (2007)
17. Liu, G.P., Wu, J.F., Wang, S.M.: The application of fuzzy predictive control for nonlinear time-delay systems. *J. Harbin Univ. Sci. Technol.* **15**(2), 24–27, 34 (2010)
18. Zhou, W.D., Zheng, L., Liao, C.Y., Cai, J.N.: Min-max robust predictive control for multi-state time-delay systems. *J. Harbin Eng. Univ.* **37**(12), 1685–1690 (2016)
19. Zhou, W.D., Zheng, L., Liao, C.Y., Cai, J.N.: Robust prediction control for multiple time delay discrete nonlinear system. *J. Harbin Inst. Technol.* **47**(9), 24–30 (2015)
20. Li, J.X., Fan, Y.M., Shi, S.L.: Robust MPC algorithm for discrete-time systems with time-varying delay and nonlinear perturbations. In: Proceedings of the 29th Chinese Control Conference, pp. 3128–3133. IEEE Press, Beijng (2010)
21. Zhang, Y.X., Liu, M., Wang, J.H.: Robust model predictive control for uncertain discrete-time system with both states and input delays. In: 2008 Chinese Control and Decision Conference (CCDC 2008), pp. 279–284. IEEE Press, Yantai (2008)
22. Su, C.L., Zhao, J.C., Li, P.: Robust predictive control for a class of multiple time delay uncertain systems with nonlinear disturbance. *Acta Automatica Sinica* **39**(5), 644–649 (2013)
23. Wang, Y.J., Zhang, D.W., Yu, J.Q.: Model predictive control for a class of nonlinear systems with time-delay based on T-S model. *Hebei J. Ind. Sci. Technol.* **35**(1), 37–42 (2018)
24. Xi, Y.G., Wang, F.: Nonlinear multi-model predictive control. *Acta Automatica Sinica* **22**(4), 456–461 (1996)

25. Ding, B., Huang, B.: Constrained robust model predictive control for time-delay systems with polytopic description. *Int. J. Control.* **80**(4), 509–522 (2007)
26. Ding, B., Xie, L., Cai, W.: Robust MPC for polytopic uncertain systems with time-varying delays. *Int. J. Control.* **81**(8), 1239–1253 (2008)
27. Li, D., Xi, Y.: Constrained feedback robust model predictive control for polytopic uncertain systems with time-delays. *Int. J. Syst. Sci.* **42**(10), 1651–1660 (2011)
28. Zheng, P., Xi, Y., Li, D.: Robust model predictive control for time-delay systems with structured uncertainty. In: The 8th World Congress on Intelligent Control and Automation, pp. 1174–1178. IEEE, Jinan (2010)
29. Zheng, P.Y., Xi, Y.G., Li, D.W.: Closed-loop robust model predictive control for time-delay systems with structured uncertainties. *Control Theory Appl.* **30**(6), 683–692 (2013)
30. Zhao, G.R., Gai, J.F., Hu, Z.G.: Evolution of nonlinear model predictive control research. *J. Naval Aeronaut. Astronaut. Univ.* **29**(3), 201–208 (2014)
31. Guo, Y.J., Liao, F.C.: Robust preview control for multirate uncertain discrete-time systems with input delay. *Control Decis.* **32**(12), 2113–2126 (2017)
32. Yuan, X.J., Gao, C.C.: Stability of observer-based Uncertain discrete systems with time-varying delays. *Period. Ocean Univ. China* **48**, 206–211 (2018)
33. Reble, M., Mahboobi, E., Sfjanji, R., Nikravesh, S.K.Y.: Model predictive control of constrained nonlinear time-delay systems. *IMA J. Math. Control Inf.* **28**(2), 183–201 (2011)
34. Graichen, K., Kugi, A.: Stability and incremental improvement of suboptimal MPC without terminal constraints. *IEEE Trans. Autom. Control* **55**(11), 2576–2580 (2010)
35. Li, S.Q., Shi, Y.J., Chen, D.Y., Wang, J.M.: Robust model predictive control for time-delay systems with polytopic uncertainty. *J. Harbin Univ. Sci. Technol.* **16**(4), 108–113, 117 (2011)
36. Gai, J.F., Zhao, G.R., Gao, C., Geng, B.L., Shi, Y.J., Chen, D.Y., Wang, J.M.: Output feedback robust predictive control for polytopic uncertain time-delay systems. *J. Naval Aeronaut. Astronaut. Univ.* **34**(5), 423–429 (2019)



# Resistivity Inversion Solving Based on a GA Optimized Convolutional Neural Network

Peng Wang and Shurong Li<sup>(✉)</sup>

Automation School, Beijing University of Posts and Telecommunications,  
Beijing 100876, China  
lishurong@bupt.edu.cn

**Abstract.** Resistivity inversion is often used to measure the structure of formations. Various optimization methods can be used to solve the resistivity inversion problem. Resistivity inversion was numerically simulated by two algorithms in this paper: convolutional neural network (CNN) and convolutional neural network optimized by genetic algorithm (COG). When CNN was used to solve the resistivity inversion problem, it was found that the hyperparameters of CNN has a greater influence on the inversion results. Because genetic algorithm (GA) has the advantage of global search, GA was used to optimize the CNN, so as to obtain a better set of hyperparameters, then CNN was trained to solve resistivity inversion. The results showed that the optimized CNN algorithm has smaller inversion errors and is easier to converge to the global optimal solution.

**Keywords:** Resistivity inversion · Forward model · Genetic algorithm · Convolutional neural network

## 1 Instruction

In the process of petroleum exploration and development, stratigraphic survey is very important. Resistivity is often used to measure formations. In resistivity formation measurement, resistivity inversion is the most important content. The resistivity inversion problem is based on the forward resistivity problem. In other words, the resistivity inversion problem is the inverse of the resistivity forward problem [1]. Most formations have a layered structure. From an electrical point of view, these formations consist of formations with various resistivities. Different formations have different resistivities. Therefore, the layered structure of the formation can be reflected by the change curve of the resistivity in the vertical direction of the formation. After understanding the stratum structure model of the stratum, the laws of geophysics can be used to calculate the resistivity change curve in the vertical direction of the stratum, which is a positive problem of resistivity. On the contrary, the resistivity curve in the vertical direction of the

© The Editor(s) (if applicable) and The Author(s), under exclusive license

to Springer Nature Singapore Pte Ltd. 2021

Y. Jia et al. (Eds.): CISC 2020, LNEE 705, pp. 634–645, 2021.

[https://doi.org/10.1007/978-981-15-8450-3\\_67](https://doi.org/10.1007/978-981-15-8450-3_67)

formation has been measured, the formation model of the formation is obtained from the resistivity change curve, which is the inverse problem of resistivity [2].

For the resistivity forward problem, it can be solved using the calculation formulas of the laws of geophysics. But for the resistivity inversion problem, it is difficult or basically impossible to solve it by calculation. The resistivity inversion problem is always solved by continuous forward iteration through optimization methods, which approximates the true model of formation. The optimization methods for solving the resistivity inversion problem can be divided into two types. One is the iterative solution of linear methods, such as Newton's method, conjugate gradient method, least square method, etc. The other is to directly use non-linear methods, such as simulated annealing, GA, artificial neural networks, etc. For the problem of resistivity inversion, a large number of scholars at home and abroad have conducted research. In 2014, Deshan Feng et al. Conducted a two-dimensional resistivity inversion study using the least squares regularization method [3]. In 2015, Gengen Qiu et al. Used a nonlinear conjugate gradient method in the two-dimensional resistivity inversion [4]. In 2016, Loke et al. Conducted a three-dimensional resistivity inversion study through a least-squares optimization method that can modify smoothness constraints [5]. In 2017, Ning Zhou et al. Adopted the adaptive regularization method to study the one-dimensional inversion of magnetotelluric resistivity [6]. In 2017, Heriyanto et al. Used singular value decomposition (SVD) and Levenberg-Marquardt (LM) technology to perform one-dimensional (1-D) DC resistivity inversion [7]. The above resistivity inversion methods rely heavily on the initial formation model, and the sensitivity matrix needs to be calculated. In 2015, He Wang et al. Used BP neural network to test and invert the two-layer and three-layer stratigraphic models [8]. In 2017, Barboza et al. Conducted a two-dimensional DC resistivity inversion experiment using particle swarm optimization algorithm [9]. In 2018, Bo Cheng et al. Used an improved genetic algorithm to simulate the electrical sounding resistivity inversion [10]. In 2019, Haibin Zhou used particle swarm optimization to conduct a resistivity inversion study [11]. The above resistivity inversion study does not depend on the initial stratigraphic model, and does not need to calculate the sensitivity matrix. However, it has some shortcomings such as easy convergence to the local optimal solution, large amount of calculation, and low accuracy.

Different algorithms have their own advantages. By combining different optimization algorithms, they can generally combine their own advantages to make the results of resistivity inversion better. In this paper, GA and CNN are combined to take advantage of the global search of the genetic algorithm [12] to optimize the CNN hyperparameters. It makes CNN resistivity inversion easier to converge to the global optimal solution, and has a faster inversion speed and higher accuracy.

## 2 Mathematical Description

### 2.1 Forward Model

In this paper, the resistivity forward model is the forward model of the electromagnetic resistivity while drilling. Through Maxwell's equations, the constitutive relationship and the boundary relationship between different media. The Helmholtz equation of the electric field  $\mathbf{E}$  can be derived as follows:

$$\nabla^2 \mathbf{E} + k^2 \mathbf{E} = i\omega\mu \mathbf{J}_T, \quad (1)$$

where  $i$  is the imaginary unit,  $\omega$  is the angular frequency of the excitation,  $\mu$  is the magnetic permeability,  $\mathbf{J}_T$  is the current density,  $k$  is the propagation long constant, the expression is:

$$k^2 = \omega^2 \mu \varepsilon + i\omega\mu\sigma, \quad (2)$$

where  $\varepsilon$  is the dielectric constant,  $\sigma$  is the conductivity. Assuming

$$k = \alpha + i\beta, \quad (3)$$

where  $\alpha$  and  $\beta$  are the real and imaginary parts of  $k$  respectively, then

$$\alpha = \omega \sqrt{\frac{1}{2}\mu \left[ \sqrt{\varepsilon^2 + \frac{\sigma^2}{\omega^2}} + \varepsilon \right]}, \quad (4)$$

$$\beta = \omega \sqrt{\frac{1}{2}\mu \left[ \sqrt{\varepsilon^2 + \frac{\sigma^2}{\omega^2}} - \varepsilon \right]}. \quad (5)$$

Through mathematical calculation, The electric field intensity at a point  $P$  in space is [13]

$$\mathbf{E} = -\frac{i\omega\mu}{4\pi} \int_V \frac{\mathbf{J}_T e^{-ikR}}{R} dV, \quad (6)$$

where  $R$  is the distance from point  $P$  to the transmitting coil. The induced electromotive force on the receiving coil can be derived as follows:

$$V = ih\omega\mu \frac{e^{ikL}}{L^3} (1 - ikL), \quad (7)$$

where  $h$  is the instrument constant,  $L$  is the source distance. Substituting  $k$  into the above formula to obtain the forward model is as follows:

$$V = ih\omega\mu \frac{e^{ikL}}{L^3} [\alpha L - i(1 + \beta L)]. \quad (8)$$

The forward model can be solved using the finite element method.

## 2.2 Inversion Model

The inverse problem is to explore the internal law of things from observable phenomena according to the evolution of things. Some classic research on inverse problems can be traced back very early. The earliest research on inverse problems originated from the problem of directional design. In recent decades, due to actual production needs, research on inverse problems has flourished. At present, the research on inverse problems has covered all fields of modern production, life and research. Inverse problems also tend to be nonlinear. To solve nonlinear inverse problems, it is usually necessary to repeat the forward and backward iterations after linearization, so the inverse problems can be solved by optimization methods.

The resistivity forward model can be expressed as:

$$y = f(x), \quad (9)$$

where  $f$  is the resistivity forward function,  $x$  is the formation model,  $y$  is the resistivity curve. So the resistivity inversion model can be expressed as:

$$x = g(y), \quad (10)$$

where  $g$  is the approximate function of the resistivity inversion model.

The loss function is defined as:

$$C = \frac{1}{2} \sum_{j=1}^m [a_j^L - f_j(x)]^2, \quad (11)$$

where  $a_j^L$  is the output value of the neuron  $j$  in the output layer  $L$ , that is, the predicted value,  $f_j(x)$  is the true value corresponding to neuron  $j$ ,  $m$  is the number of training samples.

The partial derivative of the loss function with respect to neuron  $j$  of the output layer  $L$  is  $\delta_j^L$ :

$$\delta_j^L = \frac{\partial C}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L} \frac{\partial a_j^L}{\partial z_j^L}. \quad (12)$$

According to formula (11), we can know:

$$\frac{\partial C}{\partial a_j^L} = a_j^L - f_j(x), \quad (13)$$

$$\frac{\partial a_j^L}{\partial z_j^L} = \xi'(z_j^L), \quad (14)$$

where  $\xi$  is the activation function. So formula (12) can be transformed into:

$$\delta_j^L = [a_j^L - f_j(x)\xi'(z_j^L)]. \quad (15)$$

The matrix form of the above formula is:

$$\delta^L = \frac{\partial C}{\partial a^L} \odot \xi'(z^L), \quad (16)$$

where  $\odot$  is the Hadamard product, which is used for point-to-point multiplication between matrices.

The partial derivative of the loss function with respect to neuron  $j$  of layer  $l$  is  $\delta_j^l$ :

$$\delta_j^l = \frac{\partial C}{\partial z_j^l} = \sum_i \frac{\partial C}{\partial z_i^{l+1}} \frac{\partial z_i^{l+1}}{\partial z_j^l}, \quad (17)$$

$$\frac{\partial z_i^{l+1}}{\partial z_j^l} = w_{ij}^{l+1} \xi'(z_j^l), \quad (18)$$

where  $w_{ij}^{l+1}$  represents the weight of neuron  $j$  in layer  $l$  to neuron  $i$  in the next layer, so formula (17) can be transformed into:

$$\delta_j^l = \sum_i w_{ij}^{l+1} \delta_i^{l+1} \xi'(z_j^l). \quad (19)$$

The matrix form of the above formula is:

$$\delta^l = (w^{l+1})^T \delta^{l+1} \odot \xi'(z^l). \quad (20)$$

The partial derivative of weights and biases for fully connected layers can be calculated as:

$$\frac{\partial C}{\partial w^l} = \frac{\partial C}{\partial z^l} \frac{\partial z^l}{\partial w^l} = \delta^l (a^{l-1})^T, \quad (21)$$

$$\frac{\partial C}{\partial b^l} = \frac{\partial C}{\partial z^l} \frac{\partial z^l}{\partial b^l} = \delta^l. \quad (22)$$

Regularization is used to prevent the network from overfitting when the weights and biases are updated.

$$w^l \leftarrow (1 - \frac{\eta \lambda}{m}) w^l - \eta \frac{\partial C}{\partial w^l}, \quad (23)$$

$$b^l \leftarrow (1 - \frac{\eta \lambda}{m}) b^l - \eta \frac{\partial C}{\partial b^l}, \quad (24)$$

where  $\eta$  is the learning rate,  $\lambda$  is the regularization coefficient.

The partial derivative of the pooling layer  $l$  is  $\delta^l$ , then  $\delta^{l-1}$  of the previous layer can be expressed as:

$$\delta^{l-1} = \text{upsample}(\delta^l). \quad (25)$$

The pooling process does not include an activation function. In this paper, the pooling method is max-pooling method.

Assuming that layer  $l$  is a convolution layer, the partial derivative  $\delta^{l-1}$  of the previous layer can be expressed as:

$$\delta^{l-1} = \frac{\partial C}{\partial z^{l-1}} = \frac{\partial C}{\partial z^l} \frac{\partial z^l}{\partial a^{l-1}} \frac{\partial a^{l-1}}{\partial z^{l-1}}. \quad (26)$$

After transformation, we can get:

$$\delta^{l-1} = \delta^l * \text{rot180}(w^l) \odot \xi'(z^{l-1}), \quad (27)$$

where  $\text{rot180}(w^l)$  is the convolution kernel that is flipped 180°. The partial derivative of weights for convolution layers can be calculated as:

$$\frac{\partial C}{\partial w^l} = \frac{\partial C}{\partial z^l} \frac{\partial z^l}{\partial w^l} = \delta^l * a^{l-1}. \quad (28)$$

The calculation of the biases  $b^l$  is also very different,  $\delta^l$  is a tensor, but  $b^l$  is a scalar(if  $\delta^l$  is a three-dimensional tensor,  $b^l$  is a vector), so  $b^l$  cannot be directly equal to  $\delta^l$ , the general practice is to The terms of each sub-matrix of  $\delta^l$  are summed separately to obtain an error vector, that is, the partial derivative of  $b^l$  can be expressed as:

$$\frac{\partial C}{\partial b^l} = \sum_{mn} \delta_{mn}^l. \quad (29)$$

The weights and biases are updated as in formulas (23) and (24).

After one training, the weights and biases increments of the convolutional layer, pooling layer and fully connected layer in the CNN can be calculated. Based on the initial weights and biases, the updated weights and biases can be calculated. Continuing to train and iterate until the value of the loss function value is less than the allowable error  $e$  given in advance. At this time, the trained neural network can approximately represent the resistivity inversion model function  $g$ .

### 3 Solution of Resistivity Inversion

#### 3.1 Algorithm

CNN imitates the construction of biological visual perception mechanism, which can perform supervised learning and unsupervised learning. CNN is a type of feed-forward neural network that includes convolution calculations and has a

deep structure, which learns samples by extracting features of the data. CNN consists of input layer, hidden layer and output layer. Among them, hidden layer includes convolutional layer, pooling layer and fully connected layer. CNN simulates the distinction of features through convolution, and reduces the magnitude of the neural network parameters through convolutional weight sharing and pooling operations. The basic steps of CNN algorithm are as follows:

**Convolution.** The purpose of convolution is to extract features. Assuming that the layer  $l$  is a convolution layer, the output  $a^l$  of the convolution layer can be expressed as:

$$z^l = \xi(w^l * a^{l-1} + b^l), \quad (30)$$

where  $a^{l-1}$  represents the output of the previous layer,  $w^l$  represents the weight of the layer  $l$ ,  $b^l$  represents the bias of the layer,  $*$  represents convolution symbol. A nonlinear excitation function is usually added after the convolution calculation.

**Pooling.** The purpose of pooling is to further simplify the output information of the convolution layer. In addition, this operation also has an anti-noise effect. Two common methods of pooling are mean-pooling and max-pooling. In this paper, the maximum pooling method is adopted, max-pooling is to use the maximum value of neurons in local receptive fields as the result of pooling calculation, which can reduce the offset error of the estimated mean caused by the convolution layer parameter error, and retain more feature texture information [14].

**Fully Connected Layer.** The output of the pooling layer will be mapped to a one-dimensional vector as the input of the fully connected layer. Assuming that layer  $l$  is a fully connected layer, the output  $a^l$  of the fully connected layer can be expressed as:

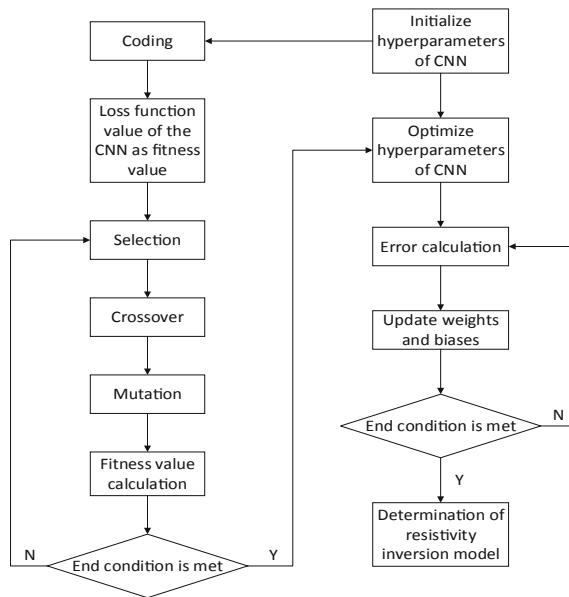
$$z^l = \xi(w^l \cdot a^{l-1} + b^l), \quad (31)$$

where  $a^{l-1}$  represents the output of the previous layer,  $w^l$  represents the weight of the layer  $l$ ,  $b^l$  represents the bias of the layer. The fully connected layer will also add an excitation function.

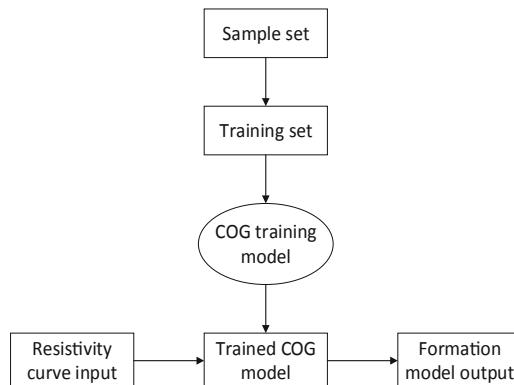
When CNN is used to solve the resistivity inversion problem, its hyperparameters have a great influence on the inversion results. Hyperparameters include the number of layers of the CNN, the number of convolution kernels in each layer, the size of the convolution kernel and pooling kernel and Learning rate [15]. In order for the algorithm to have a good inversion effect, a suitable set of hyperparameters needs to be determined. Because GA has an advantage, that is, the population in the algorithm is randomly generated, through selection, crossover and mutation operations [16, 17], so that it can effectively search the

entire solution space. Therefore, GA can be used to optimize the hyperparameters of CNN, so as to obtain a better set of hyperparameters. The loss function value of the CNN is taken as the fitness value of the individual in the GA.

The flow chart of the COG is shown in Fig. 1. The left half of Fig. 1 is the GA, the right half of Fig. 1 is the CNN.



**Fig. 1.** Flowchart of COG



**Fig. 2.** General framework for solving the resistivity inversion

### 3.2 Resistivity Inversion Through COG

The general framework for solving the resistivity inversion problem with COG is shown in Fig. 2.

First, selecting the training set from the sample set  $(Y, X)$ , where  $Y$  represents the resistivity curve,  $X$  represents the formation model corresponding to the resistivity curve. Then  $Y$  as COG input,  $X$  as COG output, training COG. Finally, The trained COG model can approximate the resistivity inversion model, which is used to solve the problem of resistivity inversion.

## 4 Numerical Simulation

### 4.1 Sample Set Generation

In this paper, the stratum is set to a three-layer H type, the stratum model is assumed to be  $[R_1, R_2, R_3, h_1, h_2]$ , the first three parameters represent the resistivity of the three layers, the latter two parameters are the layer thickness. The sample set is obtained based on the resistivity forward model. First, the Matlab program is written according to the principle of the resistivity forward model. Then  $n$  groups of formation models  $X$  are randomly generated, the corresponding forward result of the resistivity curve  $Y$  is obtained through the forward process. Finally, the two parts are combined to form the sample set  $(Y, X)$  required for resistivity inversion.

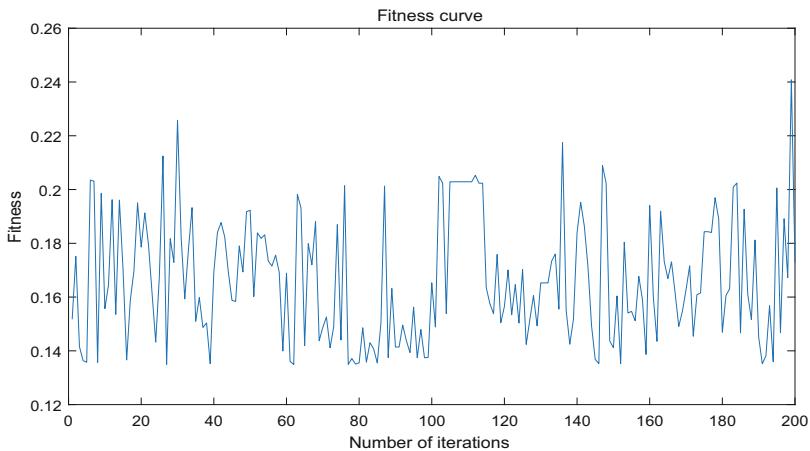
### 4.2 Results of Numerical Simulation

During the numerical simulation, two formation models were selected, namely [80, 30, 100, 100, 200] and [100, 40, 80, 200, 400]. CNN and COG algorithms are used for numerical simulation and comparison. The results of numerical simulation are as follows:

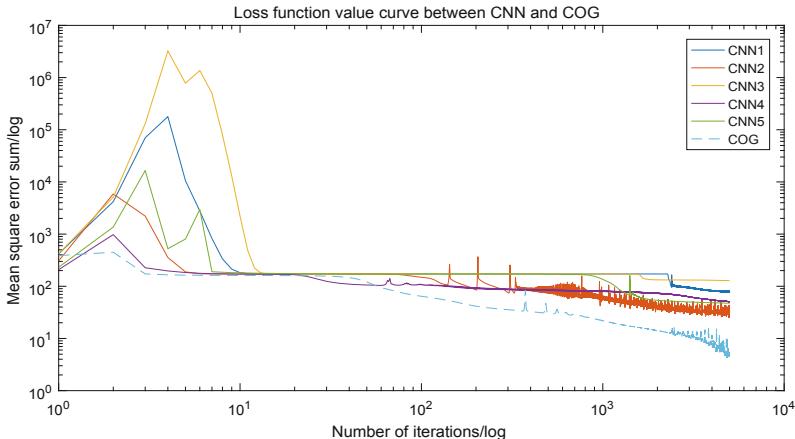
Figure 3 is the optimal individual fitness curve of COG with the formation model 1. In the 200 iterations of the algorithm, the highest fitness value of individuals is close to 0.24, that is, the inversion result of the individual corresponding to this fitness value is better. Its corresponding set of hyperparameters is used as the initial hyperparameters for CNN training.

Figure 4 is the loss function value curve of the COG with the formation model 1. The solid line represents the loss function value curve of the CNN, the dashed line represents the loss function value curve of the COG. As can be seen from the figure, the mean square error of the COG is smaller than CNN and has a faster convergence speed.

Table 1 shows the inversion results of formation model 1 and formation model 2 using CNN and COG inversion algorithms. It can be seen from the numerical results of the inversion that the relative error of the formation model obtained by COG inversion is smaller than that of CNN, that is, the inversion result of this algorithm is better.



**Fig. 3.** Optimal individuals fitness curve

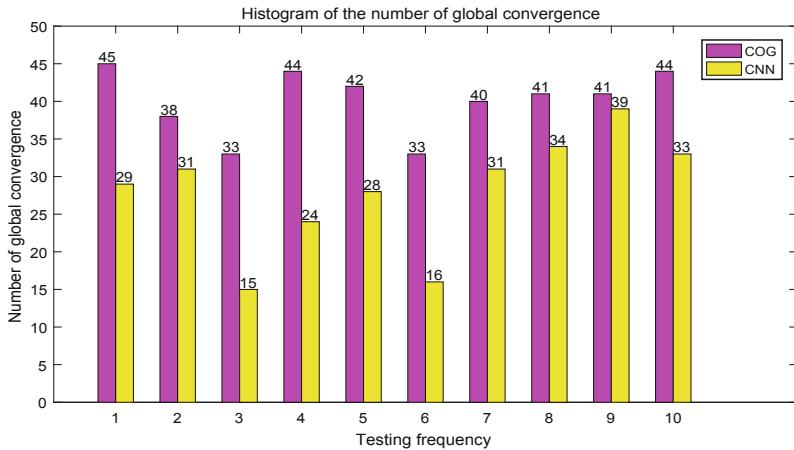


**Fig. 4.** Curve of the loss function value

Figure 5 is a comparison diagram of the global convergence between CNN and COG inversion algorithms. The two algorithms use the same training set for network training. After the training is completed, in each test, the same 50 sets of samples are used as the test set for testing, so as to obtain the number of samples where the CNN and COG algorithms converge to the global optimal solution, and the test is repeated 10 times. It can be seen from Figure 5 that the global convergence of COG is better than that of CNN, that is, the COG inversion result is easier to converge to the global optimal solution.

**Table 1.** Inversion results between CNN and COG

Model number	Parameter	Model value	CNN inversion value	COG inversion value	CNN inversion relative error/%	COG inversion relative error/%
Model 1	$R_1/(\Omega \cdot m)$	80	79.31	79.83	0.86	0.21
	$R_2/(\Omega \cdot m)$	30	29.56	29.60	1.47	1.33
	$R_3/(\Omega \cdot m)$	100	98.67	100.52	1.33	0.52
	$h_1/m$	100	102.49	100.97	2.49	0.97
	$h_2/m$	200	198.10	199.43	0.95	0.29
Model 2	$R_1/(\Omega \cdot m)$	100	97.54	99.23	2.46	0.77
	$R_2/(\Omega \cdot m)$	40	38.77	40.32	3.08	0.80
	$R_3/(\Omega \cdot m)$	80	82.28	78.99	2.85	1.26
	$h_1/m$	200	195.31	199.76	2.36	0.12
	$h_2/m$	400	393.18	401.15	1.71	0.29

**Fig. 5.** Histogram of global convergence

## 5 Conclusion

CNN optimized by GA is used to solve the resistivity inversion problem in this paper. The COG resistivity inversion is divided into three parts. The first part is that the GA is used to optimize the hyperparameters of CNN. The second part is the continuous training of CNN with optimized hyperparameters, so that the network can represent the resistivity inversion model function. The last part is that the trained CNN is used to solve the resistivity inversion problem. From the results of the inversion, it can be seen that the inversion effect of COG is better than CNN. COG not only improves the accuracy of resistivity inversion, but the inversion results are not easy to converge to the local optimal.

**Acknowledgements.** This work is supported by National Natural Science Foundation of China under Grant No. 61573378.

## References

1. Ji, Y.: Research on forward and inverse methods of electromagnetic wave resistivity logging while drilling. Master thesis, Xi'an Shiyou University, August 2019
2. Song, W., Liu, Y., Ge, F., Lu, C., Tian, D.: Resistivity inversion and its application. CT Theory Appl. Res. **24**, 377–382 (2015)
3. Feng, D., Wang, P., Yang, B.: Ultra-high density electrical method finite element method forward and generalized least squares inversion. Chin. J. Nonferrous Met. **24**, 793–800 (2014)
4. Qiu, G., Zhang, X., Pei, F., Yuan, Y., Bai, D., Zhang, P.: Comparative experiment on effectiveness of magnetotelluric sounding inversion technology. Geophys. Geochem. Explor., 118–124 (2015)
5. Loke, M.H., Wilkinson, P.B., Chambers, J.E.: 3-D resistivity inversion with electrodes displacements. ASEG Ext. Abstr. **2016**(1), 1–5 (2016)
6. Zhou, L., Ji, W., Qu, J.: Research on adaptive regularization inversion. Technol. Innov. Appl., 42–43 (2017)
7. Heriyanto, M., Srigutomo, W.: 1-D DC resistivity inversion using singular value decomposition and Levenberg-Marquardt's inversion schemes, vol. 877, no. 1 (2017)
8. Wang, H., Jiang, H., Wang, L., Xi, Z., Zhang, D.: Inversion of magnetotelluric artificial neural network. J. Central South Univ. (Nat. Sci. Ed.), 1707–1714 (2015)
9. Barboza, F.M., Medeiros, W.E., Santana, J.M.: A user-driven feedback approach for 2D DC-resistivity inversion based on PSO. Geophysics, 1–78 (2018)
10. Cheng, B., Xiong, B.: Application of improved genetic algorithm in electrical sounding inversion. Mineral Resourc. Geol. **32**, 127–130 (2018)
11. Zhou, H.: Particle swarm inversion of resistivity sounding data. Hongshuihe **38**, 84–87, 91 (2019)
12. Li, Y., Yuan, H., Yu, J., Zhang, G., Liu, K.: A review of the application of genetic algorithms in optimization problems. Shandong Ind. Technol. **180**(5), 242–243 (2019)
13. Li, H., Yan, Z., Liu, C., Jiang, Y.: Numerical simulation of azimuth resistivity logging tool response while drilling. J. China Univ. Petrol. (Nat. Sci. Ed.) **43**, 42–52 (2019)
14. Chen, A., Xia, J., Chen, Y., Tang, L.: Research on visibility inversion technology based on digital image. Comput. Simul. **35**, 252–256 (2018)
15. Neng, C., Sun, X., Xu, Y., Liu, J., Dong, L., Liu, Y.: Nonlinear inversion method of site pollution based on deep convolutional neural network. China Environ. Sci. (12) (2019)
16. Sun, C., Li, L., Wang, X., Huang, W., Zhou, F.: CSAMT one-dimensional inversion of improved genetic algorithm. J. Hunan Univ. (Nat. Sci. Ed.) **44**, 102–108 (2017)
17. Liu, B., Wang, X., Zhang, C.: Inversion calculation of seabed formation parameters based on improved genetic algorithm. Acoust. Technol. **36**, 210–216 (2017)



# GECNN-CRF for Prostate Cancer Detection with WSI

Jinfeng Dong<sup>1</sup>, Xuemei Guo<sup>1</sup>, and Guoli Wang<sup>1,2(✉)</sup>

<sup>1</sup> School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China  
isswgl@mail.sysu.edu.cn

<sup>2</sup> Key Laboratory of Machine Intelligence and Advanced Computing,  
Ministry of Education, Guangzhou, China

**Abstract.** Pathological examination is of great significance for the diagnosis and treatment of prostate cancer, but artificial pathological examination is time-consuming, laborious and error prone. Automatic pathological examination can assist doctors in diagnosis and treatment. In this paper, a deep learning convolution network based on group equivariant convolution and conditional random field is proposed, which solves the problem of inconsistent features of data after random rotation in the learning process and enhances the robustness of the network. On the other hand, the conditional random field method is used to produce the same result on the slice image, which makes the segmentation edge of the probability image clearer.

**Keywords:** WSI detection · Prostate cancer · Group equivariant convolution · Fully connected condition random field

## 1 Introduction

Prostate cancer is the leading cause of morbidity and mortality among men in the United States and Western Europe. The diagnosis of prostate cancer mainly includes digital rectal examination, prostate specific antigen (PSA) concentration examination, imaging diagnosis and prostate biopsy. However, biopsy is still the most accurate diagnostic method for prostate cancer. Through pathological examination, we can judge the origin and differentiation degree of tumor, so as to assist the pathological diagnosis and differential diagnosis of tumor. Further pathological examination of the postoperative mass can diagnose whether the mass is benign or malignant. If it is benign, surgical resection of the mass will achieve the purpose of cure. If it is malignant, surgical resection of the mass alone is not enough, and further treatment is needed. Doctors differentiate the normal cells from the tumor cells and localize the cancer area according to Whole-slide Images (WSIs) [19]. The artificial pathological examination is easily affected by various subjective factors, such as the pathologist's clinical diagnosis experience, WSI image resolution and so on. Therefore, the prostate cancer analysis method based on expert diagnosis still has considerable limitations. Computer-aided pathological examination can improve the sensitivity,

speed and consistency of prostate cancer detection. The contributions of our work are as follows:

- We propose a group equivariant convolution model for the cancer detection task in WSI images. Compared with the standard convolution model, the group equivariant convolution model can theoretically guarantee the equivariant of rotation and symmetry.
- We propose to use fully connected conditional random field method to enhance the performance of group equivariant convolution. This method can combine the label correlation between different blocks in the image to get better detection results.
- We propose methods including dropout and data jitter to enhance the generalization performance of group equivariant convolutional networks for the cancer detection task in WSI images.

## 2 Related Work

In recent years with the development of deep learning technology, CNN has shown excellent performance in the field of computer vision. At present, CNN has made great improvements in the tasks of computer vision on natural images, such as image classification [6, 8, 16, 18], object detection [5, 14, 15], object tracking [1, 23], image segmentation [2, 12, 13, 24]. Accordingly, more and more people begin to study how to apply CNN to analyze medical images [7, 9, 11, 20, 21]. Because of the super-high resolution of pathology images, it is impossible to localizes the cancer area of pathology images at pixel level, so the classification strategy based on image patch is adopted [7, 9, 11, 21]. Here some people proposed to preprocess WSIs to get small patches for training the deep model. Then they train a deep CNN model to divide these patches into normal or tumor patches. In 2013, [3] proposed to use convolutional neural network to detect cancer areas in breast histopathology. In 2014, [22] proposed a method of pathological image detection method based on the combination of cascaded neural network model and manual feature extraction method. In 2015, [17] proposed a method of pathological detection based on the combination of multi-scale convolutional neural network and graph matching. In 2018, [9] proposed a conditional random field method to enhance the accuracy of WSI cancer detection. All of these methods use the convolution model composed of standard convolution to detect the cancer area in WSI image, so these methods can not guarantee the equivariant feature layer for the possible rotation changes in WSI image, which weakens the detection performance of the model. One way to enhance the model to output consistent labels is to enhance the data. As a data enhance method, random rotation and flipping are widely used in image data processing for training better CNN model [10, 11]. But for the standard convolution, the data augment method of random rotation and flipping can not guarantee the feature results extracted from original image of the possible rotation and flipping have corresponding transformation. In other words, the standard convolution is not equivariant for rotation and flipping of input data. In [4], a group

equivariant convolution method is proposed to solve this problem, which is different from the method in [15] in which the input data is rotated eight times and flipped for averaging. This method makes the output feature map and the input data remain equivariant by rotating and flipping the convolution kernel. In this paper, we will use the group equivariant convolution method proposed in [4] to build the group equivariant convolution model, and on the basis of this model, add the conditional random field method proposed in [9] to get the pathological diagnosis model of prostate tissue sections.

### 3 Method

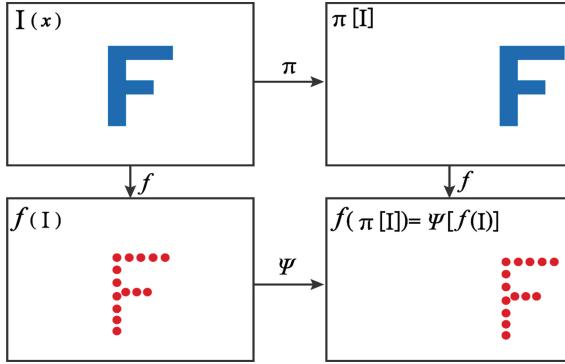
In this chapter, we will explain our detection method, Group Equivariant Convolution Neural Network Model with Condition Random Field(GECNN-CRF), of cancer in prostate WSI images. This method consists of two parts: the group equivariant residual convolution network model and the conditional random field method.

#### 3.1 Group Equivariant Convolution Network Model

This paper will be divided into two parts to introduce the group equivariant convolution neural network model, which are the mathematical principle part and the specific implementation part.

**Group Equivariant Convolution Mathematical Principle.** Group equivariant convolution is a new convolution structure different from standard convolution. Before the introduction of group equivariant convolution, the related concepts are first introduced. The first concept introduced is the equivariant of transformation. The equivariant and invariance of the transformation is the property that the change of the features of the input image changes correspondingly and remains unchanged after some transformation. For instance, one of the important features of the standard convolution neural network model is the weight sharing, which is to use a small convolution kernel to convolute the input image with a sliding window to obtain a feature map. If the input image is translated, the resulting feature map will also be translated. Suppose that the transformation of translation forms a set  $S = \{(n, m) | n \in Z, m \in Z\}$ , where  $n$  and  $m$  represent the transformation on the X axis and the transformation on the Y axis respectively. For any  $h, g \in S$ , there exists  $h + g \in S$ . That is,  $S$  forms a group for the add operation. Translation transformation makes the convoluted feature map shift correspondingly, but the value of the feature map itself does not change. In other words, the standard convolution method is equivariant for the translation (Fig. 1).

The above paragraph shows the equivariant of the standard convolution process for translation. The cancer area may appear in any possible direction in WSI images. We hope that the model can have equivariant rotation for any tissue slice, so as to obtain consistent label prediction. Therefore, we propose a method to



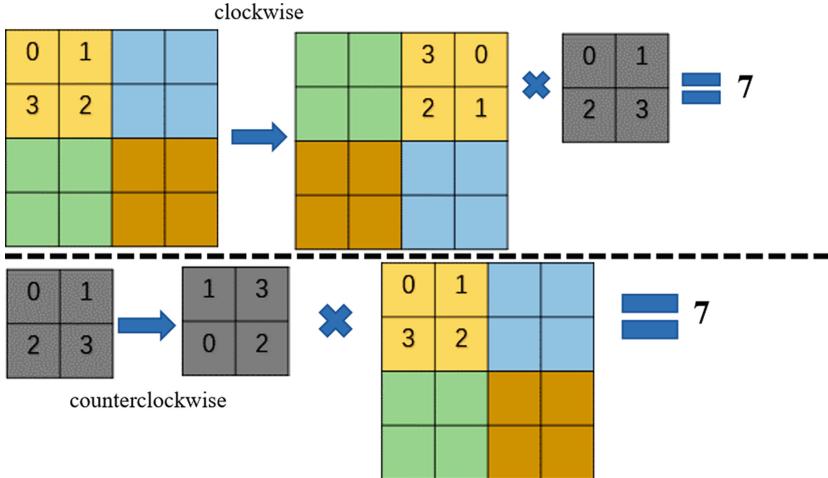
**Fig. 1.** Diagram of translation equivariant in standard convolution layer.  $I(x)$  represents the original image,  $\pi[I]$  represents the translation to the original image,  $f(I)$  represents the extracted feature from origin image through the function  $f$  and  $f(\pi[I])$  represents the corresponding extracted feature from translated image  $\pi[I]$ .

replace the standard convolution with the group equivariant convolution in the task of cancer detection in WSI images. It's equivariant for rotations of 0, 90, 180 and 270° and we will show that such a group equivariant convolution method is superior to the ordinary standard convolution method. For group equivariant convolution, the feature layer can be divided into three forms according to different convolution operations:  $Z^2$  group feature layer, P4 group feature layer and P4M group feature layer. The  $Z^2$  feature layer is the feature layer extracted from standard convolution model which is equivariant for the translated transformation. For P4 group feature layer, it is formed by group equivariant convolution that the convolution kernel  $i$  defined to represent a group formed by 90° rotation of a square region. The operations matrix of P4 group can be parameterized by three parameters ( $r, b, c$ ):

$$h(r, b, c) = \begin{bmatrix} \cos(\frac{r\pi}{2}) - \sin(\frac{r\pi}{2}) & b \\ \sin(\frac{r\pi}{2}) & \cos(\frac{r\pi}{2}) & c \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

where  $r \in \{0, 1, 2, 3\}$  and  $(b, c) \in Z^2$ . The group operation of this group is matrix multiplication. In Eq. 1, the convolution kernel is rotated and translated., and the corresponding inverse transformation can also be given by this formula. First we define a pairwise set  $S_0 = \{(n, m) | 0 \leq n < \text{height}, 0 \leq m < \text{width}, k > 0, n \in Z, m \in Z, k \in Z\}$  where height and width denote the height and width of input image. Then for a coordinate point  $(m, n)$  in the image, we can get a new coordinate point  $(m', n')$  after applying the one specific rotation operation (Fig. 2):

$$q = \begin{bmatrix} n' \\ m' \\ 1 \end{bmatrix} = hp = \begin{bmatrix} \cos(\frac{r\pi}{2}) & -\sin(\frac{r\pi}{2}) & 0 \\ \sin(\frac{r\pi}{2}) & \cos(\frac{r\pi}{2}) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} n \\ m \\ 1 \end{bmatrix} \quad (2)$$

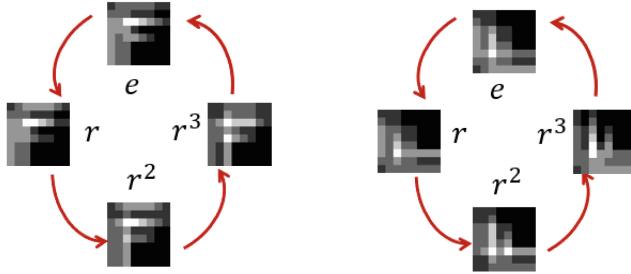


**Fig. 2.** The convolution of the input image after rotation is the same as that of the convolution kernel after inverse rotation.

Based on formula (2), we can get a new pairwise set  $S_1 = \{q|q = hp\}$  and  $h$  represents 4 types of rotation operations:  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$  and  $270^\circ$ . Finally we can get 4 images from original input by applying above operation. All these pictures can be called P4 group feature map, Because if we use some rotation operation on the input picture at the same time, it is equivalent to a new arrangement and combination of the original set of four feature maps, and the corresponding values are unchanged, so these four feature graphs are called P4 group feature maps. In order to simplify the operation, the steps of rotating input image can be changed into rotation convolution kernel and then convolution, because the results of the two are equivalent. Then based on P4 group equivariant convolution, we add symmetry operation to rotation operation to get rotation symmetry group convolution. The operation of rotation symmetry can be written by:

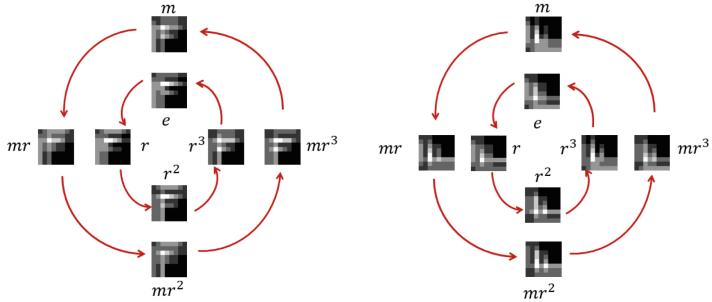
$$h(r, b, c, m) = \begin{bmatrix} (-1)^m \cos(\frac{r\pi}{2}) & (-1)^{m+1} \sin(\frac{r\pi}{2}) & b \\ \sin(\frac{r\pi}{2}) & \cos(\frac{r\pi}{2}) & c \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

where the parameters  $r$ ,  $b$  and  $c$  are defined the same as formula (2), and  $m$  indicates whether to flip. By using the convolution kernel rotation flipping operation similar to P4 group equivariant convolution, we can get the corresponding P4M group equivariant convolution feature maps. Now we get three types of group equivariant convolution feature maps:  $Z^2$  group feature maps, P4 group feature maps and P4M group feature maps, which are based on standard convolution, P4 group equivariant convolution and P4M group equivariant convolution (Figs. 3 and 4).



(a) original group equivariant feature maps (b) group equivariant feature maps after clockwise roatation

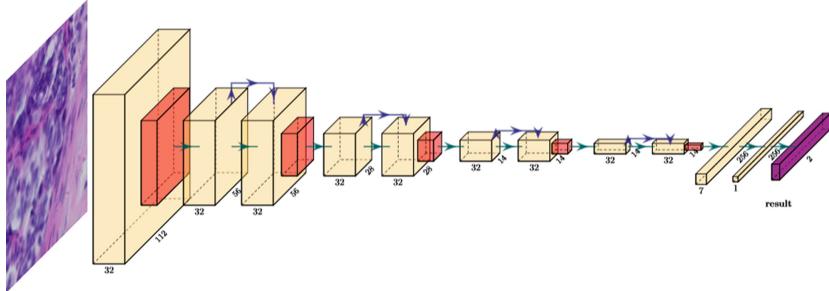
**Fig. 3.** P4 group equivariant feature maps



(a) original group equivariant feature maps (b) group equivariant feature maps after clockwise roatation

**Fig. 4.** P4M group equivariant feature maps

**The Implementation of Group Equivariant Convolution Model.** This section introduces the implementation of group equivariant convolution network model. The original network model used in this paper is Resnet network model. By replacing the standard convolution in the residual network model with the corresponding group equivariant convolution, we can get the corresponding group equivariant convolution model, such as P4 group equivariant convolution network model, P4M group equivariant convolution network model. Specifically, the group equivariant convolution model and the related baseline model structure used in this paper are shown in Table 1. In order to ensure that the models involved in the comparison have similar parameters, this paper reduces the number of convolution kernels for the group equivariant convolution model (Fig. 5).



**Fig. 5.** The group equivariant convolution network model. The group equivariant convolution network model.

**Table 1.** The standard convolution model and the group equivariant convolution model structure

Name	Output	Resnet-18	Resnet-50	Resnet-GE-P4	Resnet-GE-P4M
conv1	112 × 112	7 × 7, stride 2			
conv2_x	56 × 56	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 2$	
conv3_x	28 × 28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 2$	
conv4_x	14 × 14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 2$	
conv5_x	7 × 7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 6$	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 2$	
FC	1 × 1		Avgpool		
FLOPS	$1.8 \times 10^9$	$3.5 \times 10^9$	$0.6 \times 10^9$	$2.5 \times 10^9$	

### 3.2 Fully Connected Condition Random Field

In this section, we introduce the full connection conditional random field method. Because normal tissue and cancer area often have clear boundaries, therefore the labels of image blocks between normal tissue and cancer area tissue are correlated. In this paper, fully connected conditional random field method is introduced to improve the performance of cancer detection by using the label correlation between the image blocks. First of all, we denote the block a small grid consisting of several small images, which refers to the small slice image from input that to the network. One block represents a square area, and the size of the patch can be taken as  $3 \times 3$ . Define  $x = x_{i=1}^N$  represent the block

**Algorithm 1.** Mean field arrpoximate inference algorithm**Input:**

initialize variable  $Y = \{y_1, y_2, \dots, y_n\}$   
 initialize variable  $Q_i = \psi_u(y_i)$   
 initialize variablet  $\leftarrow 0$

**Output:** output  $Q_1 Q_2, \dots, Q_i$ 

- 1: **for**  $t < T$  **do**
  - 2:   compute  $\ln \tilde{P} = -(\sum_{i=1}^n \psi_u(y_i) + \sum_{i=1}^n \sum_{j=1}^n \phi_p(y_i, y_j))$
  - 3:    $\forall i \in \{1, 2, \dots, n\}$ , compute  $Q_i = \exp(E_{-Q_i}[\ln \tilde{P}])$
  - 4:    $\forall i \in \{1, 2, \dots, n\}$ , normlize  $Q_i$
  - 5: **end for**
  - 6: **return**  $Q$
- 

items. Define  $y = y_{i=1}^N$  as the corresponding labels of the block, where N is the number of items within the block. The distribution  $P\{y|x\}$  can be regarded as a conditional distribution:

$$P(Y = y|x) = \frac{e^{-E(y,x)}}{Z(x)} \quad (4)$$

In the formula (4),  $E(y, x)$  is a energy function given specific  $y$  and  $x$ .  $Z(x)$  represents a normalized function, which is used to ensure that the obtained value satisfies the probability distribution. The energy function can be defined as:

$$E(y, x) = \sum_i \psi_u(y_i) + \sum_{i < j} \phi_p(y_i, y_j) \quad (5)$$

where i, j ranges from 1 to N. In (5)  $\psi_u$  stands for the unitary energy function and the specific meaning here is the output probability of CNN model for the patch and  $\phi_p$  denotes the pairwise loss function, which is used to express the magnitude of the correlation between the two blocks. When the feature of the two blocks are similar, the output label should also be similar. The pairwise energy function can be defined as:

$$\phi_p(y_i, y_j) = I(y_i = y_j) \cdot w_{i,j} \cdot \left(1 - \frac{f(x_i)f(x_j)}{|f(x_i)||f(x_j)|}\right) \quad (6)$$

where  $I(y_i = y_j)$  is the indicator that ensure the label compatibility between  $y_i, y_j$  and  $w_{i,j}$  is a trainable weight that is used to represent the correlation between two blocks i and j in a patch,  $f(x_i), f(x_j)$  represents the feature maps extracted from input  $x_i, x_j$  and the corresponding parameter  $w_{i,j}$  is also used to encode spatial correlation. In order to train the G-CNN and rotated CRF end-to-end, we need to get the marginal distribution of Y, and we can use the cross-entropy loss to train the parameters. However, the real marginal distribution cannot be determined directly, so we need an algorithm to determine such a distribution. We use mean field approximate inference to solve this problem. The specific process is shown in Algorithm 1. Through the iterative iteration

of mean field algorithm method, we can get the edge distribution of each label  $Q_i$  ( $y_i$ ) to calculate the cross entropy loss, and then train the whole network with back-propagation algorithm.

## 4 Experiment

To evaluate our proposed model, we collected and labeled a prostate WSI dataset, which contains 116 H&E stained WSIs of prostate tissue split into 93 slides with pixel-level annotations for training and 23 slides with pixel-level annotations for testing.

### 4.1 Preprocessing

For the evaluation on the WSI-level prostate slides dataset, we uniformly sampling WSIs and extract tumor/normal blocks with equal probability. Each block has  $3 \times 3$  images and each image has  $256 \times 256$  pixels. Therefore each block has  $768 \times 768$  pixels. Finally, 102,502 training blocks are obtained by sampling in this data set, which includes 922,518 images, 27,082 test blocks which includes 243,738 images.

### 4.2 Result

In order to prove the effectiveness of the proposed method, we trained Resnet-18, Resnet-50, Resnet-GE-p4, Resnet-GE-p4m models with the same super parameters. Among them, the learning rate is 1e-2, and the input picture size is  $768 \times 768$ , batch size is 4. Without using any data enhancement methods, the results are shown in Table 2. In Table 2, The accuracy rate can be expressed by the formula:  $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ . Under certain threshold conditions, the test set can be divided to class normal or class tumor. If the label of a image in the test set is tumor and the predicted label is tumor, then the image is recorded as a true positive (TP). Similarly, if the label of a image is normal and the predicted label is normal, the image will be recorded as a true negative example (TN). If the real label of a image does not match the predicted label, it will be recorded as a false positive example (FP) and a false negative example (FN). AUC (area under curve) is defined as the area enclosed by the coordinate axis under ROC (Receiver Operating Characteristic) curve. As can be seen from Table 2, Resnet-GE-P4M+CRF model has obtained the best experimental results. The method of using CRF is generally better than that of not using CRF. Further, this paper tests more optimization methods, including data enhancement methods and dropout method. For data enhancement methods, we use color jitter and random flipping, rotation for data enhancement. The lightness parameter of color jitter is 0.25, the brightness parameter is 0.04, and the saturation parameter is 0.25. For dropout method, we replace the last one fully connected layer of the model with the three layers fully connected layer, and add the dropout layer to the hidden layer, respectively, taking the dropout rate of 0, 0.3, 0.7 for

**Table 2.** The comparison between group equivariant convolution model and standard convolution model.

Model	FLOPS	Accuracy(%)@threshold = 0.5	AUC(%)
Resnet-18	$1.8 \times 10^9$	77.9	85.5
Resnet-18+CRF	$1.8 \times 10^9$	81.4	88.3
Resnet-50	$3.5 \times 10^9$	82.0	89.0
Resnet-50+CRF	$3.5 \times 10^9$	82.3	89.1
Resnet-GE-P4	$0.6 \times 10^9$	78.3	85.2
Resnet-GE-P4M	$2.5 \times 10^9$	79.1	86.4
Resnet-GE-P4M+CRF	$2.5 \times 10^9$	<b>83.7</b>	<b>89.7</b>

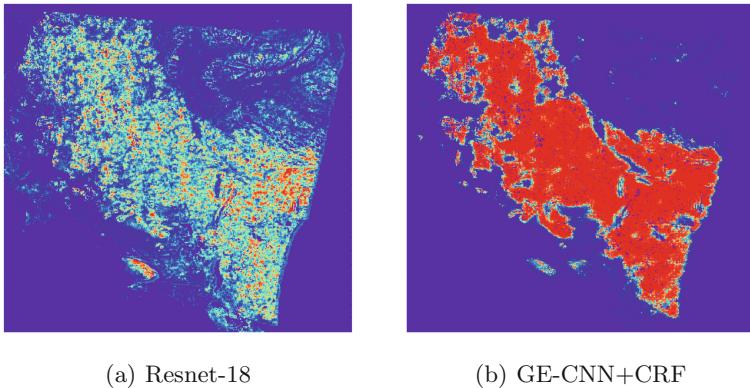
**Table 3.** The influence of data augment and dropout on the model.

Model	Accuracy(%)@threshold = 0.5	AUC(%)
Resnet-GE-P4M+CRF	83.7	89.7
Resnet-GE-P4M+CRF+AUG+D0	84.8	90.3
Resnet-GE-P4M+CRF+AUG+D0.3	84.2	90.2
Resnet-GE-P4M+CRF+AUG+D0.7	<b>85.1</b>	<b>91.0</b>



**Fig. 6.** The test WSI image and its labeled cancer area (within the blue solid line area)

experiments. The experimental results are shown in Table 3. In Table 3, **AUG** represents that the corresponding model use data augment method, **D** represents that he corresponding model use dropout method, the following number



**Fig. 7.** A comparison of the probabilistic heat maps generated by model Resnet-18 and model GECNN-CRF

indicates the dropout rate. As can be seen from Table 3, When the model uses data amplification, the performance of the model is improved. Especially, when the dropout rate is 0.7, the performance of the model reaches the optimal level. Now, we call the Resnet-GE-P4M+CRF+AUG+D0.7 model as the GECNN-CRF model (Fig. 6).

Finally, we use the visual method to test a WSI in the test set, and generates the probability heat map of cancer area. The results are shown in Fig. 7. As can be seen from Fig. 7, our proposed method, GECNN-CRF model, is much better than the baseline model, Resnet-18.

## 5 Conclusion

We have introduced our proposed method, GECNN-CRF model, which can generates group equivariant features to more corrected labels. We will further study the group equivariant convolution theory, and seek for more transformations, such as equivariant lines of any angle, so as to improve the detection performance of prostate WSI images.

**Acknowledgement.** This work was supported in part by the Key Program of the National Social Science Fund of China with Grant No. 18ZDA308 and in part by the National Natural Science Foundation of China under Grant Nos. 61772574.

## References

- Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: Fully-convolutional Siamese networks for object tracking. In: European Conference on Computer Vision, pp. 850–865 (2016)

2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2018)
3. Cirean, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2013)
4. Cohen, T.S., Welling, M.: Group equivariant convolutional networks. In: International Conference on Machine Learning, vol. 48, pp. 2990–2999 (2016)
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Hou, L., Samaras, D., Kurc, T., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2424–2433 (2016)
8. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 25 (2012)
9. Li, Y., Ping, W.: Cancer metastasis detection with neural conditional random field. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
10. Litjens, G.J.S., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Der Laak, J.A.W.M.V., Van Ginneken, B., Sanchez, C.I.: A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017)
11. Liu, Y., Gadepalli, K.K., Norouzi, M., Dahl, G.E., Kohlberger, T., Venugopalan, S., Boyko, A.S., Timofeev, A., Nelson, P.Q., Corrado, G.S., et al.: Detecting cancer metastases on gigapixel pathology images. [arXiv:1703.02442](https://arxiv.org/abs/1703.02442) (2017)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. [arXiv:1806.07064](https://arxiv.org/abs/1806.07064), pp. 3431–3440 (2015)
13. Mehta, S., Rastegari, M., Caspi, A., Shapiro, L.G., Hajishirzi, H.: ESPNet: efficient spatial pyramid of dilated convolutions for semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 552–568 (2018)
14. Redmon, J., Divvala, S.K., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556), September 2014
17. Song, Y., Zhang, L., Chen, S., Ni, D., Lei, B., Wang, T.: Accurate segmentation of cervical cytoplasm and nuclei based on multi-scale convolutional network and graph partitioning. *IEEE Trans. Biomed. Eng.* **62**, 2421–2433 (2015)
18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)

19. Teresa, A., Guilherme, A., Eduardo, C., José, R., Paulo, A., Catarina, E., António, P., Aurélio, C., Anna, S.: Classification of breast cancer histology images using convolutional neural networks. *Plos One* **12**, e0177544 (2017)
20. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T.S., Welling, M.: Rotation equivariant CNNs for digital pathology. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 210–218 (2018)
21. Wang, D., Khosla, A., Gargya, R., Irshad, H., Beck, A.H.: Deep learning for identifying metastatic breast cancer. In: Quantitative Methods (2016)
22. Wang, H., Cruzroa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., Tomaszewski, J.E., Gonzalez, F.A., Madabhushi, A.: Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *J. Med. Imaging* **1**, 034003 (2014)
23. Wang, N., Yeung, D.: Learning a deep compact image representation for visual tracking. In: Advances in Neural Information Processing Systems, pp. 809–817 (2013)
24. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations (2015)



# Design Method of Robot Welding Workstation Based on Adaptive Planing

Haofei Dai<sup>1</sup>, Zhaojiang Liu<sup>1</sup>, Yizhong Luan<sup>2</sup>, Jiyang Chen<sup>2</sup>, Wenxu Sun<sup>1</sup>, and Sile Ma<sup>1</sup>(✉)

<sup>1</sup> School of Control Science and Engineering, Shandong University,  
Jinan 250061, Shandong, China  
[masile@sdu.edu.cn](mailto:masile@sdu.edu.cn)

<sup>2</sup> Institute of Marine Science and Technology, Shandong University,  
Qingdao 266237, Shandong, China

**Abstract.** Robot welding is an important mechanical processing method in the production of modern welding industry, especially when manufacturing thick plate structural parts, robot welding plays an important role. In view of the current problems of low efficiency and low automation in the application of robot welding in multi-layer multi-path welding of thick plate structural parts, this paper proposes a design method of robot welding workstation based on adaptive planning. We integrate the laser vision system, data processing and monitoring system, and robot welding system into a whole robot welding workstation. At the same time, the control program of the welding robot was written based on the method of adaptive planning. Using this workstation, several samples of thick plate structural parts with different specifications were experimentally welded. The results show that this method is very practical for multi-layer multi-path welding of thick plate structural parts.

**Keywords:** Robot welding · Welding workstation · Multi-layer multi-path · Adaptive planning

## 1 Introduction

The traditional manual welding has high labor intensity, poor welding quality and low efficiency. Therefore, welding robots with the advantages of stable performance, high welding quality, and high production efficiency are widely used in modern manufacturing [1]. Put the welding robot into the production line and configure the welding robot workstation with corresponding peripheral equipment, which can drive the development of welding automation [2]. Robot welding workstation is a flexible processing system based on robots [3], mainly to realize the automation, information and intelligence of industrial product welding. At present, robot welding workstations have been widely used in various welding fields [4].

Due to ever increasing precision and automation demands in robotic welding [5], the relevant researchers have proposed some design and implementation solutions of robot welding workstations with different characteristics. In article [1], a design method of robot welding workstation with K-shaped positioner is proposed. This method uses the positioner to realize the dual-station welding application of the robot, which can improve the welding efficiency of the robot. However, the actual application lacks the visual system and the monitoring system, and the overall intelligence of the workstation is not high. In article [2], a design method of robot welding workstation based on welding seam tracking system is proposed. This method uses welding seam tracking system to realize high-precision welding of robot workstation, but manual intervention is needed to correct deviations during welding. The degree of automation is greatly restricted. Article [6] proposes a design method of robot welding workstation for thick-walled pipelines. This method achieves multi-layer multi-path welding of thick plates by optimizing welding strategies. However, in the welding process, robot teaching programming needs to be performed path by path, and the overall work efficiency of the workstation is low. Article [7] proposes a design method of robot welding workstation based on PLC control. This method uses PLC as the overall control system and effectively realizes real-time data interaction between the various parts of the workstation. However, during the welding process, manual teaching is still required to plan all welding paths, and the workstation cannot run autonomously throughout the process. Article [8] proposed a design method of robot welding workstation based on information monitoring system. This method integrates a variety of high-level intelligent equipment, and the workstation has a higher degree of intelligence. However, in practical applications, the problem of high cost of this method is very prominent, and it is difficult for the majority of small and medium-sized enterprises to bear, which has affected the widespread promotion of this method. There are still many problems in field welding that need to be solved [9].

In view of the advantages and disadvantages of the existing design methods of robot welding workstations, this paper proposes a design method of robot welding workstations based on adaptive planning. The innovations of this article mainly have the following two points: 1. Integrate the laser vision system, data processing and monitoring system, and robot welding system into a whole robot welding workstation. While the system has a high degree of intelligence, the complexity and cost are moderate, it is easy to be applied in actual production, and it is suitable for large-scale promotion. 2. Based on the method of adaptive planning, the welding robot control program is written, which gives the welding robot a high degree of self-adaptive ability, enabling the workstation to realize automatic welding in the whole process. The system has high production efficiency and high degree of automation.

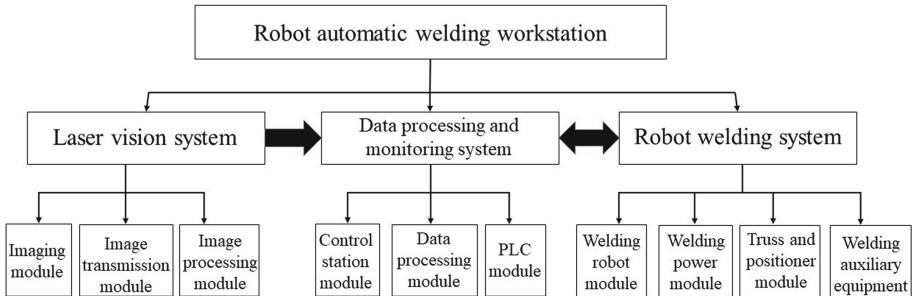
Finally, according to the overall system design scheme and adaptive planning method proposed in this paper, a robot automatic welding workstation is designed and implemented. A variety of samples of thick plate structural parts with different specifications were selected for experimental welding. The results

show that this workstation has a good welding effect for many structural parts with different specifications, and can quickly and stably complete automated welding work. The welding efficiency of this workstation is significantly higher than manual welding, which has a strong practicality.

## 2 Overall System Design

### 2.1 System Structure Framework

The robot welding workstation based on adaptive planning proposed in this paper is mainly composed of three parts: laser vision system, data processing and monitoring system, and robot welding system (Fig. 1).



**Fig. 1.** System structure framework

### 2.2 System Hardware and Software

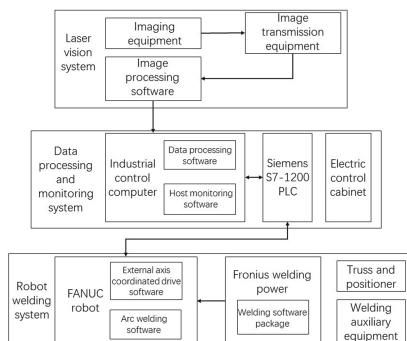
The hardware part of the laser vision system includes imaging equipment and image transmission equipment. Its main function is to collect real-time images of the weld seam contours and transmit them to the image processing module for processing; The software part includes image processing software. The main function is to process the real-time image of the welding seam contour to obtain the coordinates of several important feature points on the real-time contour of the welding seam.

The hardware part of the data processing and monitoring system includes the electric control cabinet, industrial control computer and Siemens S7-1200 PLC. The main function of the electric control cabinet is to supply power to the equipment of each part of the workstation and have the function of controlling the start and stop of the equipment; industrial control computer is mainly used to install data processing software and monitoring software; The main function of Siemens S7-1200 PLC is to provide hardware support for monitoring the running status of the system and achieving data interaction.

The software part of the data processing and monitoring system includes data processing software and monitoring software, which are installed on the industrial computer. The data processing software is developed based on python, and its main function is to calculate various weld contour information according to the coordinates of the feature points given by the laser vision system. The monitoring software is developed based on TIA Portal V15.0, which is mainly used to monitor the running status of each part of the system and realize data interaction with the robot.

The hardware part of the robot welding system includes a 6-axis FANUC robot model M-10iA, truss and positioner, Fronius TPS5000 welding power and welding auxiliary equipment. The FANUC robot is equipped with supporting welding equipment, which can perform welding operations according to the robot control program. The FANUC robot is mounted upside down on the truss, and the parts to be welded are fixed on the positioner. Five FANUC servo motors are installed on the truss and positioner, which gives the welding robot a high degree of freedom in the welding process. The Fronius TPS5000 welding power provides continuous and stable welding current as required during the robot welding process. Welding auxiliary equipment includes welding protective gas cylinders, smoke exhaust and dust removal equipment, etc. It is mainly used to provide welding protective gas, clean welding fumes, and assist welding operations.

The software part of the robot welding system includes the robot arc welding software package and the external axis coordinated drive software package in the FANUC robot controller and the welding software package in the Fronius TPS5000 welding power. The robot arc welding software package provides many convenient operation instructions for the FANUC robot to perform welding operations, and provides a platform for writing robot welding programs. The external axis coordinated drive software package is used to drive and control the five FANUC servo motors installed on the truss and positioner, so that the truss and positioner can coordinate with the FANUC robot. The welding software package is used to intelligently and dynamically adjust the welding parameters to an optimal level during the welding process (Fig. 2).

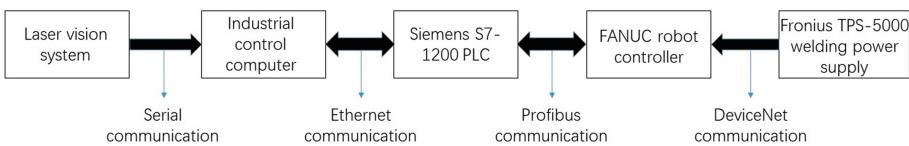


**Fig. 2.** System hardware and software

### 2.3 Communication Connection

The equipments involved in the communication connection mainly include five parts: laser vision system, industrial control computer, Siemens S7-1200 PLC, FANUC robot controller and Fronius TPS-5000 welding power supply.

Serial communication is used between the laser vision system and the industrial control computer. Ethernet communication is adopted between the industrial control computer and Siemens S7-1200 PLC. Profibus communication is used between Siemens S7-1200 PLC and FANUC robot controller. DeviceNet communication is used between the FANUC robot controller and the Fronius TPS-5000 welding power supply (Fig. 3).



**Fig. 3.** Communication connection

## 3 System Control Strategy

### 3.1 Workflow of the Workstation

The overall workflow of the workstation is mainly divided into the following steps:

Step 1: System preparation stage. Before running the system, you first need to fix the workpiece to be welded on the positioner using a fixture, and run the positioner until the workpiece reaches a suitable position. Teach a fixed scanning position P [1] at a suitable position above the workpiece, and then use the vision system to automatically calibrate the position of the robot so that the X axis of the robot tool coordinate system is parallel to the weld. Finally, perform a self-test on each part of the workstation and confirm the correctness before continuing.

Step 2: The robot performs the first layer welding at the bottom of the weld.

Step 3: The robot returns to the fixed scanning position P [1], the visual system scans the weld in real time, and transmits the real-time contour data to the robot through the PLC as a bridge.

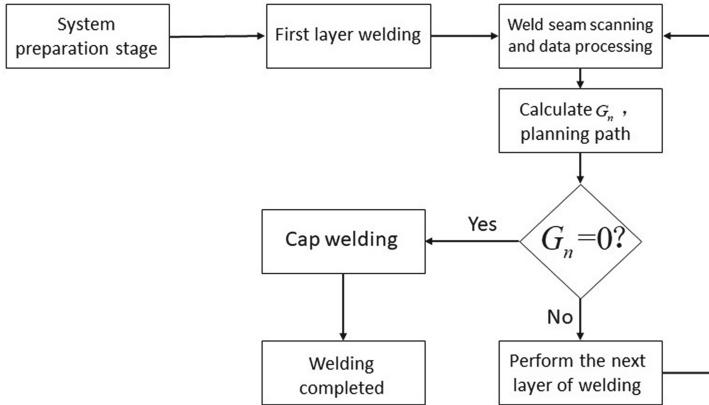
Step 4: Based on real-time contour data, the robot calculates the total number  $G_n$  of remaining welding layers. At the same time, the number  $K_n$  of welding required for the next layer and all the path starting points and postures required for the next layer of welding are calculated.

Step 5: Determine whether  $G_n$  is 0; if  $G_n \neq 0$ , skip to step 6; if  $G_n = 0$ , skip to step 7.

Step 6: The robot performs the welding of the next layer according to the path starting point and posture independently planned in real time, and jumps to step 3 after the welding is completed.

Step 7: The robot calls the pre-programmed top welding program to perform cap welding on the current weld.

Step 8: The current welding seam has been welded, the robot moves to a safe position, the equipment is adjusted to the standby state, and the welding is completed (Fig. 4).



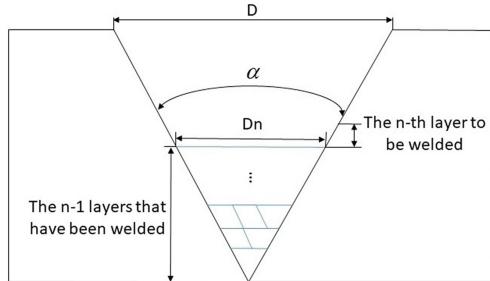
**Fig. 4.** Workflow of the workstation

### 3.2 Adaptive Planning Method

The general idea adopts the strategy of layer-by-layer planning, that is, after welding each layer, the robot returns to the fixed scanning position P [1] above the weld and performs a real-time scan of the current weld cross section. After processing by laser vision system and data processing system, the required weld contour data is obtained. Further, the number of welding paths of the next layer, the starting and ending positions of each path, the angle of the welding gun attitude of each path, the swing welding amplitude of each path, and the number of remaining welding layers can be planned in real time. Based on these data, the robot welding system can autonomously perform multi-layer multi-path welding throughout the entire welding process until the welding is completed.

The contour data of the welding seam, which is processed by the laser vision system and the data processing system, include the coordinates of the start and end points of the bottom of the weld  $(x_{11}, y_{11}, z_{11}), (a_{11}, b_{11}, c_{11})$ , maximum width of weld  $D$ , weld angle  $\alpha$ , two intersection points between the nth layer plane and the weld cross section on both sides of the weld  $(A_n, B_n, C_n), (H_n, E_n, F_n)$ , the width of the nth layer  $D_n$ .

Under the condition that the welding speed and welding current are fixed, the width of the molten pool of each welding path without swinging welding can be deduced by the user through the welding parameter table, which is set to  $H$  here (Fig. 5).



**Fig. 5.** Information map of the weld contour

After welding the  $(n - 1)$ th layer (where  $n \geq 2$ ), the robot returns to the fixed scanning position P [1] to obtain the width of the nth layer  $D_n$  and two intersection points of the nth layer plane and the weld cross section on both sides of the weld  $(A_n, B_n, C_n), (H_n, E_n, F_n)$ .

When welding to the nth layer, the number of layers needed to fill the weld is set as  $G_n$ . The calculation formula is:

$$G_n = \frac{D - D_n}{(D_n - D_2)/(n - 2)} \quad (1)$$

According to the width of the nth layer  $D_n$  and the width of each molten pool without swing welding  $H$ , the number of welding paths of the nth layer  $K_n$  can be derived.

$$K_n = \left[ \frac{D_n}{H} \right] \quad (2)$$

Where, the symbol  $\lceil \rceil$  represents the rounding operation.

Further, the amplitude value of the swing welding required for each welding of the nth layer  $S_n$  can be derived.

$$S_n = \frac{\{ \frac{D_n}{H} \} \times H}{2 \times K_n} \quad (3)$$

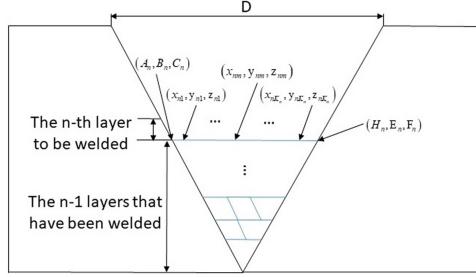
Where, the symbol  $\{ \}$  represents the remainder operation.

After the number of welding paths of the nth layer  $K_n$  is obtained, the coordinates of the starting point and the end point of each path of the nth layer can be planned. The coordinate of the starting point of the mth path of the nth layer is set as  $(x_{nm}, y_{nm}, z_{nm})$ . The calculation formula is:

$$\begin{cases} x_{nm} = A_n - \frac{(m-0.5) \times (A_n - H_n)}{K_n} \\ y_{nm} = B_n - \frac{(m-0.5) \times (B_n - E_n)}{K_n} \\ z_{nm} = C_n - \frac{(m-0.5) \times (C_n - F_n)}{K_n} \end{cases} \quad (4)$$

Further, the coordinate of the end point of the mth path of the nth layer ( $a_{nm}, b_{nm}, c_{nm}$ ) can be derived (Fig. 6). The calculation formula is:

$$\begin{cases} a_{nm} = x_{nm} - x_{11} + a_{11} \\ b_{nm} = y_{nm} - y_{11} + b_{11} \\ c_{nm} = z_{nm} - z_{11} + c_{11} \end{cases} \quad (5)$$



**Fig. 6.** Adaptive path planning

Assume that the welding attitude angle of the first path of each layer determined according to the welding process needs is  $\eta$ , the welding attitude angle of the last path of each layer is  $\lambda$ . Then the welding attitude angle of the mth path of the nth layer  $\theta_{nm}$  can be derived.

$$\theta_{nm} = \eta + \frac{(m-1) \times (\lambda - \eta)}{K_n - 1} \quad (6)$$

## 4 Test Verification

In order to verify the actual effect of this method, according to the overall design scheme proposed in this paper, a robot welding workstation based on adaptive planning was designed and implemented (Fig. 7).

In the robot controller, based on an adaptive planning method proposed in this paper, the robot control program is written, and the welding process parameters are set in advance in the program.

Several Q235 steel plate structural parts with different specifications and thickness of more than 60 mm are selected as the workpieces to be welded, and the weld length of the workpieces used for testing is 280 mm. The whole system starts to run, the workstation works automatically throughout the process, and automatically stops and returns to a safe position when welding is complete (Fig. 8).

It can be seen from the experiment that this workstation runs automatically throughout the process and the welding efficiency is very high. For thick plate structural parts of different specifications, this workstation is compared with manual welding and semi-automatic robot welding workstations. The results show that for the same workpiece, the welding time required by this workstation is



(a) Laser vision system



(b) Workstation panorama

**Fig. 7.** Workstation sketch map

(a) V-type weld



(b) K-type weld

**Fig. 8.** Welding result graph

significantly lower than manual welding or semi-automatic robot welding workstation. The efficiency can reach about 5 times of manual welding.

Observation of the welded thick plate structural parts find that the appearance of the welded joint is smooth and even, with a good molding effect. The welding quality of the welded workpieces is inspected using special testing instruments. The results show that all quality indicators meet the requirements (Tables 1 and 2).

**Table 1.** Comparison table of required welding time

Weld specifications	Manual welding	Semi-automatic workstation	This workstation
K-type 50 mm-60°	2 h	0.8 h	0.4 h
V-type 80 mm-60°	5.7 h	2.3 h	1.1 h
V-type 80 mm-90°	3.5 h	1.3 h	0.6 h

**Table 2.** Welding quality inspection results

Penetration performance	Cladding status	Interlayer porosity	Slag inclusion	Qualified rate of ultrasonic testing	Qualified rate of ray testing
High	Good	Rarely	Rarely	96%	94%

## 5 Conclusion

In view of the current problems of low efficiency and low automation in the application of robot welding in multi-layer multi-path welding of thick plate structural parts, this paper proposes a design method of robot welding workstation based on adaptive planning. We integrate the laser vision system, data processing and monitoring system, and robot welding system into a whole robot welding workstation. At the same time, the control program of the welding robot was written based on the method of adaptive planning.

Based on this workstation, an automatic welding test is carried out on thick plate structural parts. The results show that the system can realize fully automated welding, with high welding efficiency and high self-adaptability, and it has good welding effect for thick plate structural parts of different specifications. The experimental results verify the feasibility and practicability of this method.

**Acknowledgements.** This research has been supported by the Department of Science and Technology of Shandong Province (grant number 2017CXGC0913).

## References

- Li, P., Zhang, J., Liu, C.: Design of robot arc welding workstation with K-shaped positioner. *Metal Process. (Hot Process.)* **11**, 16–18 (2019). (in Chinese)
- Xu, H., Huang, D., Huang, Z.: Development and application of copper plate tuyere robot welding workstation. *Mech. Des. Manuf.* **03**, 121–123 (2016). (in Chinese)
- Bian, X.: Design and application research of robot automatic welding workstation. *Sci. Technol. Innov.* **24**, 160–161 (2019). (in Chinese)
- Zhang, C.: Scheme design of gantry welding station based on robot application. *Spec. Purp. Veh.* **11**, 91–93 (2019). (in Chinese)
- Yang, L., Liu, Y., Peng, J., Liang, Z.: A novel system for off-line 3D seam extraction and path planning based on point cloud segmentation for arc welding robot. *Robot. Comput.-Integr. Manuf.* **64** (2020)
- Li, S., Zhang, J., Liu, C., Pei, J.: Feasibility study on the development of automatic welding workstation for power station boiler main steam pipeline robot, p. 3. Physical and Chemical Inspection Branch of Chinese Mechanical Engineering Society, Failure Analysis Branch of Chinese Mechanical Engineering Society (2013). (in Chinese)
- Chen, P.: Application of robot welding workstation in automobile manufacturing. *Integr. Circ. Appl.* **37**(04), 70–71 (2020). (in Chinese)
- Zhou, J., Huang, J., Jia, X., Li, S., Wei, X.: Application of arc welding robot workstation in heavy industry. *Metal Process. (Hot Process.)* (11), 10–11+15 (2019). (in Chinese)
- Guo, J., Zhu, Z., Sun, B., Zhang, T.: A novel field box girder welding robot and realization of all-position welding process based on visual servoing. *J. Manuf. Processes* (2020)



# Road Intersection Path Planning Based on Q-learning for Unmanned Ground Vehicle

Lingxue Zhao, Chaofang Hu<sup>(✉)</sup>, Yao Guo, and Patrick Tjan

School of Electrical and Information Engineering, Tianjin University, Tianjin, China  
cfhu@tju.edu.cn

**Abstract.** This paper concentrates on the road intersection path planning problem of unmanned ground vehicle (UGV). First, the interaction between UGV and environment is established as a Markov decision process (MDP) model. Considering the feasibility of path, the kinematic model is also utilized to update the states of UGV, such as position, velocity and attitude. Then, the optimal driving strategy and path are generated by Q-learning algorithm. Reward function is designed to reflect the gain and loss of the chosen action. Finally, simulations demonstrate the feasibility of Q-learning in path planning of UGV.

**Keywords:** Unmanned ground vehicle · Path planning · Q-learning · Reinforcement learning

## 1 Introduction

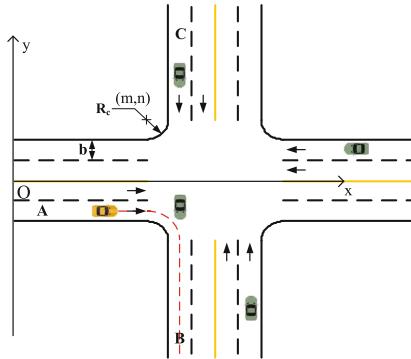
Road intersection is one of the most dangerous road conditions, which accounts for a large amount of traffic accidents [1,2]. Due to the advantages of safe driving and quick response, intelligent driving has become an effective way to avoid traffic accidents [3,4]. Unmanned ground vehicle (UGV) has become an important part of intelligent transportation system [5,6], and is attracting more and more attention from researchers. As a basic problem of UGV research, path planning is vital for safe driving. Various methods have been applied to path planning, such as Rapid-exploring Random Tree [7,8], A\* [9] and Artificial Potential Field [10]. However, all of the methods aforementioned require certain models, and ignore the kinematic characteristics of UGV. These limitations may lead to infeasible or unsafe path [11].

Reinforcement learning provides a new way to satisfy the kinematic characteristics and path planning requirements [12,13]. Q-learning is the most commonly used reinforcement learning method, which does not need any human knowledge. UGV can find the feasible path by continuously interacting with the environment, which is similar with human intelligence [14]. Gao et al. [15] address the decision making of car-following using a reinforcement Q-learning method. Guenca et al. [16] use Q-learning to train the UGV to navigate roundabouts appropriately.

In this paper, Q-learning algorithm is utilized for path planning of UGV. The kinematic model and Markov decision process (MDP) model are established, and the immediate reward functions are designed according to the requirements of path planning at road intersection. The article is organized as follows: Sect. 2 introduces the demonstration of road intersection and kinematic model of UGV. Sect. 3 presents MDP model and Q-learning algorithm. The simulations are presented at Sect. 4. The conclusions are drawn finally.

## 2 Problem Statement

### 2.1 Description of Road Intersection

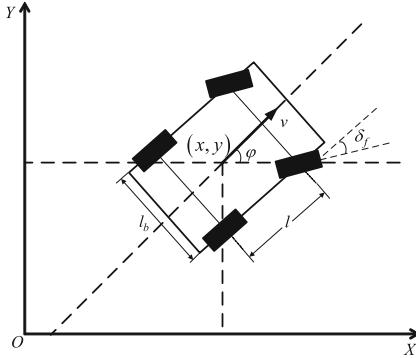


**Fig. 1.** Demonstration of road intersection

Figure 1 shows the demonstration of road intersection. At the road intersection, the corners are quarter arcs with radii of  $R_c$ . UGV conducts a right-turn maneuver, while obstacle vehicles pass the road intersection along a straight line. All of the straight roads have double lanes, and the width of the lane is  $b$ . UGV is traveling at lane A, and the target lane is lane B.

### 2.2 Vehicle Kinematic Model

In order to describe the motion of UGV, the vehicle kinematic model is established, shown in Fig. 2.  $(x, y)$  denotes the coordinates of the center of gravity (CG) of UGV in the inertial coordinate system  $XOY$ ;  $\varphi$  represents the yaw angle;  $\delta_f$  is the front tire steering angle;  $v$  is the velocity of UGV at the CG, and  $acc$  is the acceleration;  $l$  is the wheelbase;  $l_b$  is the width of UGV. The vehicle is a front tire steering vehicle. The position, velocity and attitude of UGV can be calculated by following equation:



**Fig. 2.** Vehicle kinematic model

$$\begin{aligned}\dot{x} &= v \cos \varphi \\ \dot{y} &= v \sin \varphi \\ \dot{v} &= acc \\ \dot{\varphi} &= \frac{v \tan \delta_f}{l}\end{aligned}\tag{1}$$

The state variables are  $\xi = [x, y, v, \varphi]^T$  and the control inputs are  $u = [acc, \delta_f]^T$ .

### 3 Q-learning Algorithm

In this paper, path planning is considered as an optimization problem under the trade-off between gain and loss. Q-learning is a model-free reinforcement learning method, which provides UGV with the capability of learning to find the optimal actions by experience.

#### 3.1 MDP Modeling

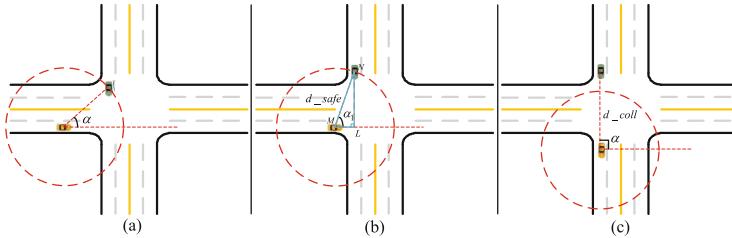
Q-learning is a learning method for discrete state space, so the path planning problem is formulated as a MDP problem. MDP can be described as five variables: finite set of possible states  $S$ ; finite set of actions  $A$ ; state transition probability matrix  $P$ ; discount factor  $\gamma$  ( $\gamma \in [0, 1]$ ) and reward function  $R$ .

#### 3.2 Q-learning for Path Planning

**State Set and Action Set.** In Table 1, the continuous states and actions are discretized with equal intervals. In state set,  $r$  denotes the position of UGV. When UGV is approaching the road intersection,  $r = 1$ . When UGV is crossing the road intersection, the value of  $r$  is 2.  $r = 3$  means UGV is traveling at the straight road after crossing the intersection.  $d$  represents the lateral displacement

**Table 1.** State set and action set of MDP

State set	Action set
$r \in \{1, 2, 3\}$	$a \in \{-5, -2.5, 0, 2.5, 5\} \text{ m/s}^2$
$d \in \{0, 0.05, 0.1, \dots, 1.75\} \text{ m}$	
$V \in \{-1, 0, 1, \dots, 16\} \text{ m/s}$	$\delta_f \in \{-35^\circ, -30^\circ, -5^\circ, 0^\circ, 5^\circ, 30^\circ, 35^\circ\}$
$d_{coll} \in \{0, 1\}$	
$\alpha \in \{-1, 1\}$	

**Fig. 3.** Typical conditions of obstacle avoidance

from road centerline. Velocity is designed according to the traffic rules. The velocity range is 5–15 m/s at the straight road and 0–15 m/s at road intersection for safety. If  $v$  is negative, we set  $V = -1$ ; if  $v$  is larger than the upper limit, the value of  $V$  is set as 16 m/s.

$d_{coll}$  and  $\alpha$  are related to the obstacle avoidance. In Fig. 3, the red circle with a radius of safety distance  $d_{safe}$  is the safety area of UGV. The distance between UGV and obstacle vehicle is  $d_{coll}$ , which is simplified by 0 and 1 in state set:  $d_{coll} = 0$  means that the distance is larger than safety distance; otherwise, the value of  $d_{coll}$  is 1. Three typical conditions are presented in Fig. 3. Line  $MN$  is the connection line of UGV and obstacle vehicle.  $\alpha$  is the angle between line  $MN$  and the horizontal direction. It can be seen that as UGV approaches road intersection,  $\alpha$  becomes larger and larger. In Fig. 3 (b), UGV is located at the end of the straight road.  $NL$  is the heading direction of obstacle vehicle, which is perpendicular to  $ML$ . Hence,  $\alpha_1$  can be calculated as:

$$\alpha_1 = \arccos \left( \frac{|ML|}{|MN|} \right) = \arccos \left( \frac{R_c + 2/b}{d_{safe}} \right) \quad (2)$$

If  $\alpha < \alpha_1$ , UGV is far from the road intersection while obstacle vehicle is close to road intersection, shown as Fig. 3 (a). In this case, we set  $\alpha = -1$ , and UGV should decelerate to avoid collision. On the contrary, if  $\alpha \geq \alpha_1$ , shown as Fig. 3 (b) and (c), UGV is required to accelerate to avoid collision, and  $\alpha = 1$ .

**Reward Function.** The reward contains two parts: position reward and velocity reward. Since  $d$  represents the lateral displacement from the road centerline,

we hope  $d$  could be as small as possible. Besides, if UGV collides with road boundary, it will receive a punishment. However, the requirements of velocity are different according to the position of UGV. According to the actual driving condition, the total reward is discussed in three situations.

### (1) Approaching road intersection ( $r = 1$ )

In this situation, UGV should decelerate to prepare for crossing the road intersection. If UGV is traveling at the centerline of the road with allowable velocity, it will receive a positive reward when decelerating. Considering the comfortability, lower deceleration will lead to larger reward. When the velocity is slow enough to enter the road intersection, it will get the largest reward of 50. But if UGV is not at the centerline, or the velocity is higher than the upper limit, it will receive a punishment of  $-20$ . Besides, if the velocity is lower than the limit and UGV continues to decelerate, it will also receive a large punishment; but if UGV accelerates to try to reach the velocity limit, it will obtain a small punishment. In summary, when the distance between UGV and obstacle vehicle is larger than safety distance, i.e.  $d\_coll = 0$ , the reward is designed as follows:

$$R_1 = \begin{cases} 50 & d = 0 \text{ and } V = 5 \\ 30 & d = 0 \text{ and } 5 < V \leq 15, \text{ and } -1 \leq V - V_0 < 0 \\ 2 & d = 0 \text{ and } 5 < V \leq 15, \text{ and } V - V_0 < -1 \\ -1 & d = 0 \text{ and } 5 < V \leq 15, \text{ and } V - V_0 \geq 0 \\ -1 & d = 0 \text{ and } V < 5, \text{ and } V - V_0 > 1 \\ -20 & d = 0 \text{ and } V < 5, \text{ and } V - V_0 \leq 1 \\ -20 & d \neq 0 \text{ or } V > 15 \end{cases} \quad (3)$$

where  $V_0$  is the velocity at the last instant. When UGV encounters the obstacle vehicle, i.e.  $d\_coll = 1$  and  $\alpha = -1$ , the most important task is to avoid collision. Hence, UGV should decelerate, and the quicker it decelerates, the larger reward it will receive. So the reward is:

$$R_{coll1} = \begin{cases} 20|V - V_0| & d = 0, \text{ and } V - V_0 \leq 0, \text{ and } V \geq 0 \\ -1 & d \neq 0, \text{ and } V - V_0 \leq 0, \text{ and } V \geq 0 \\ -20 & d \neq 0, \text{ and } V - V_0 > 0, \text{ and } V \geq 0 \\ -20 & V < 0 \end{cases} \quad (4)$$

### (2) At road intersection ( $r = 2$ )

UGV is required to travel at 5 m/s at road intersection for safety, so the reward for  $V = 5$  m/s is the highest. Considering obstacle avoidance, the velocity can be less or larger than 5 m/s, but the reward is negative. If the velocity is beyond the limit, UGV will receive the largest punishment of  $-20$ . For the position reward, the closer UGV is to the centerline, the greater reward it will get. If UGV collides with the road boundary, it will receive a punishment of  $-20$ . The reward is designed as Eq. (5):

$$R_2 = \begin{cases} 100 - \frac{50d}{0.85} & d \leq 0.85 \text{ and } V = 5 \\ -1 & d \leq 0.85 \text{ and } 5 < V \leq 15 \\ -1 & d \leq 0.85 \text{ and } V < 5, \text{ and } 0 < V - V_0 \leq 1 \\ -1 & d \leq 0.85 \text{ and } V > 15, \text{ and } -1 \leq V - V_0 < 0 \\ -20 & d \leq 0.85 \text{ and } V < 5, \text{ and } V - V_0 \leq 0 \cup V - V_0 > 1 \\ -20 & d \leq 0.85 \text{ and } V > 15, \text{ and } V - V_0 < -1 \cup V - V_0 \geq 0 \\ -20 & d > 0.85 \end{cases} \quad (5)$$

When UGV encounters obstacle vehicle at road intersection, the value of  $\alpha$  is 1. So UGV should accelerate to avoid collision, and the higher velocity is, the greater reward UGV will obtain. The reward function is:

$$R_{coll2} = \begin{cases} 20|V - V_0| & d \leq 0.85, \text{ and } V - V_0 \geq 0, \text{ and } V \geq 0 \\ -1 & d > 0.85, \text{ and } V - V_0 \geq 0, \text{ and } V \geq 0 \\ -20 & d > 0.85, \text{ and } V - V_0 < 0, \text{ and } V \geq 0 \\ -20 & V < 0 \end{cases} \quad (6)$$

### (3) Leaving road intersection ( $r = 3$ )

After crossing the road intersection, UGV should accelerate to the highest velocity. So the acceleration should be positive, and the selection of deceleration action should be punished. But considering the comfortability, smaller acceleration is better. The rewards are shown as Eq. (7).

$$R_3 = \begin{cases} 50 - \frac{25d}{0.85} & d \leq 0.85 \text{ and } V = 15 \\ 30 - \frac{15d}{0.85} & d \leq 0.85 \text{ and } 5 \leq V < 15, \text{ and } 0 < V - V_0 \leq 1 \\ 2 - \frac{d}{0.85} & d \leq 0.85 \text{ and } 5 \leq V < 15, \text{ and } V - V_0 > 1 \\ -1 & d \leq 0.85 \text{ and } 5 \leq V < 15, \text{ and } V - V_0 \leq 0 \\ -1 & d \leq 0.85 \text{ and } V > 15, \text{ and } V - V_0 < -1 \\ -20 & d \leq 0.85 \text{ and } V > 15, \text{ and } V - V_0 \geq -1 \\ -20 & d > 0.85 \text{ or } V < 5 \end{cases} \quad (7)$$

When UGV meets obstacle vehicle, the reward is shown as follows:

$$R_{coll3} = \begin{cases} 20|V - V_0| & d \leq 0.85, \text{ and } V - V_0 \geq 0, \text{ and } V \geq 0 \\ -1 & d > 0.85, \text{ and } V - V_0 \geq 0, \text{ and } V \geq 0 \\ -20 & d > 0.85, \text{ and } V - V_0 < 0, \text{ and } V \geq 0 \\ -20 & V < 0 \end{cases} \quad (8)$$

**Reward Value Update.** In each learning iteration, UGV will randomly select a certain action, and obtain the immediate reward  $R(s, a)$ . The state-action value function, also called Q-value, is calculated by the following equation:

$$Q(s, a) = R(s, a) + \gamma \max Q(s', a') \quad (9)$$

where  $Q(s, a)$  is the value of an action  $a$  under the state  $s$ , and  $a'$  is the next action under next state  $s'$ . The Q-value update function is presented as:

$$Q(s, a) \leftarrow Q(s, a) + \beta [R(s, a) + \gamma \max_{a'} Q(s', a') - Q(s', a')] \quad (10)$$

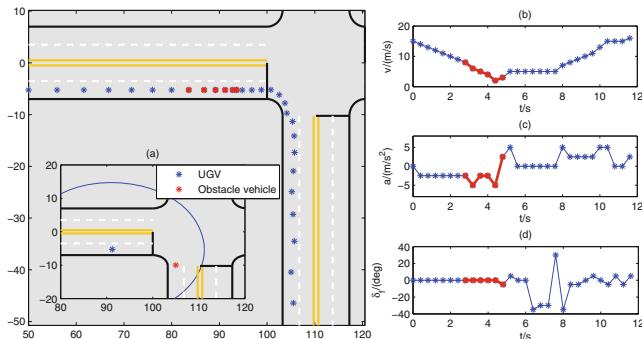
where  $\beta \in [0, 1]$  is the learning factor. At each iteration, UGV chooses an action by  $\varepsilon$ -greedy algorithm.

**Algorithm.** The algorithm for path planning at road intersection is summarized as following steps:

- Step 1. Initialize Q-table,  $\beta$ ,  $\gamma$  and  $\varepsilon$ .
- Step 2. Observe the current states of UGV.
- Step 3. Select an action by  $\varepsilon$ -greedy algorithm.
- Step 4. Perform the action and obtain immediate reward according to (3)–(8).
- Step 5. Update the states of UGV by kinematic model (1). If UGV has not reached the target state, update Q-table according to Eq. (10), and go back to Step 2. Otherwise, quit the algorithm.

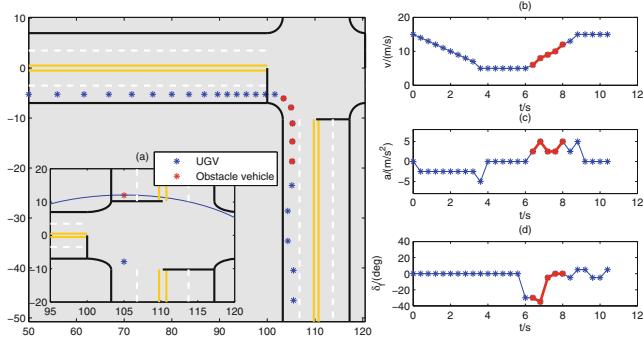
## 4 Simulation

Three typical scenarios are discussed in this section.  $b$  is 3.5 m,  $R_c$  is 5 m,  $l$  is 2.7 m and  $l_b$  is 1.8 m. The initial position, velocity and attitude for UGV are  $\xi = [50, -5.25, 15, 0]^T$ , and the initial state set for Q-learning is  $S = [1, 0, 15, -1, -1]$ .  $d_{safe}$  is 20 m,  $\beta$  is 0.4,  $\gamma$  is 0.9, and  $\varepsilon$  is 0.01. The obstacle vehicle travels from lane  $C$  to lane  $B$ . The target states are  $r = 3$ ,  $d = 0$  and  $V = 15$ .

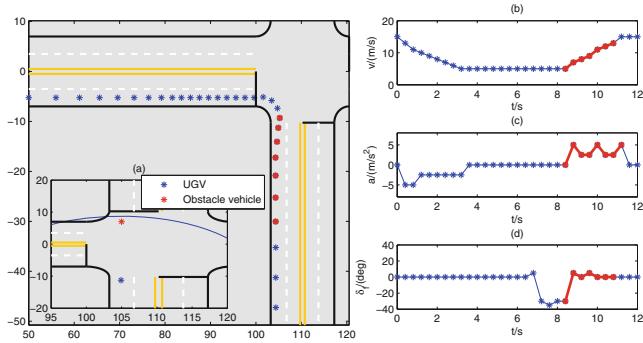


**Fig. 4.** Path planning results in Scenario 1

In Scenario 1, the initial information for obstacle is  $\xi_1 = [105, 30, 10, -\frac{\pi}{2}]^T$ . When UGV is approaching road intersection, it encounters with the obstacle,



**Fig. 5.** Path planning results in Scenario 2



**Fig. 6.** Path planning results in Scenario 3

shown in Fig. 4 (a). Figure 4 (b), (c) and (d) are the velocity, acceleration and steering angle of UGV, respectively. The red points represent that the obstacle vehicle is within the safety area of UGV and UGV is trying to avoid collision.

The initial information of obstacle vehicle is  $\xi_2 = [105, 80, 10, -\frac{\pi}{2}]^T$  in Scenario 2. UGV meets the obstacle vehicle when crossing the road intersection. The results are presented in Fig. 5.

In Scenario 3, after crossing the road intersection, UGV encounters with the obstacle vehicle. The initial position of obstacle vehicle is  $\xi_3 = [105, 95, 10, -\frac{\pi}{2}]^T$ . The results are shown in Fig. 6.

## 5 Conclusion

In this paper, Q-learning method is used for path planning of UGV at road intersection. First, the vehicle kinematic model is introduced for the state update of Q-learning. Then, a MDP model is established. Several states and actions are selected to describe the movement of UGV, and the collision avoidance is also considered. By designing the reward function and update function, Q-learning

algorithm is conducted to find the optimal control policy for path planning. Finally, three scenarios are simulated to illustrate the feasibility of the method. In the future, the reward function can be improved, and the computation burden should be further reduced.

**Acknowledgments.** This work is supported by National Natural Science Foundation of China under Grant (61773279, 61873340), Key Technologies Program of Tianjin (19YFHBQY00040), and Joint Science Foundation of Ministry of Education of China (No. 6141A0202304).

## References

1. Chen, Y., Zha, J., Wang, J.: Autonomous T-intersection driving strategy considering oncoming vehicles based on connected vehicle technology. *IEEE/ASME Trans. Mechatron.* **24**(6), 2779–2790 (2019)
2. Chen, L., Englund, C.: Cooperative intersection management: a survey. *IEEE Trans. Intell. Transp. Syst.* **17**(2), 570–586 (2016)
3. Wu, B., Qian, L., Lu, M., Qiu, D.: Optimal control problem of multi-vehicle cooperative autonomous parking trajectory planning in a connected vehicle environment. *IET Intel. Transp. Syst.* **13**(11), 1677–1685 (2019)
4. Hu, C., Cao, L., Zhao, L., Wang, N.: Model predictive control-based steering control of unmanned ground vehicle with tire blowout. *J. Tianjin Univ. (Sci. Technol.)* **52**(5), 468–474 (2019)
5. Kong, L., Khan, M.K., Wu, F., Chen, G., Zeng, P.: Millimeter-wave wireless communications for IoT-cloud supported autonomous vehicles: Overview, design, and challenges. *IEEE Commun. Mag.* **55**, 62–68 (2017)
6. Katrakazas, C., Quddus, M., Chen, W., Lipika, D.: Real-time motion planning methods for autonomous on-road driving: state-of-the-art and future research directions. *Transp. Res. Part C0 Emerg. Technol.* **60**, 416–442 (2015)
7. Shi, Y., Li, Q., Bu, S., Yang, J., Zhu, L.: Research on intelligent vehicle path planning based on rapidly-exploring random tree. *Math. Prob. Eng.* (2020). Article ID 5910503
8. Rasekhipour, Y., Khajepour, A., Chen, S., Litkouhi, B.: A potential field-based model predictive path-planning controller for autonomous road vehicles. *IEEE Trans. Intell. Transp. Syst.* **18**(5), 1255–1267 (2017)
9. Wang, Y., Liu, Z., Zuo, Z., Li, Z.: Local path planning of autonomous vehicles based on A\* algorithm with equal-step sampling. In: 37th Chinese Control Conference, Wuhan, pp. 7828–7833 (2018)
10. Huang, Y., Ding, H., Zhang, Y., Wang, H., Cao, D., Xu, N., Hu, C.: A motion planning and tracking framework for autonomous vehicles based on artificial potential field elaborated resistance network approach. *IEEE Trans. Industr. Electron.* **67**(2), 1376–1386 (2020)
11. Chen, C., Chen, X., Ma, F., Zeng, X., Wang, J.: A knowledge-free path planning approach for smart ships based on reinforcement learning. *Ocean Eng.* **189**, 106299 (2019)
12. You, C., Lu, J., Filev, D., Tsiotras, P.: Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning. *Robot. Auton. Syst.* **114**, 1–18 (2019)

13. Makantasis, K., Kontorinaki, M., Nikolos, I.: Deep reinforcement-learning-based driving policy for autonomous road vehicles. *IET Intell. Transp. Syst.* **14**(1), 13–24 (2020)
14. Kontoudis, G., Vamvoudakis, K.: Kinodynamic motion planning with continuous-time Q-learning: an online, model-free, and safe navigation framework. *IEEE Trans. Neural Netw. Learn. Syst.* **30**(12), 3803–3817 (2019)
15. Gao, Z., Sun, T., Xiao, H.: Decision-making method for vehicle longitudinal automatic driving based on reinforcement Q-learning. *Int. J. Adv. Rob. Syst.* **16**(3), 1–13 (2019)
16. Cuneca, L., Puertas, E., Andres, J., Aliane, N.: Autonomous driving in roundabout maneuvers using reinforcement learning with Q-Learning. *Electronics* **8**(12), 1536 (2019)



# Univariate ReLU Neural Network and Its Application in Nonlinear System Identification

Xinglong Liang and Jun Xu<sup>(✉)</sup>

Harbin Institute of Technology, Shenzhen, China

**Abstract.** ReLU (rectified linear units) neural network has received significant attention since its emergence. In this paper, a univariate ReLU (UReLU) neural network is proposed to both model the nonlinear dynamic system and reveal the insights about the system. Specifically, the neural network consists of neurons with linear and UReLU activation functions, and the UReLU functions are defined as the ReLU functions respect to each dimension. The UReLU neural network is a single hidden layer neural network, and the structure is relatively simple. The initialization of the neural network employs the decoupling method, which provides a good initialization and some insight into the nonlinear system. Compared with normal ReLU neural network, the number of parameters of UReLU network is less, but it still provide a good approximation of the nonlinear dynamic system. The performance of the UReLU neural network is shown through a Hysteretic benchmark system: the Bouc-Wen system. Simulation results verify the effectiveness of the proposed method.

**Keywords:** Neural network · Identification · Univariate ReLU function · Decoupling method

## 1 Introduction

System identification deals with the problem of building mathematical models of dynamical systems based on observed data from the system [1], which can be applied in industrial processes, economic and financial systems, biology and life sciences, medicine, social systems, and so on [2]. Since the theory of linear time-invariant (LTI) systems has been extensively studied over decades [3], a huge amount of effective system identification methods have been developed for linear systems [2,4]. However most industrial systems are nonlinear systems, for which accurate descriptions can not be built by just using the identification methods developed for linear systems. Hence the nonlinear system identification is currently a field of active research [5–8].

As an effective method, neural networks based method has played important roles in different fields, which is due to that they are conceptually simple and easy

to train and use [2]. In spite of the excellent performance of neural networks, they still receive criticism as they are basically black-box models, making it difficult to draw any insights from the identified model. Moreover, the parameters of the neural network may be too much, and its training requires sophisticated skills and tunings.

The rectified linear units (ReLU)  $\max(0, x)$  as an activation function has been used in neural networks widely. One advantage of ReLU is its non-saturating nonlinearity. In terms of training time with gradient descent, these non-saturating nonlinearity are much faster than the saturating nonlinearity like sigmoid function [9]. Besides, many ReLU variants have been proposed to boost the performance of networks, such as Leaky ReLU (LReLU) [10], the parametric rectified linear unit (PReLU) [11], the exponential linear Unit (EReLU) [12] etc.

In this paper, we propose a univariate ReLU (UReLU) neural network based on the UReLU activation function, which is ReLU function with respect to each dimension. The initialization of the UReLU network can be fulfilled by a decoupling method, which was based on tensor decomposition and proposed by [5] and [13]. After initialization, the parameters of the UReLU neural network can be estimated by the Variable Projection Method [14]. The initialization of the neural network employs the decoupling method, which provides a good initialization and some insight into the nonlinear system.

The rest of the paper is organized as follows. Section 2 gives a detailed description of the UReLU function, and then the structure as well as the training of the UReLU neural network are provided in Sect. 3 and 4, respectively. Section 5 describes the interpretability of the UReLU neural network, which will facilitate the subsequent control or optimization after nonlinear system identification. Simulation studies are shown in Sect. 6. Finally, Sect. 7 ends with conclusions.

## 2 Univariate ReLU Function

In this paper, the univariate ReLU (UReLU) is introduced as the activation function of the neural network, which can be simply expressed as

$$\max\{0, x_i - \beta_{i1}\}, \max\{0, x_i - \beta_{i2}\}, \dots, \max\{0, x_i - \beta_{i,q_i}\}$$

where  $i \in \{1, \dots, n\}$ , and the bias  $\beta_{ij} (j = 1, \dots, q_i)$  is chosen based on the training data and satisfies

$$\beta_{i1} < \beta_{i2} < \dots < \beta_{i,q_i}.$$

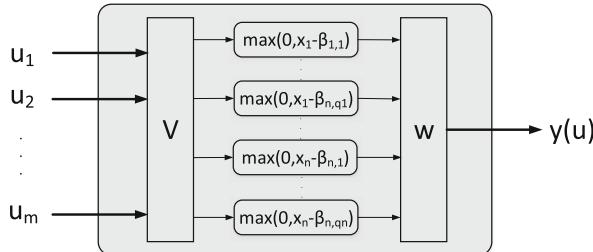
Specifically, assume the training data is

$$(\mathbf{x}(k), y(k))_{k=1}^N,$$

we can choose the bias  $\beta_{i1}, \dots, \beta_{i,q_i}$  according to the distribution of  $x_i(k), i = 1, \dots, n, k = 1, \dots, N$ . For example, for the dimension  $i$ , we can choose  $\beta_{i1}, \dots, \beta_{i,q_i}$  to be the  $q_i$ -quantiles of  $x_i(k), k = 1, \dots, N$ . Compared with the ReLU, the UReLU focuses on the single variable and the bias parameters are determined according to the data distribution and need not to be trained during the network training.

### 3 Structure of the UReLU Neural Network

The structure of the UReLU neural network is shown in Fig. 1, in which the input vector  $\mathbf{u} = [u_1, \dots, u_m] \in \mathbb{R}^m$  and the output is  $y(\mathbf{u}) \in \mathbb{R}$ . A linear transformation  $\mathbf{V}$  is introduced to transform the input vector  $\mathbf{u}$  to the intermediate vector  $\mathbf{x} \in \mathbb{R}^n$ , which is then sent to the UReLU neurons. Finally, the output is derived as the weighted sum of the UReLU neurons. Generally speaking, the dimension of the intermediate vector  $\mathbf{x}$  is lower than that of the input vector  $\mathbf{u}$ , i.e.,  $n < m$ .



**Fig. 1.** Structure of the UReLU neural network

The following describes the UReLU neural network with respect to 3 parts, the connection between the input and the intermediate vector, the generation of the UReLU nonlinearities, and the connection to the output.

#### 3.1 Connection Between the Input and the Intermediate Vector

The intermediate vector  $\mathbf{x} \in \mathbb{R}^n$  is obtained through a linear transformation of the input vector  $\mathbf{u} \in \mathbb{R}^m$ , which is fulfilled through a linear transformation matrix  $\mathbf{V}$ . The linear transformation matrix  $\mathbf{V}$  can be expressed as

$$\mathbf{V} = \begin{bmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{m1} & \cdots & v_{mn} \end{bmatrix}, \quad (1)$$

then we have

$$\mathbf{x} = \mathbf{V}^T \cdot \mathbf{u}.$$

After the linear transformation, the dimension of the intermediate vector is less than that of the input vector, i.e.,  $n < m$ .

For  $N$  samples, define the input data matrix and the intermediate data matrix as

$$\mathbf{U} = \begin{bmatrix} u_1(1) & \cdots & u_m(1) \\ \vdots & \ddots & \vdots \\ u_1(N) & \cdots & u_m(N) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_1(1) & \cdots & x_n(1) \\ \vdots & \ddots & \vdots \\ x_1(N) & \cdots & x_n(N) \end{bmatrix},$$

then we have

$$\mathbf{X} = \mathbf{U}\mathbf{V} \quad (2)$$

### 3.2 Generation of the UReLU Nonlinearities

Based on the intermediate data matrix  $\mathbf{X}$ , the UReLU nonlinearities can be represented as follows,

$$\begin{aligned} \mathbf{B} = & [\max\{0, \mathbf{x}_1 - \beta_{1,1}\}, \dots, \max\{0, \mathbf{x}_1 - \beta_{1,q_1}\} \\ & \dots, \max\{0, \mathbf{x}_n - \beta_{n,1}\}, \dots, \max\{0, \mathbf{x}_n - \beta_{n,q_n}\}], \end{aligned} \quad (3)$$

in which  $\mathbf{x}_i = [x_i(1), \dots, x_i(N)]^T \in \mathbb{R}^N$ . According to (2), we have  $\mathbf{x}_i = \mathbf{U} \cdot [v_{1i}, \dots, v_{mi}]^T$ . As is mentioned before, the bias  $\beta_{ij}, i = 1, \dots, n, j = 1, \dots, q_i$  can be determined according to the sampled data distribution. For balanced data, we set  $q_i = q, \forall i \in \{1, \dots, n\}$  and employ the following simple method to obtain the bias,

$$\beta_{ij} = [\max(\mathbf{x}_i) - \min(\mathbf{x}_i)] \cdot s_j + \min(\mathbf{x}_i), \quad (4)$$

in which  $\mathbf{s} = [0, 1/q, \dots, (q-1)/q]$ . In the training of the UReLU neural network,  $q$  is preset and not changed, while  $\beta_{ij}$  fluctuates as  $\mathbf{V}$  changes. Hence the data matrix  $\mathbf{B}$  is basically dependent on  $\mathbf{V}$ , i.e.,  $\mathbf{B}(\mathbf{V})$ .

### 3.3 Connection to the Output

The output of the UReLU neural network can be written as:

$$\hat{\mathbf{y}} = [\mathbf{1}, \mathbf{B}(\mathbf{V})] \cdot \mathbf{w} \quad (5)$$

where  $\mathbf{1} \in \mathbb{R}^N$  is a vector with all entries being 1,  $\mathbf{w} = [w_0, w_1, \dots, w_M]^T$  is the weight vector,  $w_0$  is for the constant neuron,  $M = nq$  denotes the number of UReLU neurons.

## 4 Training of the UReLU Neural Network

As is mentioned before, the predicted output is a function of the linear transformation matrix  $\mathbf{V}$ , the weights vector  $\mathbf{w}$  and the input  $\mathbf{u}$ . Hence the training of the UReLU neural network deals with the problem of finding the optimal  $\mathbf{V}$  and  $\mathbf{w}$ . Here the gradient-based optimization method is employed, which includes the procedures of parameter initialization and optimization.

### 4.1 Parameter Initialization

A good initialization parameter is vital for the gradient based method, which will make the convergence faster and the result more stable. The initialization of the parameters  $\mathbf{V}$  and  $\mathbf{w}$  is fulfilled through a tensor decomposition method proposed in [13].

Specifically, the initialization of the linear transformation matrix  $\mathbf{V}$  consists of the following 3 steps:

- 1) An NARX polynomial model is established based on the input and output of nonlinear system.
- 2) Obtain the Hessian of NARX polynomial model. And then Hessian is evaluated at  $N$  different operating points, the results can be stacked into a 3 dimensional tensor  $\mathcal{H}$ .
- 3) The 3 dimensional tensor  $\mathcal{H}$  can be written as

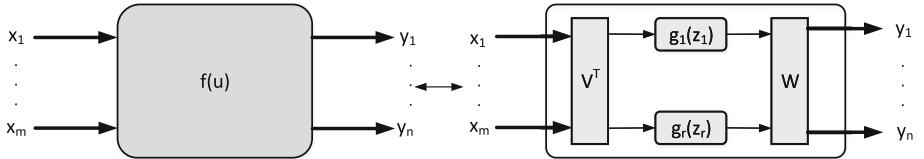
$$\mathcal{H} = [\mathbf{V}, \mathbf{V}, \mathbf{W}] \quad (6)$$

This is the Canonical Polyadic Decomposition (CPD) [16] of the tensor  $\mathcal{H}$ . The CPD can be implemented by the tensorlab toolbox [17]. we only use the obtained  $\mathbf{V}$  to initialize the linear transformation matrix. Actually by using this initialization, the Hessian information can provides some insight into the nonlinear system, thus facilitate the interpretability [13]. It is noted that we can use the forward regression with orthogonal least squares (FROLS) developed by [15] to reduce the number of polynomial terms.

The reason of using this kind of parameter initialization strategy is that the UReLU structure is similar to the decoupling structure proposed in [13], which is shown in Fig. 2, in which

$$\mathbf{f}(\mathbf{x}) = \mathbf{W}\mathbf{g}(\mathbf{V}^T\mathbf{x}) \quad (7)$$

where  $\mathbf{W}$  and  $\mathbf{V}$  are transformation matrices, the vector function  $\mathbf{g}(z)$  is composed of univariate functions  $g_i(z_i)$  in its  $r$  components.



**Fig. 2.** A typical decoupling structure.

## 4.2 Optimization

Given the inputs and outputs of a nonlinear system at  $N$  sampling points, our objective is to minimize the following criterion

$$\|r(\mathbf{V}, \mathbf{w})\|_2^2 = \|\mathbf{y} - \hat{\mathbf{y}}(\mathbf{V}, \mathbf{w})\|_2^2 \quad (8)$$

with respect to  $\mathbf{V}$  and  $\mathbf{w}$ .

It should be noted that  $\hat{\mathbf{y}}(\mathbf{V}, \mathbf{w}) = \mathbf{B}(\mathbf{V})\mathbf{w}$  is a nonlinear function of  $\mathbf{V}$  and a linear function of  $\mathbf{w}$ . If the matrix  $\mathbf{V}$  is fixed,  $\mathbf{w}$  can be easily obtained by solving the linear least squares problem:

$$\mathbf{w} = \mathbf{B}(\mathbf{V})^\dagger \mathbf{y} \quad (9)$$

where  $\mathbf{B}(\mathbf{V})^\dagger$  is the Moore-Penrose generalized inverse of  $\mathbf{B}(\mathbf{V})$ . Substituting (9) into (8) and we obtain:

$$\min_{\mathbf{V}} \frac{1}{2} \|r(\mathbf{V}, \mathbf{w})\|_2^2 = \min_{\mathbf{V}} \frac{1}{2} \|\mathbf{y} - \mathbf{B}(\mathbf{V})\mathbf{B}(\mathbf{V})^\dagger \mathbf{y}\|_2^2 \quad (10)$$

This is the Variable Projection functional [14]. In this case, the linear parameter  $\mathbf{w}$  depends on the nonlinear parameters  $\mathbf{V}$ , hence we only need to focus on optimizing  $\mathbf{V}$ .

To solve the problem (10), the most reliable nonlinear least-squares algorithms require the Jacobian matrix  $\frac{\partial r}{\partial \mathbf{V}}$ . In 1973, Golub and Pereyra showed how the Jacobian matrix  $\frac{\partial r}{\partial \mathbf{V}}$  could be computed exactly from the derivative  $\frac{\partial \mathbf{B}(\mathbf{V})}{\partial \mathbf{V}}$ . This was an important step for efficiency and reliability of the method. Thanks to the structure of the UReLU neural network, the derivative  $\frac{\partial \mathbf{B}(\mathbf{V})}{\partial \mathbf{V}}$  is easy to be obtained. For the element  $\max\{x_i(k) - \beta_{ij}\}$  in  $\mathbf{B}$ , the derivative can be expressed as

$$\frac{\partial \beta_{ij}}{\partial v_{st}} = \begin{cases} s_j [u_s(k_{\max,i}) - u_s(k_{\min,i})] + u_s(k_{\min,i}) & t = i \\ 0 & t \neq i. \end{cases} \quad (11)$$

where

$$x_i(k_{\min,i}) = \min(\mathbf{x}_i), x_i(k_{\max,i}) = \max(\mathbf{x}_i),$$

Finally, the problem can be solved by quasi-Newton method or Levenberg-Marquardt method, in which the number of iterations can be tuned to prevent overfitting.

## 5 Interpretability of the UReLU Neural Network

As is mentioned before, although most nonlinear models can capture the nonlinear phenomenon very well, the number of parameters used is large and at the same time, the model interpretability is lost. In this section, the model interpretability is illustrated through 2 aspects: dimensionality reduction and the easy-get piecewise linear (PWL) relationship.

### 5.1 Dimensionality Reduction

The transformation  $\mathbf{X} = \mathbf{U}\mathbf{V}$  maps a data vector  $\mathbf{u}_i$  from an original space of  $m$  variables to a new space of  $n$  variables. As mentioned before, we require that  $n < m$ . After the optimization process in the training of the UReLU neural network, the new generated data matrix  $\mathbf{X}$  is always uncorrelated, which will be shown clearly in the simulation study. Hence, through this process, the dimensionality of the problem can be greatly reduced, which makes the subsequent training or identification problem easier.

This dimensionality reduction procedure is similar to the principle analysis (PCA), in which as much of the variance in the dataset as possible is retained.

The major difference between our linear transformation and the PCA transformation is that what we focus on is the training efficiency, i.e., the variable selection procedure in this paper is a supervised process. In special, the initial linear transformation is obtained through the stacked Hessian of the nonlinear dynamic system, which reflects the insights about the system to some extent.

## 5.2 Linear Relationships on Subregions

The system input  $\mathbf{U}$  passes through the dimensionality reduction module and is sent to the UReLU module. And from the expression of the UReLU functions, we can know the domain partition of the input  $\mathbf{U}$ , and in each subregion, the predicted output is affine. For the UReLU neural network, the subregions and linear relationships defined on the subregions are clear. In special, when  $\beta_{ij}$  are chosen according to Sect. 3.2, there are totally  $(q - 1)^n$  subregions, which is the Cartesian product of the sets

$$\begin{aligned} & \{\beta_{11} \leq x_1 \leq \beta_{12}, \dots, \beta_{1,q-1} \leq x_1 \leq \beta_{1q}\} \\ & \vdots \\ & \{\beta_{n1} \leq x_n \leq \beta_{n2}, \dots, \beta_{n,q-1} \leq x_n \leq \beta_{nq}\}, \end{aligned}$$

i.e., we have the subregions  $\Gamma_{k_1 \dots k_n}$  in the  $\mathbf{x}$  space with  $k_i \in \{1, \dots, q - 1\}$  and

$$\Gamma_{k_1 \dots k_n} = \{x \in [\beta_{1,k_1}, \beta_{1,k_1+1}], \dots, x_n \in [\beta_{n,k_n}, \beta_{n,k_n+1}]\}.$$

For any  $\mathbf{x} \in \Gamma_{k_1 \dots k_n}$ , we have

$$\hat{y} = w_0 + \sum_{i=1}^{k_1} w_i(x_1 - \beta_{1i}) + \dots + \sum_{i=1}^{k_n} w_{c_{n-1}+i}(x_n - \beta_{ni}), \quad (12)$$

in which  $c_s = sq, s = 1, \dots, n$  and  $c_n = M$ .

The linear function (12) can be expressed with respect to  $\mathbf{u}$  according to the linear transformation matrix  $\mathbf{V}$  which will facilitate greatly the control and optimization of the nonlinear system after system identification.

## 6 Benchmark Result

The experiment approximates a Bouc-Wen Benchmark Model, which described as [18]. There are totally 40,960 training data points and two validation sets, respectively 8192 multi-sine input and 153000 swept sine input. According to [18], we use the simulated RMSE to evaluate the performance of system, and report it as the dB form, i.e.,  $20\lg(\text{RMSE})$ . The same as [5], we choose 30 regressors and all of the 40960 samples are used for training. Then the initial value of the linear transformation matrix is identified according to Sect. 4.1. The Variable Projection Method was used to estimate all the parameters.

It is noted that the condition numbers of  $\mathbf{U}$  and  $\mathbf{X}$  are  $1.74 \times 10^7$  and 28.26, respectively. After the linear transformation, uncorrelated variables are

formed, making the following identification easier, confirming the effectiveness of the dimensionality reduction. Table 1 lists the simulation error when using the UReLU neural network, with both the multi-sine and swept sine inputs, denoted by RMSE(mul) and RMSE(swe), respectively. The results are compared with other state-of-the-art results as well as those obtained with a single layer ReLU neural network which contains 50 neurons. The number of parameters used for each method are also listed in Table 1, in which the notation “–” indicates that the corresponding value is not available.

**Table 1.** Comparison of the test performance of several approaches on Bouc-Wen system.

Method	RMSE (mul)	RMSE (swe)	# Parameter
NARX [19]	–75.73	–77.20	–
EHH [20]	–83.00	–88.78	3530
D-NARX [5]	–85.42	–95.55	206
ReLU	–74.64	–73.99	1550
UReLU	–87.18	–96.41	201

It can be seen that the result of UReLU neural network is quite encouraging. Compared with other method, we have achieved the best accuracy. Besides, the number of parameters employed in the UReLU neural network is also the smallest.

## 7 Conclusions

In this paper, the UReLU neural network is proposed based on the linear transformation and UReLU function, which can be seen as the decoupled ReLU neural network and the interpretability can be shown clearly. The training of the UReLU neural network follows a decoupled strategy, and the easy obtained derivative of the UReLU neural network facilitates the training process. In the simulation study, the UReLU neural network is used to approximate a complex nonlinear system and the performance is excellent.

**Acknowledgement.** This work is jointly supported by the National Natural Science Foundation of China (U1813224), and Science and Technology Innovation Committee of Shenzhen Municipality (JCYJ2017-0811-155131785).

## References

1. Lennart, L.: System Identification: Theory for the User, pp. 1–14. Prentice Hall, Upper Saddle River (1999)

2. Billings, S.A.: Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-temporal Domains. Wiley, New York (2013)
3. Lathi, B.P.: Signal Processing and Linear Systems. Oxford University Press, New York (1998)
4. Nelles, O.: Nonlinear system identification: from classical approaches to neural networks and fuzzy models. Springer, Cham (2013)
5. Westwick, D.T., Hollander, G., Karami, K., et al.: Using decoupling methods to reduce polynomial NARX models. IFAC-PapersOnLine **51**(15), 796–801 (2018)
6. Schoukens, J., Vaes, M., Pintelon, R.: Linear system identification in a nonlinear setting: nonparametric analysis of the nonlinear distortions and their impact on the best linear approximation. IEEE Control Syst. Mag. **36**(3), 38–69 (2016)
7. Abdalmoaty, M.R., Hjalmarsson, H.: Application of a linear PEM estimator to a stochastic Wiener-Hammerstein benchmark problem. IFAC-PapersOnLine **51**(15), 784–789 (2018)
8. Dreesen, P., Westwick, D.T., Schoukens, J., et al.: Modeling parallel Wiener-Hammerstein systems using tensor decomposition of Volterra kernels. In: International Conference on Latent Variable Analysis and Signal Separation, pp. 16–25. Springer, Cham (2017)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
10. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of ICML, p. 3 (2013)
11. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
12. Clevert, D.-A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint [arXiv:1511.07289](https://arxiv.org/abs/1511.07289)
13. Dreesen, P., De Geeter, J., Ishteva, M.: Decoupling multivariate functions using second-order information and tensors. In: International Conference on Latent Variable Analysis and Signal Separation, pp. 79–88. Springer, Cham (2018)
14. Golub, G., Pereyra, V.: Separable nonlinear least squares: the variable projection method and its applications. Inverse Prob. **19**(2), R1 (2003)
15. Billings, S.A., Korenberg, M.J., Chen, S.: Identification of non-linear output-affine systems using an orthogonal least-squares algorithm. Int. J. Syst. Sci. **19**(8), 1559–1568 (1988)
16. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**(3), 455–500 (2009)
17. Vervliet, N., Debals, O., Sorber, L., et al.: Tensorlab 3.0
18. Noël, J.P., Schoukens, M.: Hysteretic benchmark with a dynamic nonlinearity. In: Workshop on Nonlinear System Identification Benchmarks, pp. 7–14 (2016)
19. Belz, J., Münker, T., Heinz, T.O., et al.: Automatic modeling with local model networks for benchmark processes. IFAC-PapersOnLine **50**(1), 470–475 (2017)
20. Xu, J., Tao, Q., Li, Z., et al.: Efficient hinging hyperplanes neural network and its application in nonlinear system identification. Automatica **116**, 108906 (2020)



# Development of Multiply Magnetic Field Generator Combined with Living Cell Workstation

Jiansheng Xu<sup>(✉)</sup>, Chuanfang Chen, Deyu Kong, Linfei Ye, and Ming Xu

Institute of Electrical Engineering, Chinese Academy of Sciences, Beijing, China  
jsxu@mail.iee.ac.cn

**Abstract.** With the development of society, people pay more and more attention to their health. More and more attention has been paid to the study of the effect of magnetic field on life. In this paper, it is proposed to add a variety of magnetic field generating devices such as uniform magnetic field, rotating magnetic field and swinging magnetic field on living cell workstation to study the long-term effects of various magnetic fields on living cells. The maximum value of the developed magnetic field strength can reach 10 mT, the frequency of rotating and swinging magnetic field can be 100 Hz, and the higher frequency can be achieved when the magnetic field strength is reduced.

**Keywords:** Living cell workstations · Rotating magnetic field · Uniform magnetic field · Electromagnetic devices

## 1 Introduction

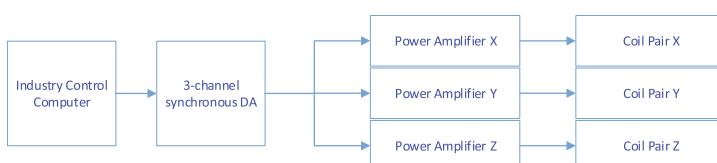
With the development of electrical technology, human beings are taking advantage of a variety of electromagnetic fields technology and the equipment to promote life and production progress. At the same time, there are more and more threats to health for the exposure to complex electromagnetic environment were recognized by the people [1–5]. To research the bioeffects of the magnetic fields, various electromagnetic devices have been developed in the laboratory to simulate different magnetic fields. Dimici et al. gave the bioeffects on growth and biomass composition [6]. Falone et al. explored the function of improved mitochondrial and methylglyoxal-related metabolisms support hyperproliferation induced 50 Hz magnetic field in neuroblastoma cells [7]. Pan et al. studied the effects on the formation of magnetosomes in Magnetospirillum sp. strain AMB-1 by 50 Hz, 2 mT pulsed electromagnetic field [8]. Sun et al. investigated the effects of extremely low frequency magnetic fields on circadian rhythms of cryptochrome in mouse embryonic fibroblast cell [9]. Yang et al. reported the bioeffect of the cochlear stria marginal cells exposed to 1,800 MHz mobile radiofrequency radiation [10].

With the deepening of cell research, real-time observation of the effect of magnetic field on living cells have gradually become a hot topic. The results are of

great significance to reveal the relationship between living cells and environmental magnetic field. The living cell imaging system is mainly used for long-term collection and timing shooting of white light or fluorescence microscopic imaging of living cells under the condition of in vitro simulation. It can realize qualitative and quantitative analysis at the cell and molecular level, living cell image processing and dynamic tracing of living cells. So it plays an important role in the research of life science. In order to meet the experimental requirements of living cell microimaging system to the greatest extent, living cell microimaging system includes high configuration inverted electric fluorescence microscope, living cell culture environment control system and high-speed and high-sensitivity image acquisition device. Real time imaging of living cells has become an important experimental means to promote the most active research of life sciences such as cell biology, neurobiology and developmental biology. However, the existing living cell workstation has no magnetic field generating device, so it is impossible to directly study the effect of magnetic field on cells under the microscope for a long time. Therefore, this paper proposes to develop an accurate three-dimensional magnetic field generation and control device which can be combined with living cell workstation.

## 2 Scheme Design

There are many ways to realize magnetic field, such as using permanent magnet, electromagnetic coil, superconducting magnet and so on. In order to facilitate the realization of various forms of magnetic field and increase the flexibility of magnetic field regulation, we adopt the form of electromagnetic coil. The signal generating module is used to generate multiple kinds of current signals, and the corresponding coils are connected after amplifiers to generate the required magnetic field in the center of the microscope. The overall system design block diagram is shown in Fig. 1. The existing microscopes are generally commercially purchased special equipment, which does not reserve space for the external magnetic field device. We need to design the mechanical scheme in combination with the specific microscopes. In order to generate a given magnetic field in the observation area of the microscope, two mechanical design schemes are proposed. One is to embed three pairs of orthogonal coils directly on the stage. The advantage of this scheme is that the current needed to generate the magnetic field is small and the power consumption is low, but at the same time, there are also disadvantages, such as limited operating space, to avoid the space where the cell



**Fig. 1.** System composition block diagram

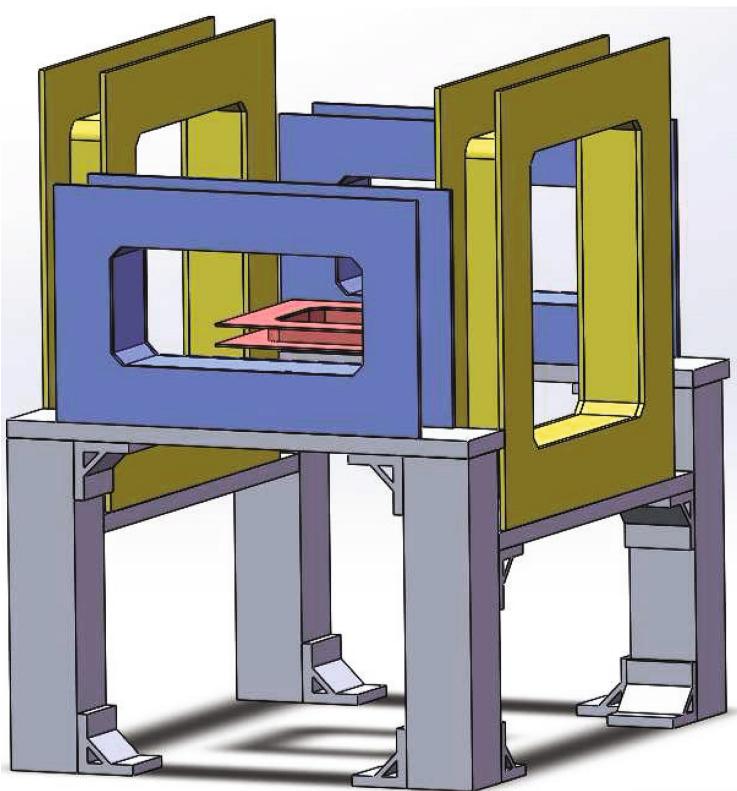
culture chamber is located, so it is necessary to modify the three-dimensional electric stage, which will directly affect the structure and performance of the stage. The other is to place two sets of orthogonal coils of the horizontal axis outside the microscope, and to place two coils above and below the electric stage. The advantage of this scheme is that it has enough operating space, but the problem is that with the increase of coil size, the currents needed to produce a given magnetic field become larger and the power consumption becomes higher. After comparing the advantages and disadvantages of the two schemes, we adopt the second scheme to construct the required system. The microscope of our living cell workstation is the inverted microscope ix83 of Olympus. After measurementing the available space and doing some magnetic field simulation optimization, we design the mechanical structure shown in Fig. 2. Taking the observation point of microscope objective as the coordinate origin, three sets of mutually orthogonal coils are respectively three axes of the coordinate system. The Cartesian coordinate system  $OXYZ$  is established, in which the axis of front and rear coil pairs is  $OX$ , the axis of left and right coil pairs is  $OY$ , and the axis of upper and lower coil pairs is  $OZ$ . The directions of the three axes conform to the right-hand rule.

For the uniform magnetic field generator, Helmholtz coils are usually used according to the requirements of magnetic field uniformity. In order to facilitate the mechanical installation, we directly design three sets of parallel rectangular coils. Because the two horizontal coil pairs are installed on the periphery of the microscope so their sizes are large, they can not be simply equivalent to current rings. We use Maxwell to simulate the magnetic field, and optimize the uniform range of the magnetic field with a 35mm diameter petri dish at the observation point of the microscope. After many times of optimization simulation, the optimal parameters are selected. Take the left and right coils for example, the magnetic field simulation is shown in Fig. 3. The maximum and the minimum of the magnetic field are 10.723 mT and 10.511 mT.

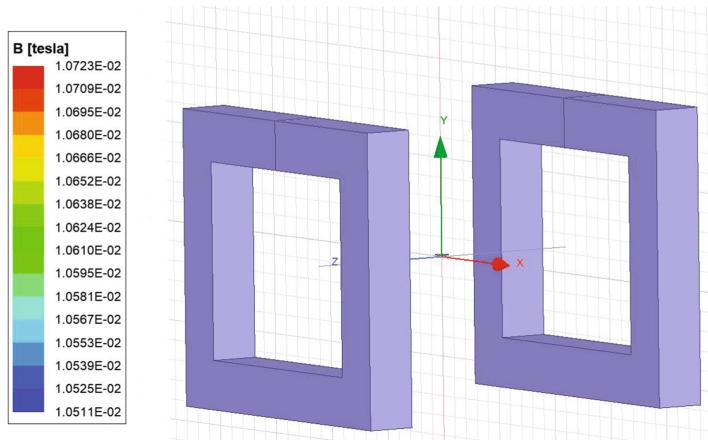
### 3 Circuit Design

In order to realize a variety of different magnetic fields, we use an industrial computer and a USB card to generate the required current waveforms. In the overall scheme, the currents in three pairs of coil are controlled. The USB3020 data acquisition card of Beijing Altai Technology Co., Ltd. is selected. The board can realize 4 channels of 16bit DA synchronous output, with each channel having 256 kW RAM, and the output frequency within 1 mHz–1 MHz. Three specified current amplifiers are used to amplify the currents.

To use the USB3020 data acquisition card, first to assign the parameters in the parameter structure of the board, include output range, frequency, loop-count, triggermode, triggersource, triggerdir bsingleout, clocksource. Then to initialize `USB3020_initdeviceda(hdevice, segmentcount, segmentinfo(0), dapara, ndachannel);` according to the current waveform that needs to be generated, place the current waveform data into the array `dabuffer()` according to a certain



**Fig. 2.** The structure of three pairs of coils



**Fig. 3.** Magnetic field simulation of Y coils pair

sampling rate, and then use `USB3020_writedevicebulkda(hdevice, dabuffer(0), nwritesizewords, nretsizewords, ndachannel)` to writes the current waveform data to the channel where the current waveform needs to be generated.

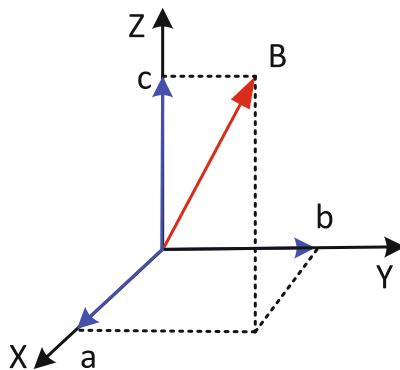
## 4 The Realization of Multiple Magnetic Fields

According to the Biotsavart's Law, the magnetic field near the coordinates origin, i.e. observation area of microscope produced by three pairs of coils is directly proportional to the currents passing through the coils. The magnetic field components  $B_x, B_y, B_z$  on the three axes are

$$B_x = k_x i_x, B_y = k_y i_y, B_z = k_z i_z. \quad (1)$$

Where  $k_x, k_y, k_z$  are the magnetic field scale factors of the three pairs of coils, and  $i_x, i_y, i_z$  are the currents of the three pairs of coils. For each pair of coils, the positive direction of current is defined as the direction of current which generate a positive magnetic field.

The generation method of uniform magnetic field. When the amplitude  $B_0$  of the magnetic field needs to be generated, the direction vector of the magnetic field is  $(a, b, c)$ , as shown in Fig. 4. According to the vector decomposition method, the magnetic field is decomposed into three coordinate axes, i.e. the magnetic field size  $B_x, B_y, B_z$  that three pairs of coils need to generate. Then the magnetic field formula generated by the coils can be used to deduce the current in the three pairs of coils  $i_x, i_y, i_z$ . Then industrial control computer can generate the corresponding current through the data acquisition card. The following current control methods of rotating and oscillating fields are the same. The currents corresponding to three pairs of coils are:



**Fig. 4.** Vector decomposition of uniform magnetic field

$$i_x = \frac{B_0 a}{k_x \sqrt{a^2 + b^2 + c^2}} \quad (2)$$

$$i_y = \frac{B_0 b}{k_y \sqrt{a^2 + b^2 + c^2}} \quad (3)$$

$$i_z = \frac{B_0 c}{k_z \sqrt{a^2 + b^2 + c^2}} \quad (4)$$

The production method of a rotating magnetic field. When the amplitude of the designed rotating magnetic field is  $B_0$ , the rotation axis is  $n$ , with its frequency  $\omega$ , let  $n$  have azimuth angle  $(\theta, \varphi)$ , where  $\theta$  be the angle between the direction of the magnetic field rotation axis  $n$  and the positive direction of the  $OZ$  axis, and  $\varphi$  be the angle between the projection of the direction of the magnetic field rotation axis  $n$  on the  $XOY$  plane and the positive direction of the  $OX$ . The currents in the three pairs coils are [11]:

$$i_x = B_0 (\sin \varphi \cos \omega t - \cos \theta \cos \varphi \sin \omega t) / k_x \quad (5)$$

$$i_y = B_0 (-\cos \varphi \cos \omega t - \cos \theta \sin \varphi \sin \omega t) / k_y \quad (6)$$

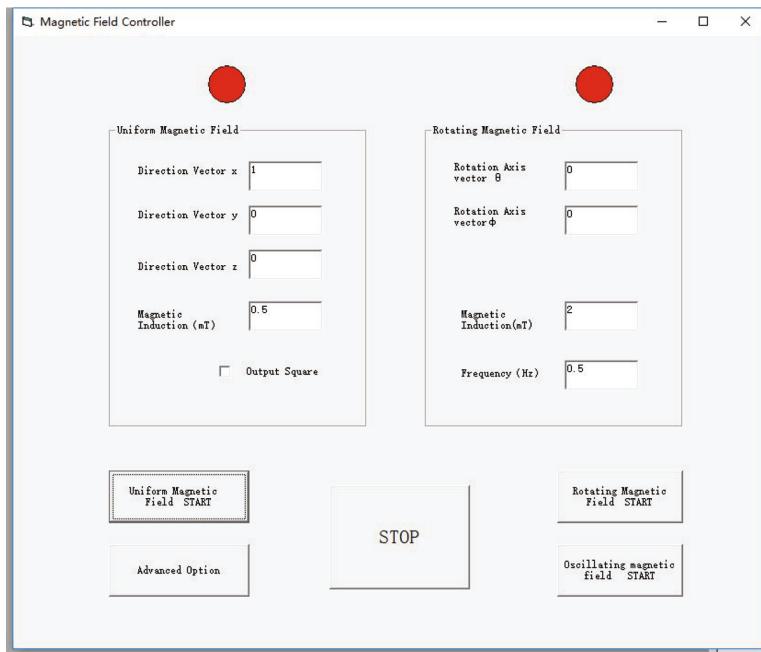
$$i_z = B_0 \sin \theta \sin \omega t / k_z \quad (7)$$

Where  $k_x, k_y, k_z$  are the magnetic field scale factors of the three pairs of coils, and  $i_x, i_y, i_z$  are the currents of the three pairs of coils.

The method of producing swing magnetic field. If a swinging magnetic field is generated between the  $+OX$  and  $+OY$  axes, the realization method is to apply 1/2 cycle of positive current in the  $X$  coil alignment; after the current in the  $X$  coil alignment is turned off, apply 1/2 cycle of positive current in the  $Y$  coil; then apply 1/2 cycle of current in the  $X$  coil; after the current in the  $X$  coil alignment is turned off, apply 1/2 cycle of current in the  $Y$  coil, and then repeat again. In the same way, 1/3 cycle of current can be successively applied in  $X, Y$  and  $Z$  coils. That is to say, the swinging magnetic field among  $+OX$ ,  $+OY$  and  $+OZ$  axes can be realized. Similarly, the oscillating magnetic field of any axes sequence can be realized.

The using method of the data acquisition card is briefly described as the flowing: set up the parameters of the card, initialize the board, output the signal data and control the currents in three pairs of coils, close the output, release the board. The designed user interface on industrial control computer is shown in Fig. 5.

After proper selection of components, the multi magnetic field generating device we designed is achieved as shown in Fig. 6. After the completion of the device, we use F. W. Bell Gaussmeter to measure the magnetic field, including uniform magnetic field, rotating magnetic field and swinging magnetic field. It can be verified that the maximum value of magnetic induction is 10 mT, the rotating magnetic field and swinging magnetic field can be realized with 100 Hz.



**Fig. 5.** User interface



**Fig. 6.** The designed multiply magnetic fields generator

## 5 Conclusion

In order to meet the needs of bioelectromagnetic sensing research, this paper proposes a scheme to load multiple magnetic fields on living cell workstation, develops a multiple magnetic field generating devices, and the experiments results show that the designed magnetic fields can reach the proposed targets. In the future, we will carry out more forms of magnetic field biological experiments.

## References

1. Shuguang, Yu., Peng, S.: A review of bioeffects of static magnetic field on rodent models. *Prog. Biophys. Mol. Biol.* **114**(1), 14–24 (2014)
2. Pilla, A.A., Markov, M.S.: Bioeffects of weak electromagnetic fields. *Rev. Environ. Health* **10**(3–4), 155–169 (1994)
3. International Commission on Non-Ionizing Radiation Protection: ICNIRP statement-guidelines for limiting exposure to time-varying electric and magnetic fields (1 Hz to 100 kHz). *Health Phys.* **99**(6), 818–836 (2010)
4. Bin, Z., Yangli, X., Zhenhong, N., Lin, C.: Effects and mechanisms of exogenous electromagnetic field on bone cells: a review. *Bioelectromagnetics* **41**(4), 263–278 (2020)
5. Soumaya, G., Aida, L., Mohsen, S., Hafedh, A.: Bioeffects of static magnetic fields: oxidative stress, genotoxic effects, and cancer studies. *Biomed Research International* (2013)
6. Deamici, K.M., Cardias, B.B., Costa, J.A.V., Santos, L.O.: Static magnetic fields in culture of chlorella fusca: bioeffects on growth and biomass composition. *Process Biochem.* **51**(7), 912–916 (2016)
7. Falone, S., Santini, S., Di Loreto, S., Cordone, V., Grannonico, M., Cesare, P., Cacchio, M., Amicarelli, F.: Improved mitochondrial and methylglyoxal-related metabolisms support hyperproliferation induced by 50Hz magnetic field in neuroblastoma cells. *J. Cell. Physiol.* **231**(9), 2014–2025 (2016)
8. Pan, W., Chen, C., Wang, X., Ma, Q., Jiang, W., Lv, J., Wu, L., Song, T.: Effects of pulsed magnetic field on the formation of magnetosomes in the Magnetospirillum sp. strain AMB-1. *Bioelectromagnetics*, **31**(3), 246–512 (2013)
9. Sun, Z., Geng, D., Chen, C., Wang, P., Song, T.: The extremely low frequency magnetic fields affected the circadian rhythms of cryptochrome in mouse embryonic fibroblast cell. *Chin. J. Ind. Hyg. Occup. Di.* **6**(35), 459–462 (2017)
10. Yang, H., Zhang, Y., Wang, Z., Zhong, S., Hu, G., Zuo, W.: The effects of mobile phone radiofrequency radiation on Cochlear Stria Marginal Cells in Sprague-dawley Rats. *Bioelectromagnetics* **41**(3), 219–229 (2020)
11. Xu, J.: A rotating magnetic field generating system and its realization method, China, Patent no: CN102820118B



# TIO Loss: A Transplantable Inversed One-Hot Loss for Imbalanced Multi-classification

Lin Wang and Chaoli Wang<sup>(✉)</sup>

University of Shanghai for Science and Technology,  
Shanghai 200093, China  
[clwang@usst.edu.cn](mailto:clwang@usst.edu.cn)

**Abstract.** In image classification, class imbalance is a common problem when training neural networks. It is partly because collecting a great quantity of images in reality is a difficult task. Class imbalanced datasets usually lead to imbalanced learning, especially in multi-classification. In this paper, we introduce our inversed one-hot learning method by inverting the encoding way of labels and present a transplantable inversed one-hot loss, named TIO loss, which can be added to current existing loss functions. Our main idea is using the penalty property of logarithm function by taking the misclassified classes into consideration. We conduct the ablation experiments by adding TIO loss to cross-entropy loss and focal loss. Finally, we verify our method on two imbalanced datasets and the experimental results show the significant improvements.

**Keywords:** Class imbalance · Inverse · One-hot · Loss function

## 1 Introduction

In recent years, convolutional neural networks (CNNs) are gaining significantly importance in the fields of computer vision, including image classification, segmentation, and object detection [1–6]. CNNs have strong advantages of feature learning comparing to artificial extraction of features. However, there are some limitations for CNNs. One of the most challenging limitations is the imbalanced training dataset [7,8]. For imbalanced datasets, some classes have a significantly larger number of samples in training set than other classes. As a result, the balance among training frequencies of different classes will be broken [9]. Previous works [7,8,10] show that the performance of CNNs tends to degrade inextricably when training set exists class-imbalance phenomenon. To solve the problems mentioned above, many great researches have been studied [11].

CNN is one of the supervised learning methods, which means the performance of network heavily relies on numbers of accurately labeled data [3]. However, the real world datasets inevitably exist class imbalance problem. In an imbalanced dataset, classes can be generally divided into two categories, the major classes

(classes with more training data) and the minor classes (classes with less training data) [7]. Class imbalance problem will make training network difficult to extract features from minor classes. Naturally, training network is harder to classify those classes [13, 16]. As is illustrated in Fig. 1, this difficulty reflecting in the network's output is that the predictions of minor classes return smaller confidence coefficients.



**Fig. 1.** Confidence coefficients for major class and minor class trained with ResNet-18.

The imbalanced training dataset will make the network learn more features from major classes and less features from minor classes. Consequently, the major classes will have the higher probabilities to become the well-classified classes and the minor classes will have the higher probabilities to became the poorly-classified classes [12]. And this challenge is even more critical in multi-classification tasks, because the imbalanced degree of multi-classification are more complicated [13].

The recent studies have aimed to alleviate the challenge of class-imbalanced problems [9, 14]. In general, there can be summarized as two strategies: re-sampling data and re-weighting loss. As re-sampling strategies is one of manipulating datasets methods, three problems may arise. First, under-resampling or over-sampling may discard some useful data and duplicate some useless data. Second, interpolating or synthesizing samples will inevitably includes noise. In addition, re-sampling strategies may make a little difference in the face of extreme class-imbalanced dataset. Re-weighting is a more widely studied strategy, because it is aiming at distributing different losses among different classes without changing the original dataset. However, most of current works are based on one-hot encoding and cross-entropy loss, which only consider the loss of ground truth and ignore the loss of non-ground truth.

In this paper, we introduce inversed one-hot learning method by considering the loss between the labels of non-ground truth and the predictions of non-ground

truth. Based on this point, we further proposed a transplantable inversed one-hot loss (TIO loss) to reduce the impact of class-imbalanced training data in multi-classification. TIO loss has an easy transplantable property to the current existing loss functions.

Our key contributions can be summarized as follows: (1) We propose a novel cost-sensitive TIO loss to deal with the imbalanced multi-classification problem. (2) We show the significant performance improvements on two datasets. (3) We conduct ablation experiments with cross-entropy and focal loss to explain the transplantable property of TIO. We believe our study on imbalanced multi-classification can offer useful guidelines for peers research works.

## 2 Related Works

Most of previous works on imbalanced multi-class classification are focus on two approaches: data re-sampling and cost-sensitive learning [7, 14].

### 2.1 Data Re-sampling

For the re-sampling strategies [15, 16], most of approaches are redistribution the data distribution by over-sampling and under-sampling. Those direct sampling methods are more likely to discard some useful data and duplicate some useless data, which is adverse to the network training. In extreme cases, over-sampling and under-sampling will not work. Other re-sampling approaches, including blending and regenerating images [17] is aiming at increasing the image number of minor classes. However, those novel samples will undoubtedly bring noise into training data [18, 19], and could reduce the presentation ability of models.

### 2.2 Cost-Sensitive Learning

Cross-entropy loss is a good quantitative evaluation between labels and predictions of ground truth and it has a good performance on balanced datasets, such as MNIST [20], CIFAR [21]. However, models trained with cross-entropy loss performance poorly on imbalanced multi-classification datasets. To address this problem, many fantastic researches have been studied. In 2017, Lin et al. proposed the focal loss [5], by reducing the loss value of well-classified samples and making the network focus on the poorly-classified samples. However, like cross-entropy loss, focal loss only considers the evaluation between labels of ground truth and predictions of ground truth. In 2019, Cui et al. [14] proposed CB loss based on effective number of samples. However, none of these methods concentrates on considering the loss between the labels of non-ground truth and the predictions of non-ground truth. One of the reasons is all those methods are based on one-hot encoding learning. The specific approach of one-hot learning is generating a set of all zero vectors (except the position of ground truth label is one) for all input images and a set of all zero vectors (except the position of ground truth prediction is one). After the processing of one-hot encoding, all the input images and output predictions are transformed to the value of zero or one, which is easily to computing the training networks loss.

### 3 Proposed Method

There are three steps to transform cross-entropy loss to inversed one-hot loss. One is inversing labels from one-hot encoding to inversed one-hot encoding. Another is making cross-entropy loss symmetric with  $p = 0.5$  (shown in Fig. 2). Finally, we multiply and sum each of the corresponding elements of the inverse one-hot vector and the symmetric cross-entropy function.

#### 3.1 Inversed One-Hot Vector

In machine learning, one-hot encoding is a widely used encoding method to convert categorical data to integer data (zero or one). An one-hot form is an encoding vector with each element equaling to zero except the label element (equaling to one). One-hot learning, as an important work, makes the loss computing between the input labels and the output predictions feasible and convenient.

**Definition 1.** Let  $h(h = [0, \dots, h_i, \dots, 0], h_i = 1)$  be a one-hot encoding vector, then the inverse one-hot vector can be expressed as:

$$\text{inv}(h) = [1, \dots, h_i, \dots, 1](h_i = 0) \quad (1)$$

Suppose the input label has a one-hot encoding form of  $y = [y_1, y_2, \dots, y_n]$ . We can easily get its inversed one-hot vector as  $\text{inv}(y)$ . Once we define the inverse one-hot vector, we need to consider all the non-ground truth's loss (where the value is one in  $\text{inv}(y)$ ).

#### 3.2 Deformation of Cross-Entropy Loss

Suppose the output prediction is  $p = [p_1, p_2, \dots, p_n]$  ( $n$  is the number of classes and the index of ground truth label is  $i(i \in [0, 1, \dots, n])$ ). Cross-entropy loss is written as:

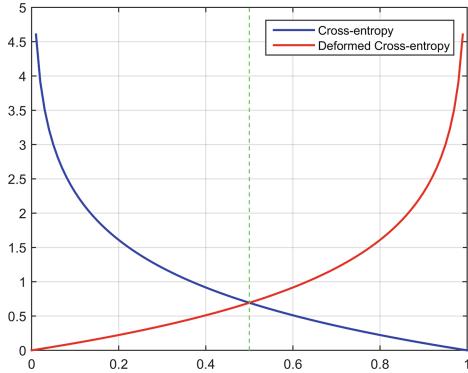
$$CE(p_i) = -\log(p_i) \quad (2)$$

where  $p_i$  is the prediction probability of ground truth.

Figure 2 shows cross-entropy loss function curve. One of the limitations for cross-entropy loss is that it only computes the loss between ground truths label and ground truths prediction. Instead, cross-entropy loss dismisses the information of false labels. Here we use an easy strategy to transform loss of ground truth to loss of non-ground truth by performing an axisymmetric ( $p = 0.5$ ) operations on cross-entropy loss. The reason we directly perform axisymmetric on cross-entropy loss is exponential function has a good property of computing loss. The deformed cross-entropy loss is:

$$DCE(p_j) = -\log(1 - p_j) \quad (3)$$

where  $p_j$  is the prediction probability of non-ground truth.



**Fig. 2.** Cross-entropy loss and deformed cross-entropy loss.

### 3.3 TIO Loss

Once finishing the two steps above, there return an inverse one-hot vector and a deformed cross-entropy function. TIO loss aims at computing loss of all non-ground truth. Therefore, we need to multiply and sum each of the corresponding elements of the inverse one-hot vector and the deformed cross-entropy function, which has the following form:

$$TIO(y, p) = -\text{inv}(y)\log^T(1 - p) \quad (4)$$

The current existing loss functions ignore the losses of misclassified classes. And the inversed one-hot learning shows the penalty of misclassified classes. It is easy to combine the existing loss functions with TIO loss considering the transplantable property of TIO. In this paper, we respectively add TIO to cross-entropy loss and focal loss, which can be shown as:

$$CE + TIO(y, p) = CE(p_i) + TIO(y, p) \quad (5)$$

$$FL + TIO(y, p) = FL(p_i) + TIO(y, p) \quad (6)$$

where  $CE$  represents cross-entropy loss and  $FL$  represents focal loss. This reformulation is motivated by the property cross-entropy loss: (1) Logarithm function has a smooth monotonicity. Specifically, Logarithmic function assigns higher loss if  $p_i$  is relatively smaller and assigns smaller loss if  $p_i$  is relatively larger. (2) For multi-classification, cross-entropy loss only computes difference between ground truth label and prediction probability. Once the training data appears class imbalance phenomenon, neural network have a bigger probability to misclassify the ground truth to a wrong class. In this case, adding a punitive loss to the misclassified class will make the network learn better.

## 4 Experimental Results and Analysis

### 4.1 Dataset Introduction

To verify the effectiveness of TIO loss, we use two datasets including the manually sampled imbalanced CIFAR dataset and the Kaggle Mushroom dataset. Both of datasets exist class-imbalanced problem with different degree and are available on Internet.

#### 4.1.1 Imbalanced CIFAR Dataset

CIFAR is a widely used benchmark datasets in image classification. Many excellent researches are based on them, such as ResNet, CB loss. CIFAR-10 contains 50000 training images and 10000 test images. Each of the ten classes has the 5000 training images and 100 test images with the same  $32 \times 32$  pixels. CIFAR-100 shares all the same training and test images with CIFAR-10 but it is divided into 100 classes in more detail. In this work, we manually sample the CIFAR-10 training set and CIFAR-100 training set to obtain the class-imbalanced datasets. The Imbalanced CIFAR dataset takes the same sampling strategy with [7]. According to the paper, we adopt the imbalance ratio to measure class-imbalance degree. Imbalanced ratio formula is expressed as:

$$\rho = \frac{\max\{S_1, S_2, \dots, S_i\}}{\min\{S_1, S_2, \dots, S_i\}} \quad (7)$$

where  $i$  is the index of each class.  $S_i$  is the number of class  $i$ .  $\max\{S_1, S_2, \dots, S_i\}$  denotes the largest image number of all classes, and  $\min\{S_1, S_2, \dots, S_i\}$  is the smallest image number of all classes. A set of  $\rho$  ( $\rho \in 10, 100, 200, 500$ ) are use to sample the imbalance datasets. The imbalanced CIFAR dataset follows the sampling strategies of power functions below:

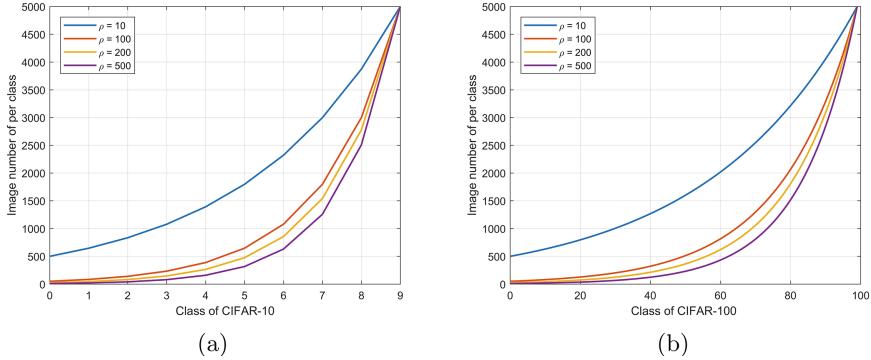
$$N_i = 5000\rho^{\frac{i-9}{9}} \quad (8)$$

$$N_j = 500\rho^{\frac{j-99}{99}} \quad (9)$$

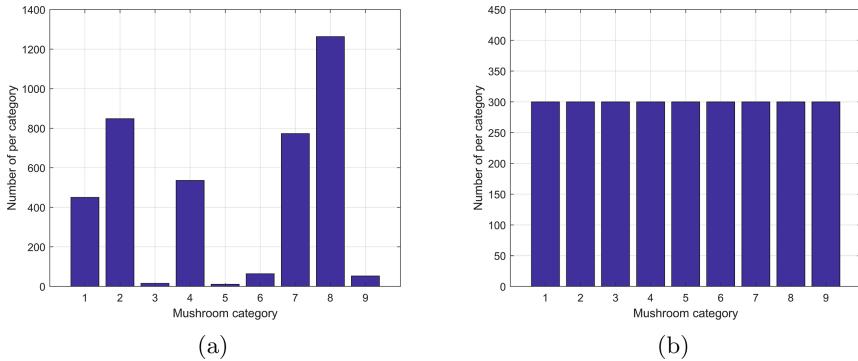
where  $i \in [0, 9]$  ( $i \in N^*$ ) and  $j \in [0, 99]$  ( $j \in N^*$ ) are the label indexes of CIFAR-10 and CIFAR-100.  $N_i$  and  $N_j$  denote the image numbers of each class in the manually sampled imbalanced CIFAR dataset. After the exponential sampling with different  $\rho$ , the distributions of new training set are shown in Fig. 3. In addition, the test set keeps the same size with the original CIFAR dataset without class imbalance phenomenon.

#### 4.1.2 Kaggle Mushroom Dataset

The other class-imbalanced dataset used in our experiments is the Kaggle Mushroom dataset. It is available on <https://www.kaggle.com/uciml/mushroom-classification>. The dataset includes 9 classes, and each class contains around 300 to 1500 images. The test set is defined by selecting 300 images from each class, and the rest part of the dataset constitutes the training set. Figure 4 shows the distributions of the training and test set.



**Fig. 3.** Data distribution of Imbalanced CIFAR training set. Left: CIFAR-10. Right: CIFAR-100.



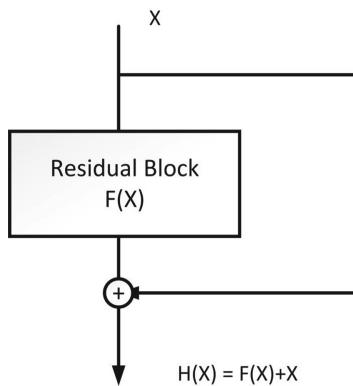
**Fig. 4.** Data distribution of Kaggle Mushroom dataset. Left: Train set. Right: Test set.

## 4.2 Implementation

Deep residual networks (ResNet) [3] is a landmark work to make CNNs come to a deeper network structure. As is shown in Fig. 5, ResNet has jump short-cuts, which effectively overcome gradient disappeared problem with the increasing depth of network. ResNet has gained great success on benchmark datasets including ImageNet, CIFAR-10/CIFAR-100. In this work, we conduct our experiments on ResNet-18.

Our deep learning frameworks are based on PyTorch. Table 1 shows several data augmentation strategies implemented in the training sets.

For the first dataset, we trained our model on a single GTX 2080 GPU for 100 epochs. The batch size is 128 with input size  $32 \times 32$  pixels. The learning rate was initially set to 0.05. Then, the learning rate was divided by 10 at the 60th and 80th epoch. For the second dataset, we trained our model on a single GTX 2080 GPU for 100 epochs. The batch size is 64 with input size  $224 \times 224$



**Fig. 5.** Residual block of ResNet.

**Table 1.** Data augmentations for training sets

Augmentation strategy	Imbalanced CIFAR dataset	Kaggle Mushroom dataset
Resize	✓	✓
Random Crop		✓
Random Horizontal Flip	✓	
Color Jitter	✓	✓
Random Rotation		✓

pixels. The learning rate was initially set to 0.01. Then, the learning rate was divided by 10 at the 60th and 80th epoch.

### 4.3 Experimental Results

#### 4.3.1 Experiments on Imbalanced CIFAR Dataset

A series of extension experiments conducted on Imbalanced CIFAR dataset with various imbalance ratios and the experiments. We present results by using cross-entropy loss, focal loss, TIO loss and its combinations with cross-entropy loss and focal loss. Table 2 and Table 3 show the results on CIFAR-10/CIFAR-100 test set. It can be judged that the performance of single TIO loss is between cross-entropy loss and focal loss. However, the combination of TIO and cross-entropy or focal loss will clearly improve the accuracies for Imbalanced CIFAR dataset.

#### 4.3.2 Experiments on Kaggle Mushroom Dataset

We use top-1 *accuracy* and *F1\_score* as the metrixes. Table 4 shows the results of Kaggle Mushroom dataset. It can be seen the addition of TIO can improve the performance of cross-entropy and focal loss. Table 5 lists the average confidence

**Table 2.** Results on Imbalanced CIFAR-10 dataset

Imbalance ratio	10	100	200	500	Average accuracy
CE	0.8484	0.6559	0.5927	0.4971	0.6485
FL( $\gamma = 0.5$ )	0.8522	0.6644	0.6015	0.5302	0.6621
FL( $\gamma = 1$ )	0.8560	0.6773	0.6224	0.5358	0.6729
FL( $\gamma = 2$ )	0.8514	0.6561	0.6210	0.5209	0.6624
TIO	0.8508	0.6583	0.6011	0.5113	0.6554
CE+TIO	0.8574	0.6745	0.6134	0.5352	0.6701
FL( $\gamma = 0.5$ )+TIO	0.8623	0.6891	0.6296	0.5440	0.6813
FL( $\gamma = 1$ )+TIO	<b>0.8701</b>	<b>0.6962</b>	<b>0.6405</b>	<b>0.5576</b>	<b>0.6911</b>
FL( $\gamma = 2$ )+TIO	0.8669	0.6899	0.6401	0.5534	0.6876

CE represents cross-entropy loss, FL represents focal loss, CE+TIO represents the combination of cross-entropy loss and TIO loss, FL+TIO represents the combination focal loss and TIO loss.

**Table 3.** Results on Imbalanced CIFAR-100 dataset

Imbalance ratio	10	100	200	500	Average accuracy
CE	0.5568	0.3653	0.3249	0.2665	0.3784
FL( $\gamma = 0.5$ )	0.5728	0.3876	0.3403	0.2955	0.3991
FL( $\gamma = 1$ )	0.5777	0.3906	0.3434	0.2972	0.4022
FL( $\gamma = 2$ )	0.5730	0.3856	0.3428	0.3037	0.4013
TIO	0.5689	0.3802	0.3379	0.2877	0.3937
CE+TIO	0.5792	0.3947	0.3498	0.3063	0.4075
FL( $\gamma = 0.5$ )+TIO	0.5887	0.4023	0.3523	0.3155	0.4147
FL( $\gamma = 1$ )+TIO	<b>0.5921</b>	<b>0.4145</b>	<b>0.3669</b>	<b>0.3287</b>	<b>0.4256</b>
FL( $\gamma = 2$ )+TIO	0.5901	0.4111	0.3613	0.3256	0.4220

**Table 4.** Results on Kaggle Mushroom dataset

Loss function	Accuracy	F1_score
CE	0.5829	0.5567
FL( $\gamma = 0.5$ )	0.6219	0.5935
FL( $\gamma = 1$ )	0.6277	0.6080
FL( $\gamma = 2$ )	0.6253	0.6064
TIO	0.6101	0.5787
CE+TIO	0.6254	0.5922
FL( $\gamma = 0.5$ )+TIO	0.6413	0.6213
FL( $\gamma = 1$ )+TIO	<b>0.6452</b>	<b>0.6286</b>
FL( $\gamma = 2$ )+TIO	0.6425	0.6259

**Table 5.** Average confidence coefficients of each class on Kaggle Mushroom dataset

Class of dataset	CE	FL	TIO	CE+TIO	FL+TIO
Amanita	0.5045	0.5156	0.5102	0.5276	0.5323
ressula	0.7134	0.7152	0.7147	0.7206	0.7204
Hygrocybe*	<b>0.1807</b>	<b>0.2043</b>	<b>0.1926</b>	<b>0.2378</b>	<b>0.2515</b>
Cortinarius	0.6355	0.6468	0.6404	0.6517	0.6556
suillus*	<b>0.1296</b>	<b>0.1434</b>	<b>0.1339</b>	<b>0.1765</b>	<b>0.1967</b>
Entoloma*	<b>0.2629</b>	<b>0.2811</b>	<b>0.2707</b>	<b>0.3278</b>	<b>0.3499</b>
Boletus	0.6911	0.7017	0.6964	0.7099	0.7054
Lactarius	0.7401	0.7404	0.7402	0.7411	0.7430
Agaricus*	<b>0.2327</b>	<b>0.2540</b>	<b>0.2424</b>	<b>0.2879</b>	<b>0.3032</b>

**Fig. 6.** Test confidence coefficients of different loss functions on ResNet-18.

coefficients of all classes on test set. We use different loss functions to obtain the predicted confidence coefficients of major class and minor class. Table 5 and Fig. 6 show the additions of TIO loss to cross-entropy loss and focal loss contribute to improving the confidence coefficients of minor class.

## 5 Conclusion

In this work, we present a theoretically and technically sounded TIO loss function to reduce the impact of class-imbalanced training data in multi-classification. From experimental results, TIO loss performances between cross-entropy and

focal loss, while the combination of TIO and cross-entropy loss or the combination of TIO and focal loss performances much better than each single loss. The main idea is to take non-ground truth into consideration. This property allows TIO loss has a good transplantable trait and can be added to some existing loss functions. We argue that TIO loss can be applied to multi-class segmentation or multi-class object detection in further research.

## References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
2. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
4. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
5. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
6. Ren, S., He, K., Girshick, R., Sun J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
7. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **106**, 249–259 (2018)
8. He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**(9), 1263–1284 (2009)
9. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5375–5384 (2016)
10. Drummond, C., Holte, R.C., et al.: C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: Workshop on Learning from Imbalanced Datasets II, vol. 11, pp. 1–8. Citeseer (2003)
11. Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R.: Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Trans. Neural Netw. Learn. Syst.* **29**(8), 3573–3587 (2017)
12. Elkan, C.: The foundations of cost-sensitive learning. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence, 4–10 August 2001, Seattle, no. 1, May 2001
13. Zhou, Z.-H., Liu, X.-Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans. Knowl. Data Eng.* **18**(1), 63–77 (2005)
14. Cui, Y., Jia, M., Lin, T.-Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9268–9277 (2019)

15. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
16. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328. IEEE (2008)
17. Zou, Y., Yu, Z., Vijaya Kumar, B.V.K., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 289–305 (2018)
18. Jia, Y.: Robust control with decoupling performance for steering and traction of 4WS vehicles under velocity-varying motion. *IEEE Trans. Control Syst. Technol.* **8**(3), 554–569 (2000)
19. Jia, Y.: Alternative proofs for improved LMI representations for the analysis and the design of continuous-time systems with polytopic type uncertainty: a predictive approach. *IEEE Trans. Autom. Control* **48**(8), 1413–1416 (2001)
20. Burges, C.J.C., LeCun, Y., Cortes, C.: The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>. Accessed 12 July 2016
21. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Technical report. Citeseer (2009)



# Image Classification Method Based on Generative Adversarial Network

Longhui Hu and Chaoli Wang<sup>(✉)</sup>

University of Shanghai for Science and Technology, Shanghai 200093, China  
clwang@usst.edu.cn

**Abstract.** Image classification algorithms based on deep learning often require a large amount of training data with labels, but sometimes it is difficult to obtain so much training data that meet the requirements. The current methods for data augmentation only operate on the original image and does not change the deep information of the image, so the improvement of models effectiveness is limited. Referring to the idea of CGAN (Conditional Generative Adversarial Networks), we propose an image classification method based on WGAN (Wasserstein GAN). We add category labels to the data, then sent it into WGAN's generator to train them. Finally, the generator can output the specified category samples, and the ability of discriminator is optimized. Since the discriminator has a part to extract image features, Softmax classifier is added to the last layer of the discriminator so that it can output the category of image and whether it is true or false. Both real samples and generated samples are sent to the discriminator during training, so the number of training samples is increased, the accuracy and robustness of classification model are improved, and the convergence speed of network are accelerated. Experiments on MNIST and CIFAR-10 data sets demonstrate the effectiveness of our method.

**Keywords:** Convolutional neural network (CNN) · Category label · Generative adversarial network (GAN) · WGAN · Image classification

## 1 Introduction

The current image classification algorithms mainly include the traditional Machine learning algorithms K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Multilayer Perceptron (MLP), as well as the popular image classification algorithms based on the deep learning and Convolutional Neural Network (CNN) in recent years [11].

The traditional machine learning algorithms (KNN, SVM, MLP, etc.) have a large amount of computation and can only learn the shallow features of images, so the classification accuracy of RGB images with higher pixels is not good. Therefore, image classification algorithms based on deep learning have emerged. Deep learning models include Deep Belief Network (DBN), Convolutional Neural Network (CNN), Recurrent Neural Networks (RNN), and so on. These models

build a deeper level of network, and on this basis, the deeper level of feature information of images and the internal relations implicit in the data can be learned by the computer, so that the features learned can be more expressive and the classification can be more accurate [6].

However, algorithms based on deep learning often need a large amount of labeled training data, and it is difficult to obtain these data in many situations. Therefore, data augmentation method is extremely important in the training of deep learning model [5]. At present, the methods used for data augmentation mainly include random clipping, image inversion, adding noise, etc. But these methods only operate on the original image without changing the deep information of the image, so the improvement of model effect is limited. This paper presents a method of image classification based on WGAN, this method can not only generate a similar image of the sample training set, and improved the diversity of samples compare to other methods based on Generative Adversarial Networks (GAN) [1], and training process more stable. This method generates an image similar to the training sample through a generator, sends it into the discriminator together with the original image, and finally adds the Softmax classifier to classify. Since the sample sent to the discriminator contains both the original training image and the generated image, the training set is expanded. Experiments were carried out on MNIST and CIFAR-10 data sets to verify the effectiveness of the proposed method.

The main contributions of this paper are as follows:

1. Proposed a new image generation method which called CWGAN, its training process is more stable than traditional GAN;
2. Modified the discriminator of WGAN, and added Softmax classifier in the last layer, so that it could recognize the samples and output the classification results;
3. Modified the loss function of WGAN and add a classification loss term to it, so that the classifier parameters can be updated at the same time when the error is propagated back;
4. The samples generated by CWGAN and the original samples are sent into the discriminator for training, which expanded the training samples and higher classification accuracy was obtained.

## 2 Related Work

### 2.1 Image Classification Algorithms Based on Deep Learning

Image classification algorithms based on deep learning is mainly convolutional neural network (CNN). The original model of CNN is Lenet-5 model proposed by LeCun in 1998 [6]. The model uses the gradient-based Back Propagation (BP) algorithm to conduct supervised training on the network, converts the images into feature graphs through a series of convolution pooling operations, and then classifies or recognizes the image feature graphs through the full connection layer. This method has been successful in the classification of handwritten digits [6],

on the basis of which more powerful CNN models such as AlexNet, VGGNet, GoogLeNet and ResNet [8–11] have emerged.

All models based on CNN need a large number of labeled training data sets, which are often difficult to obtain and expensive to label, which has become a major problem restricting deep learning image classification algorithm. In October 2014, Ian J. Goodfellow et al. proposed a new framework to generate the model through the estimation of the confrontation process, that is generate adversarial network(GAN) [1]. GAN has ability to generate new images based on samples, in recent years, more and more GAN models have been used to expand the training data set of the image classification model [4], and use it for unsupervised or semi-supervised learning for image classification [7], or for image style conversion. However, the traditional GAN is prone to model collapse during training, which was also proved by our experiments on MNIST.

## 2.2 Generative Adversarial Network GAN and CGAN

Traditional GAN [1] mainly consists of two parts: generator and optimizer. Firstly, the generator can generate images close to the real sample through optimization. The discriminator is then optimized to distinguish whether the image being sent is original or generated by the generator. Generator and discriminator are trained alternately to optimize the model in the process of game playing. The best probability output of the discriminator is 0.5, that's mean it is impossible to distinguish the real data or the generated data. GAN objective function is formula (1):

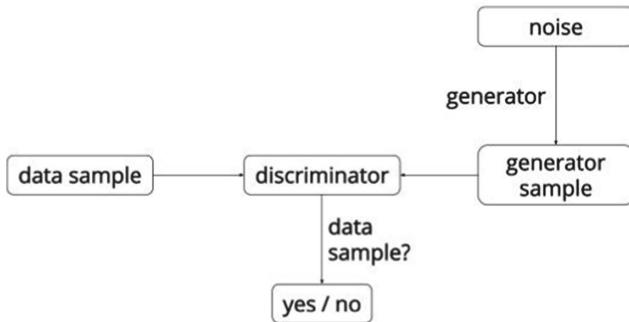
$$\min_G \max_D V(D, G) = P\mathbf{E}_{x \sim P_{data}(x)}[\log D(x)] + \mathbf{E}_{z \sim P_Z(z)}[\log(1 - D(G(z)))] \quad (1)$$

where  $x$  represents real data and  $P_{data}(x)$  represents the distribution of real data,  $z$  represents the noise of input network and  $P_Z(z)$  represents the distribution of noise, which is generally a gaussian distribution.

Traditional GAN can only randomly generate samples similar to the training set according to the distribution of the training set, and cannot generate multiple types of data at one time according to people's requirements. Structure of GAN is shown in Fig. 1. In order to solve the problem that GAN can only generate data randomly from simulated training samples, Mirza [2] proposed a CGAN model. The same category label is added to the generator and discriminator of GAN to guide the generation of samples, so that GAN has the ability to generate multiple types of data. The objective function of CGAN is:

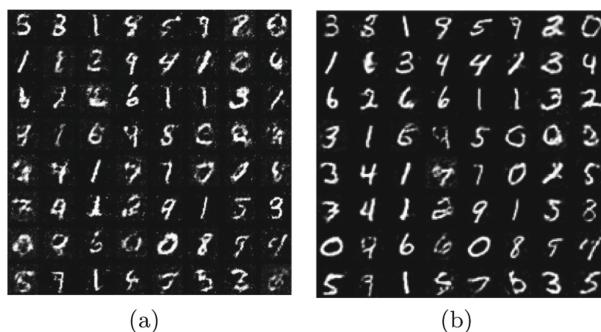
$$\begin{aligned} \min_G \max_D V(D, G) &= \mathbf{E}_{x \sim P_{data}(x)}[\log D(x|y)] \\ &+ \mathbf{E}_{z \sim P_Z(z)}[\log(1 - D(G(z|y)))] \end{aligned} \quad (2)$$

In formula (2),  $y$  is the category label added in the generator and discriminator to control the generation of sample category, and other parameters are consistent with GAN.



**Fig. 1.** Structure of GAN

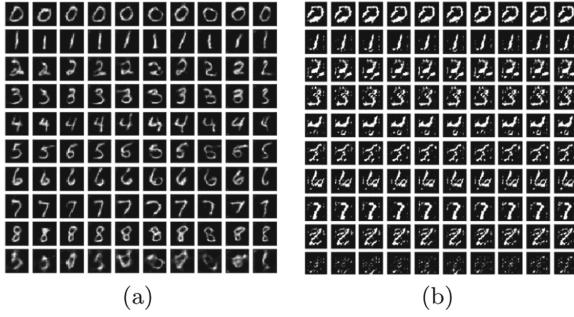
As the loss function of GAN and CGAN is based on JS/KL(jenson-shannon, kullback-leibler) divergence [3,8], the training process is very unstable, and the model training is difficult and even prone to collapse [3,7,9]. We conducted experiments use GAN and CGAN on MNIST data set, and the results were as follows:



**Fig. 2.** The training results of GAN on MNIST

Figure 2(a) shows the results of 10 epoches trained by GAN, and Fig. 2(b) shows the results of 20 epoches trained by GAN. It can be seen that with the increase of training times, although the output picture becomes clearer, but the type of picture generated is completely the same, even the shape of handwritten digits is the same. It can be seen that the model has been collapse.

Figure 3(a) are generated images that CGAN trained 10 epochs. In the training, we found that since the 12th epoch, the images generated were more and more blurred (noise increased), as shown in Fig. 3(b). As a result, Loss increased and the model crashed, resulting in the same shape of the images generated. Thus, it can be seen that the training of GAN was very difficult. Moreover,



**Fig. 3.** The training results of CGAN on MNIST

fuzzy images will have an impact on the accuracy of image classification [2], so we need to consider combining other methods to produce clearer images.

### 2.3 WGAN Based on Wasserstein Distance

The original GAN training is more difficult, because under its optimal discriminant, the generator loss has such problems as gradient disappearance, gradient instability, etc. The main reason for these problems is that the equivalent optimized distance measurement method (JS/ KL divergence) is not reasonable [3]. In order to solve these problems, Martin Arjovsky et al. [3] proposed WGAN in 2017. The core of WGAN is to change the JS/ KL divergence when the original GAN calculates loss to Wasserstein distance, which is also called earth-mover (EM) distance:

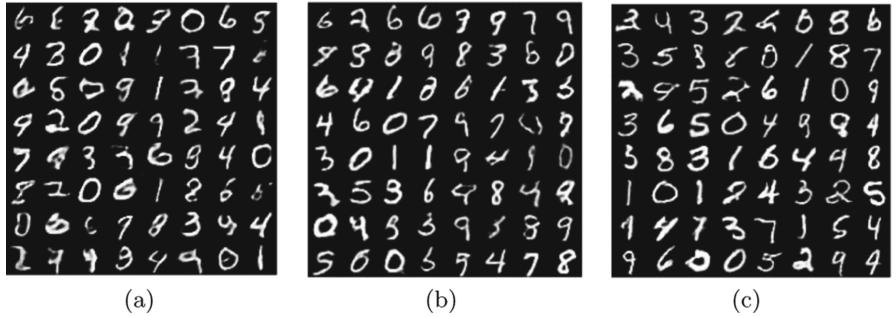
$$W(P_r, P_g) = \inf_{\gamma \sim \prod(P_r, P_g)} \mathbf{E}_{(x,y) \sim \gamma} [\|x - y\|] \quad (3)$$

Here,  $\prod(P_r, P_g)$  is the set of all possible joint distributions combined with  $P_r$  and  $P_g$ , and for each possible distribution, we can sample from  $(x, y) \sim \gamma$  obtain a real sample  $x$  and a generated sample  $y$  and calculate the distance  $\|x - y\|$  between the samples, so you can calculate the expected value  $\mathbf{E}_{(x,y) \sim \gamma} [\|x - y\|]$  of the samples for the distance under this joint distribution  $\gamma$ . It can take a lower bound  $\inf_{\gamma \sim \prod(P_r, P_g)} \mathbf{E}_{(x,y) \sim \gamma}$  on this expected value in all possible joint distributions which is defined as the Wasserstein distance. The optimization function of WGAN is as follows:

$$L = \mathbf{E}_{x \sim P_r} [f_w(x)] - \mathbf{E}_{z \sim P_g} [f_w(g(z))] \quad (4)$$

By minimizing the distance, WGAN can shorten the distribution between the generated data and the real data, which completely solves the problems existing in traditional GAN such as model collapse and lack of diversity.

We also experiment use WGAN on MNIST data set, the results were as follows: In Fig. 4, (a) (b) (c) are the images obtained by WGAN trained 12,20,30



**Fig. 4.** The training results of WGAN on MNIST

epochs on MNIST. It can be seen that with the increase of training times, the image quality generated is getting more higher, and the shape of handwritten digits generated each time is different, its means that there is no model collapse, which shows the advantages of WGAN. However, WGAN also does not have the ability to generate multiple specified samples. Drawing on the ideas of CGAN, in order to enable WGAN to generate multiple samples we want at one time, we add conditional labels to the generator of WGAN and modify its discriminator, so that it can output the classification results while producing the trueness and falsehood of samples. We call this new structure CWGAN.

### 3 CWGAN Model Structure and Training

#### 3.1 CWGAN Model Structure

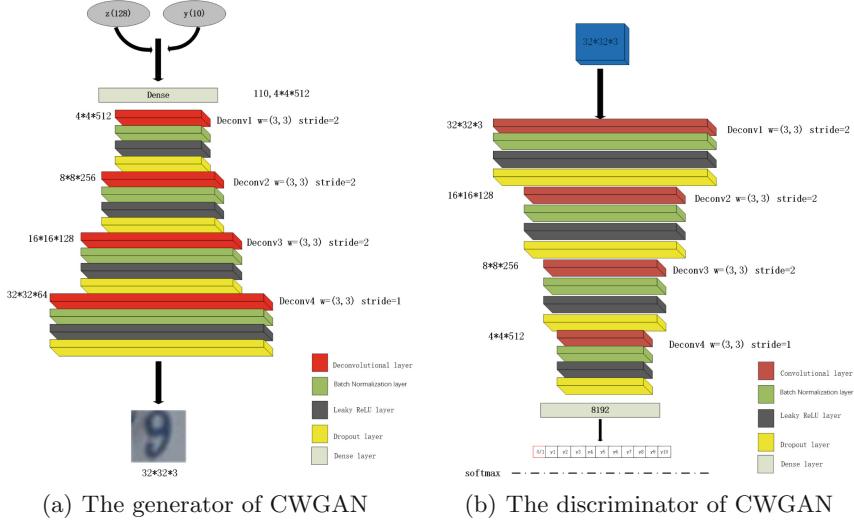
CWGAN designed in this paper combines the advantages of WGAN and CGAN, making it easier to train the model, and at the same time, it can also get samples of specified categories according to the input conditions. The models generator and discriminator are both based on CNN architecture. The input of the model is the random noise of the specified dimension and the conditional features of the corresponding data set. After the generator, the pictures of the corresponding category are output, and then they are fed into the discriminator, finally output the discriminate and classify results of each picture. The structure of CWGAN are show in Fig. 5.

#### 3.2 Model Training

Compared with WGAN's loss function, CWGAN's loss function adds a conditional label to the generator to guide the generation of samples:

$$L_W = \mathbf{E}_{x \sim P_r}[f_w(x|y)] - \mathbf{E}_{z \sim P_g}[f_w(g(z))] \quad (5)$$

The purpose of training is to make it smaller and smaller, and at the same time to show the training process. The smaller the value, the smaller the distance

**Fig. 5.** The structure of CWGAN

between the real data and the generated data distribution of Wasserstein, and the better the model effect.

Since we added a classifier to the discriminator, we also need to construct a classification loss function. The classification loss function in this paper adopts the common cross entropy loss function:

$$L_C = -\hat{y} \log y - (1 - \hat{y}) \log (1 - \hat{y}) \quad (6)$$

where,  $\hat{y}$  is the probability of the corresponding real category of network output, and  $y$  is the label of an image.

By adding Eqs. (5) and (6), the loss function for classification of CWGAN in this paper can be obtained:

$$L_G = L_W + \lambda L_C \quad (7)$$

The purpose of training is to make the loss function as small as possible. Where,  $\lambda$  is a parameter to balance  $L_W$  and  $L_C$ . After many experiments, it is the best while  $\lambda$  is 0.4. In this paper, the RMSProp optimizer is used to optimize the loss function.

## 4 Experimental Results and Analysis

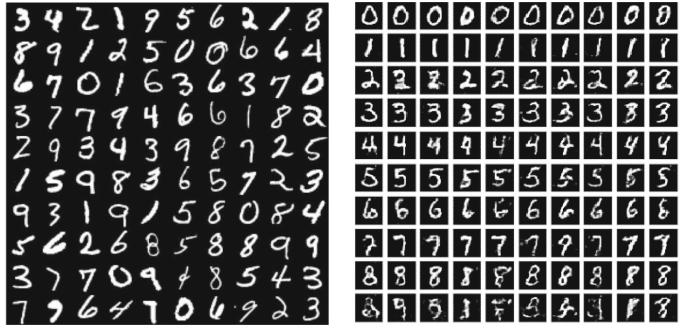
The data sets used in this experiment are the MNIST and CIFAR-10 data sets. The experimental environment is configured as follows: the CPU is Intel Core i7-8250h, the memory is 8GB, and the GPU is NVIDIA GTX1050Ti. Software: the operating system is Windows10 64bit version, the programming framework is

Tensorflow framework based on Python, and the compiler software is Pycharm. The learning rate during training was 0.00005, and the alpha parameter of Leaky ReLU activation function was set to 0.2.

#### 4.1 Experiment on MNIST

MNIST data set [5] is a 28\*28 pixel gray (single color channel) handwritten digital pictures, the training set has 60,000 pictures, the test set has 10,000 pictures.

Figure 6(a) is the sample image of the MNIST data set, and Fig. 6(b) is the MNIST image generated after the 40 epoches trained in this paper. It can be seen that after a certain number of iterations, the generator has been able to generate very smooth and diverse samples.

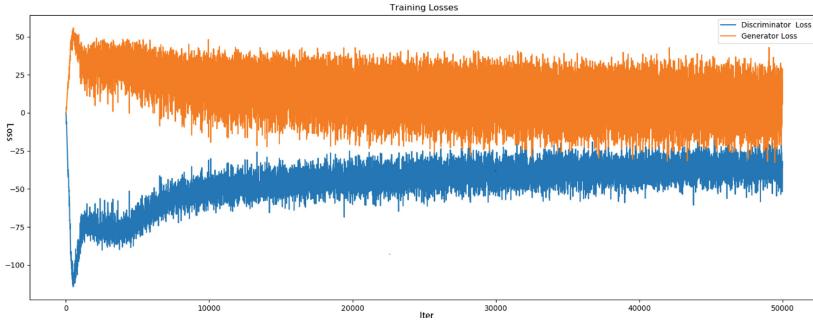


(a) MNIST sample images (b) CWGAN generated pictures

**Fig. 6.** The sample images and generated images of MNIST

In order to show the training process more clearly, we drew the loss curve of generator and discriminator during the training process, as shown in Fig. 7. It can be seen that at the early stage of training, the losses of generator and discriminator are relatively smooth. After several iterations, the model tends to be stable, and generator and discriminator start to fight each other. At the same time, the generator loss is in a decreasing state, while the discriminator loss is in an increasing state, which also conforms to the design idea of CWGAN loss function. The Wasserstein distance between the generated sample and the discriminant sample gradually decreases, it means the value of the loss function gradually decreases and the quality of the generated sample gradually increases.

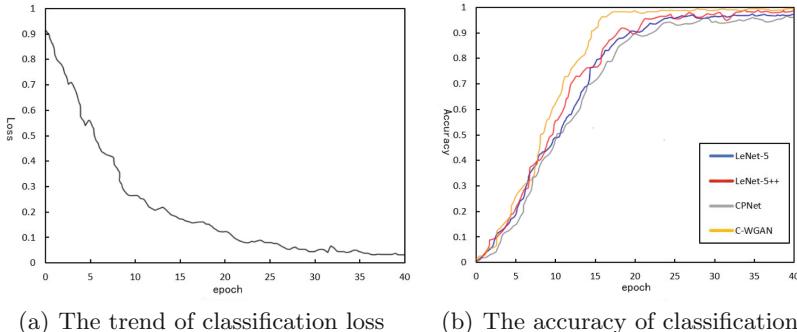
After the original sample and the generated sample are sent into the discriminator, the discriminator will output the classify and discriminate results of the sample, and then the discrimination loss and classification loss will be backpropagated simultaneously, and the generator and discriminator will be updated. The classify loss of CWGAN on the MNIST data set is shown in Fig. 8(a). It can be seen that the loss decreased rapidly in the first 10 epochs, and the loss decreased



**Fig. 7.** The trend of generation and discrimination loss on MNIST

slowly from the 10th epoch and finally tended to converge, it means that the model training was successful.

In order to verify the validity of the model in this paper, we designed a number of control experiment, including LeNet-5 [6] model (the experiment are carried out in advance enhancement and without enhancement the data set), and with in this paper, we designed the discriminant architecture identical CPNet model, and the input data was enhanced, such as cropping, scaling, add noise. The final classification training accuracy curve is shown in Fig. 8(b). It can be seen that the network designed in this paper tends to converge faster, and finally gets a higher classification accuracy than other networks.



**Fig. 8.** The training results on MNIST

At the same time, we extracted the classifier of the trained model and verified it on the test set of MNIST. Table 1 is the final accuracy comparison of these methods on the MNIST test set. Compared with the first three lines, we can see the advantages of this data augmentation method, compared with the fourth and fifth lines, we can see the advantages of our network. Generally, it can be seen that the classification accuracy of the method in this paper is improved compared

with other traditional network structures and data augmentation methods, which also proves the effectiveness of the method in this paper.

**Table 1.** Comparison of each methods on MNIST

Classification method	Data augmentation	Accuracy
LeNet-5	NULL	95.85
LeNet-5	Crop, scale, add noise	96.56
CPNet	Crop, scale, add noise	94.34
CGAN+Softmax	NULL	96.98
WGAN+Softmax	NULL	97.45
<b>CWGAN+Softmax</b>	NULL	<b>98.13</b>

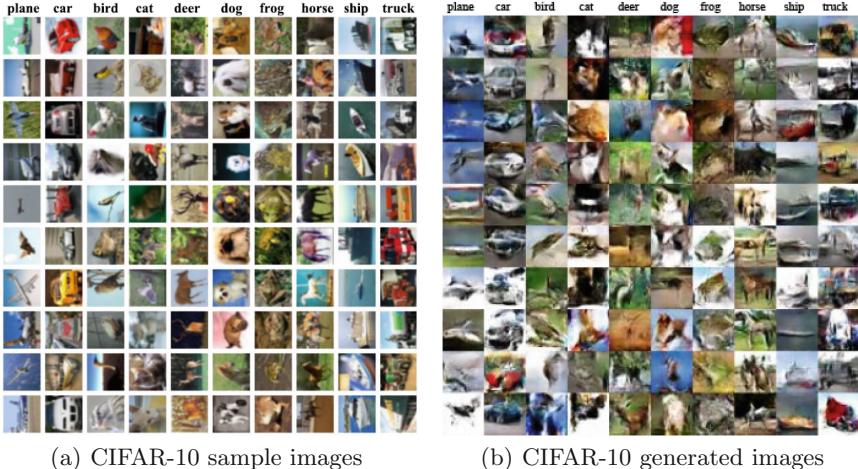
## 4.2 Experiment on CIFAR-10

The CIFAR-10 data set is a small data set for the classification of common objects compiled by Hinton's students Alex Krizhevsky and Ilya Sutskever. There are 10 categories of RGB color images: plane, car, bird, cat, deer, dog, frog, horse, ship and truck. The data set contains a total of 60,000(32 \* 32) color images for these 10 categories, with 6,000 images for each category, of which 5,000 are used for training and 1,000 for testing. The 10 categories are completely mutually exclusive and do not overlap.

Figure 9 shows the sample images of CIFAR-10 and the images generated after 40 epoches were trained in this paper. It can be seen that the generated image is relatively clear and similar to the image of the original data set.

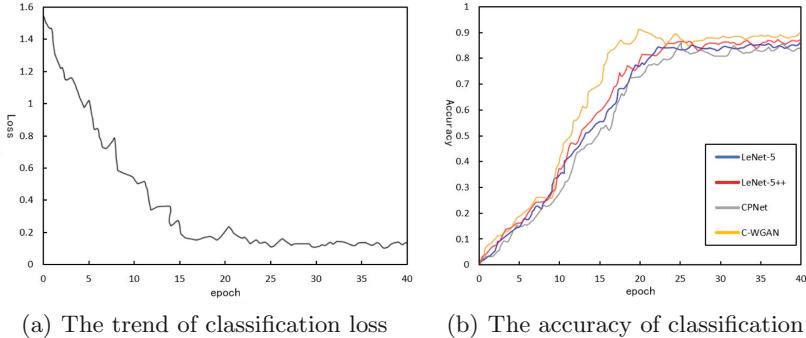
Figure 10(a) shows the loss variation curve of CIFAR-10 classification of the network designed in this paper. The classification loss gradually declined from the beginning, and gradually converged to the 20th epoch. The loss function did not diverge, and the classification loss was stable at about 0.12 in the end.

Figure 10(b) shows the variation curve of accuracy of training on CIFAR-10 when CWGAN designed in this paper is used for classification. We also conducted a number of comparative experiments, and used data enhancement methods such as random flip, clipping, mirroring and adding noise in other networks. However, the method in this paper did not use such data augmentation methods. It can be seen that the classification accuracy of the method proposed in this paper is higher than that of other methods in most of the time, which again proves the good effect of this method in data enhancement.



(a) CIFAR-10 sample images

(b) CIFAR-10 generated images

**Fig. 9.** The sample images and generated images of CIFAR-10

(a) The trend of classification loss

(b) The accuracy of classification

**Fig. 10.** The training results on CIFAR-10

Table 2 shows the comparison between the classification results of the method in this paper and other methods, which is consistent with the method adopted in Fig. 10. After the training of these networks, we conducted tests on the test set of CIFAR-10. Compared with the first three lines, we can see the advantages of this data augmentation method, compared with the fourth and fifth lines, we can see the advantages of our network. As can be seen from Table 2, the classification accuracy of the method proposed in this paper is higher than that of other methods.

**Table 2.** Comparison of each methods on CIFAR-10

Classification method	Data augmentation	Accuracy
LeNet-5	NULL	81.45
LeNet-5	Crop, scale, mirror, add noise	83.56
CPNet	Crop, scale, mirror, add noise	83.28
CGAN+Softmax	NULL	83.85
WGAN+Softmax	NULL	84.62
<b>CWGAN+Softmax</b>	NULL	<b>85.65</b>

## 5 Conclusion

This paper proposes an image classification method based on CWGAN to improve the accuracy of image classification. Compared with traditional network training, CWGAN is more stable, and its classifier converges faster than other traditional networks. Experiments on MNIST and CIFAR-10 data sets demonstrate the effectiveness of the proposed method. In addition, modifying the architecture and layers of this article's generator and discriminator (for example, changing them to U-Net and ResNet-18) can improve the classification effect on high resolution images. Whether the trained generator can be used to expand the data set and achieve better results in the field of object detection and semantic segmentation will be the next research direction.

## References

1. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial networks. In: Proceedings of International Conference on Neural Information Processing Systems, pp. 2672–2680. MIT Press (2014)
2. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint, p. 1784. [arXiv: 1411](https://arxiv.org/abs/1411.1784) (2014)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. arXiv preprint [arXiv: 1701.07875](https://arxiv.org/abs/1701.07875) (2017)
4. Salimans, T., Goodfellow, I., Zaremba, W., et al.: Improved techniques for training GANs. In: Proceedings of International Conference on Neural Information Processing Systems, pp. 2234–2242. MIT Press (2016)
5. LeCun, Y., Cortes, C., Burges, C.J.C.: The MNIST database of handwritten digits, 12 July 2016. <http://yann.lecun.com/exdb/mnist/>
6. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proceedings of IEEE 1998 (1998)
7. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Computer Science (2015)
8. Krizhevsky, A., Sutskever, I., Hinton, G: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 25, no. 2 (2012)

9. Simonvan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
10. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. arXiv preprint [arXiv: 1409.4842](https://arxiv.org/abs/1409.4842)
11. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)



# Design and Development of Integrated Device for Wireless Detection of Flue Gas in Cremation Equipment

Fengguang Huang<sup>1,2,3(✉)</sup>, Lin Tian<sup>1</sup>, and Wei Wang<sup>1,2,3</sup>

<sup>1</sup> 101 Research Institute of Ministry of Civil Affairs, Beijing 100070, China  
huangfengguang@163.com

<sup>2</sup> Key Laboratory of Pollution Control, Ministry of Civil Affairs,  
Beijing 100070, China

<sup>3</sup> Institute of the Ministry of Civil Affairs Environmental Monitoring Center station,  
Beijing 100070, China

**Abstract.** An integrated device for wireless detection of flue gas in cremation equipment is designed in this paper, which consists of a signal acquisition and processing system, a flue gas sampling system and a power supply system. By designing, selecting, constructing and experimental debugging, a more optimized integrated device for flue gas wireless detection is finally formed, which is verified the effectiveness and stability by field tests. This research is an important part of cremation equipment Internet of Things (IoT) and informatization, which provides a certain research basis for the development of intelligence.

**Keywords:** Cremation equipment · Wireless detection · Hardware integration · Internet of Things (IoT)

## 1 Introduction

Recently, with the vigorous development of the information industry, various industries and fields have welcomed the development opportunities brought by IoT and informatization. As a traditional industrial equipment, cremation equipment in funeral and interment industry combines with information industry to realize the IoT, which can promote performance improvement and product upgrading [1]. Cremation equipment mainly refers to cremation machines, relics and sacrifice incinerators in funeral places. Cremation equipment can produce a large amount of toxic and harmful air pollutants such as flue gas dust, sulfur dioxide, nitrogen oxides, carbon monoxide, hydrogen chloride and dioxin during incineration, which is easy to cause harm to the health of the surrounding people [2,3]. Pollutants discharged from incineration of cremation equipment are important funeral and interment pollution sources. The national mandatory emission standard specifies in detail the emission limits of pollutants in flue gas from incineration of cremation equipment [4]. At the same time, due to the low degree of

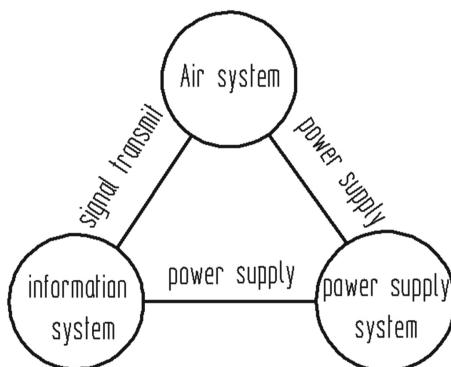
informatization of traditional cremation equipment and other reasons, the operators of funeral parlors cannot intuitively burn the specific values of flue gas emissions during daily operation, which can only be understood when the environmental monitoring personnel are present with special monitoring equipment. Funeral parlor personnel and cremation equipment manufacturers urgently need to obtain the corresponding values of flue gas emission of cremation equipment under various working conditions, so as to improve the operation performance and manufacturing technology of cremation equipment [5,6].

In order to realize the IoT of cremation equipment, an integrated device for wireless detection of flue gas in cremation equipment is designed according to the characteristics of frequent start-stop, intermittent incineration and small emission, which can synchronize with the working state of cremation equipment, continuously sample, analyze and collect the flue gas at the discharge port during operation, and send wireless signals as required.

## 2 Design of Wireless Detection Integrated Device

### 2.1 Overall Design of Wireless Detection Integrated Device

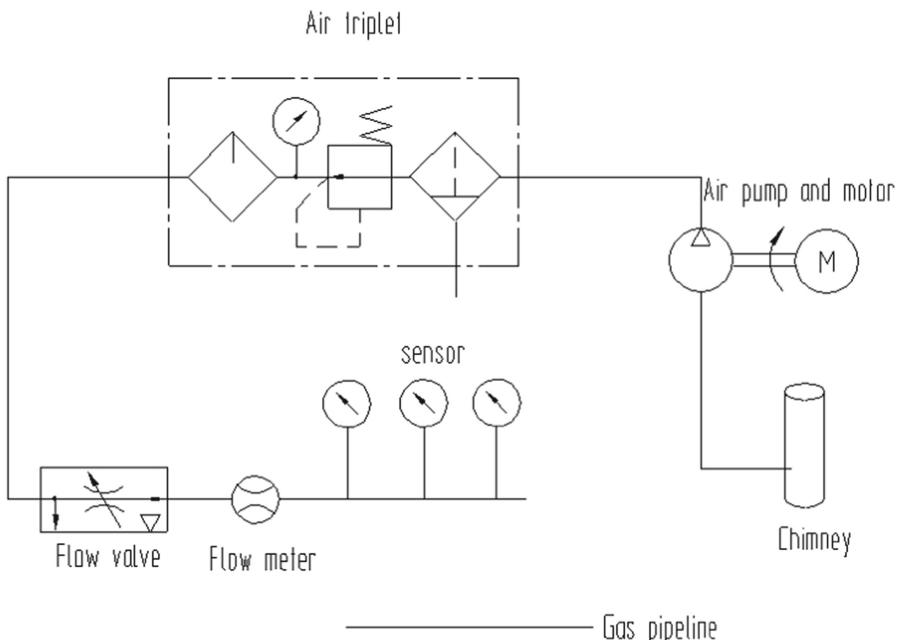
The device is mainly composed of flue gas sampling system, power supply system and signal processing system. The overall device composition is shown in Fig. 1. The flue gas sampling system samples the flue gas generated by combustion of cremation equipment and introduces it to a gas sensor for analysis after treatment. The power supply system provides power and voltage stabilization for the acquisition device. The signal processing system is the core of the whole device. It collects and processes the data signals of sensors and actuators in real time, and sends and receives wireless signals as required.



**Fig. 1.** System composition of centralized acquisition device

## 2.2 Design of Gas Path System for Flue Gas Sampling

The flue gas sampling system samples and collects the flue gas discharged from cremation equipment and introduces it to the gas sensor after treatment. The flue gas discharged from cremation equipment has the characteristics of high temperature, large flow rate, high flue gas concentration and strong corrosiveness. These high temperature, high concentration, high flow rate and high corrosiveness flue gas would reduce the service life of the sensor and even lead to sensor failure. Therefore, pretreatment of flue gas must be considered when designing the gas circuit system.

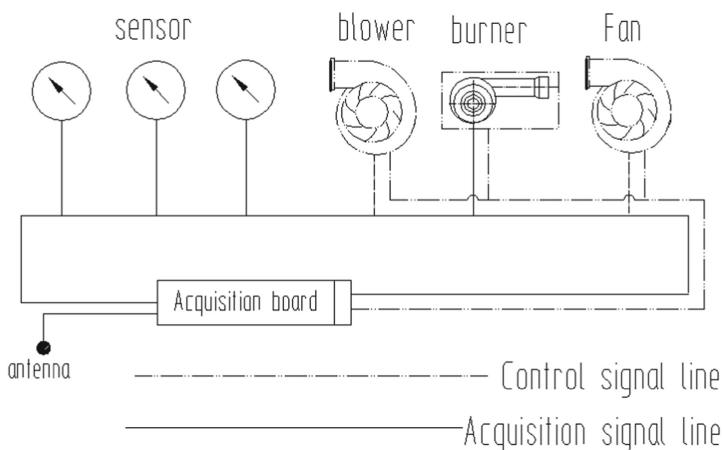


**Fig. 2.** Principle of flue gas sampling system of acquisition device

The flue gas sampling system (Fig. 2) is mainly composed of electric air pump, pneumatic triplet, flow valve, flow meter, gas sensor, gas connection management, etc. The electric air pump extracts flue gas samples from the flue gas discharge port of cremation equipment. After the flue gas samples are adjusted by the pneumatic triple piece to adjust the pressure and remove solid impurities and oil mist, the flue gas samples are adjusted to the appropriate flow rate through the flow regulating valve, and finally flow into the gas detection sensor to analyze and obtain the pollutant value. Table 1 shows some component selection parameters (extensible).

**Table 1.** Selection of partial components of flue gas sampling system (extensible)

No.	Name	Main parameters	Layout position
1	Electric air pump	Flow rate: 12 L/min; Voltage: 24 V; Vacuum: -80 KPa	Inside
2	Pneumatic triplet	Voltage regulation range: 0.1–0.9 MPa; Accuracy of filter element: 40 $\mu\text{m}$	Inside
3	Flow valve and flow meter	Measurement and adjustment range: 6 mL/min–45 L/min; Manual setting	Inside
4	O <sub>2</sub> sensor	Measurement range: 0–30%; Signal: 4–20 mA/RS485	Inside
5	SO <sub>2</sub> sensor	Measurement range: 0–100 ppm; Signal: 4–20 mA/RS485	Inside
6	CO sensor	Measurement range: 0–5000 ppm; Signal: 4–20 mA/RS485	Inside
7	Pipelines and joints	6 mmPU 64 mm	Inside

**Fig. 3.** Information system principle of acquisition device

### 2.3 Design of Signal Processing System

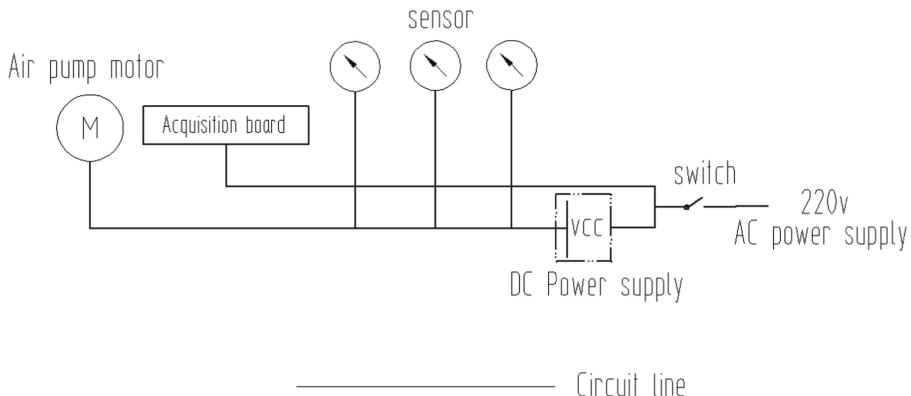
The information processing system uses the signal processing control board to collect signals of sensor parameters arranged in various parts of the cremation equipment, and communicates with the external control end through the wireless transmission module. At the same time, the received control instructions are also sent to the field actuator of the cremation equipment for control, as shown in Fig. 3.

The signal processing system needs to collect pollutant sensor signals, combustion chamber air volume signals, temperature signals, pressure signals, burner status signals, wind speed signals, etc. In order to ensure the stability of the information communication process, RS485 is mainly used for communication. The selection of some actuators and working condition sensors is shown in Table 2 (Extensible).

**Table 2.** Selection of some actuators and working condition sensors (extensible)

No.	Name	Main parameters	Layout position
1	Inverter	PI500	Furnace body
2	Oil valve	DN15, signal: RS485	Tubing
3	Inverter	Schneider ATV61	Post-treatment
4	Temperature sensor	Measurement range: 0 to 1200; Signal: 4–20 mA/RS485	Combustion chamber, pipe, chimney
5	Flue gas flowmeter	0–40/s (customizable); Signal: 4–20 mA/RS485	Pipeline

## 2.4 Design of Power Supply System

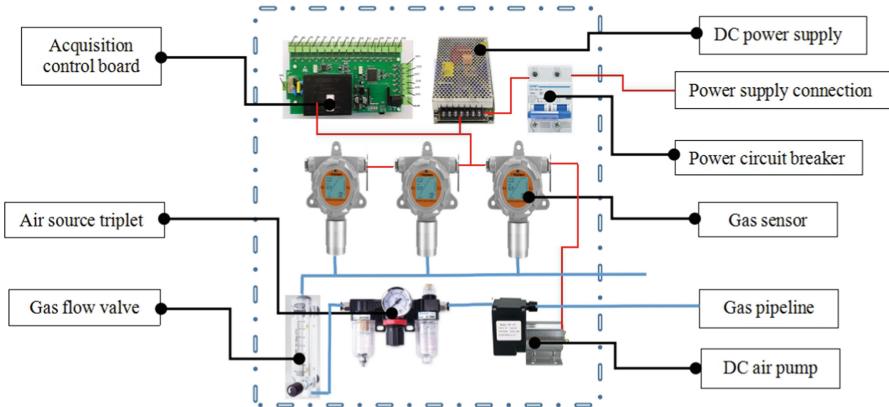


**Fig. 4.** Power supply system principle of acquisition device

## 3 Assembly and Debugging of Wireless Detection Integrated Device

Each of the above devices is trial-produced and assembled. The appearance and internal structure of the actual assembled flue gas wireless detection integrated

device are shown in Fig. 5. The device has simple appearance, compact internal structure and reasonable arrangement, which can ensure stable and safe operation of the system. At the same time, through on-site debugging of the wireless detection integrated device, the experimental results show that the integrated system can effectively collect sensor signals and accurately control the stroke of each actuator.

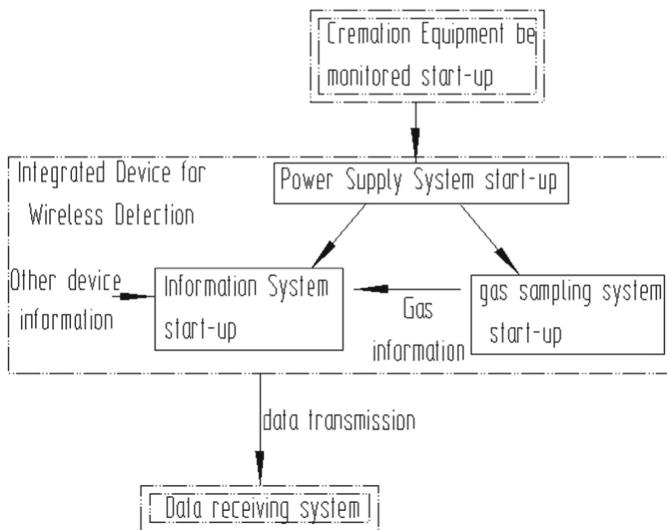


**Fig. 5.** Flue gas wireless detection integrated device

#### 4 The Working Process and Advantages of the Integrated Device

The integrated device for flue gas detection is specially designed for the operation and the discharge parameters of cremation equipment. Its working process is related to the start-up of the cremation equipment. Among others, the start-up and working process are shown as Fig. 6.

When start the cremation equipment be monitored, the power supply system of the integrated device for flue gas detection starts as well. The power supply system provides power for the information system and the gas sampling system for the integrated device so that they can start to work. The gas sampling system will collect the gas at the designated position to the flue gas collection system and then convert to the gas information, which will be conveyed to the information process system so that the gas information and the working information of other equipment can be processed. Then, the processed information will be transmitted to the online data receiving device or be stored to the internal storage of the information process system so as to complete the data collection. Meanwhile, information process system will analyze the orders from online so as to perform feedback operation of the cremation equipment.



**Fig. 6.** The working process of the Integrated Device



**Fig. 7.** The working condition of the Integrated Device

According to the analysis of the working condition (as shown in Fig. 7) traditional cremation equipment, the one with a centralized information collection device has significant advantage than the one without such device, which can be seen in the Table 3.

**Table 3.** The advantage analysis on the equipment with a centralized information collection device and the one without such device

No.	Advantages	The equipment with centralized information collection device	The equipment without centralized information collection device
1	Collect all information during the burning process	Yes	No (Fixed-point observation only)
2	The real-time control of the burning process of the cremation equipment	Yes	No (Manual operation)
3	The storage and transmission of the cremation data	Yes	No
4	Whether the structure of the cremation device will be destroyed	Yes	No

## 5 Conclusion

According to the characteristics of frequent start-stop of cremation equipment, intermittent incineration and small amount of pollutant emission points, an integrated device for wireless flue gas detection is developed. Through the classified design of the flue gas sampling system, signal processing system, power supply system and other devices of the centralized collection device, and the main selection of the main components of each system, the assembly of the complete set of devices is finally completed. Field tests have verified that the data collection of the flue gas wireless detection device is effective and the equipment runs stably. It can be directly installed on the existing cremation equipment to provide intuitive digital display of flue gas emission for field operation, and can also provide research basis for the subsequent IoT and intelligent development of cremation equipment.

## References

1. Huang, F., Tian, L., et al.: Design and implementation of remote monitoring system for working conditions of cremation equipment. In: Proceedings of 2018 Chinese Intelligent Systems Conference; 13–14 October 2018, pp. 259–271. Springer Publishing House (2018)
2. Gautam, J.V., Prajapati, H.B., et al.: Empirical study of job scheduling algorithms in hadoop MapReduce. *Cybern. Inf. Technol.* **17**(1), 146–163 (2017)

3. Huang, F., Wang, J.: Analysis of temperature trend characteristics of multi-group combustion chamber of incinerator in funeral industry. *J. Environ. Eng.* **37**(Supplement), 1088–1090 (2019)
4. Ministry of Environmental Protection of the people's Republic of China. Emission Standard of Air Pollutants for Crematoria GB 13801-2015. Standards Press of China, Beijing (2015)
5. Parham, N., Chu, K.L., Liew, W.S.: Robust remote heart rate estimation from multiple asynchronous noisy channels using autoregressive model with Kalman filter. *Biomed. Signal Process. Control* (2019)
6. Huang, F., Tian, L., et al.: Hardware development of flue gas acquisition monitoring and control system for incineration equipment. In: 2019 Chinese Automation Congress, 22–24 November 2019, pp. 1900–1903. IEEE Publishing House (2019)



# A Hierarchical Fuzzy Comprehensive Evaluation Algorithm for Running States of a Mine Hoist Synchronous Motor Drive System (MHSS)

Wei Liu<sup>1</sup>, Fuzhong Wang<sup>2</sup>, Ao Hou<sup>2(✉)</sup>, and Sumin Han<sup>2</sup>

<sup>1</sup> Guangzhou Railway Polytechnic, Guangzhou 510430, China

<sup>2</sup> School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 45400, China  
651575456@qq.com

**Abstract.** The paper proposes a fuzzy hierarchical comprehensive evaluation method including fuzzy matrix and comprehensive evaluation matrix calculation thought to evaluate the running state of a mine hoist synchronous motor drive system (MHSS). Based on fault analysis of each component of the MHSS, the index system of state evaluation is constructed, including three primary indexes, several secondary and tertiary indexes, etc. A hierarchical fuzzy comprehensive evaluation model is constructed. The paper carried out the simulation experiment for the MHSS with a model number of JKMD-44(Z). The experimental results show that the established state evaluation model can accurately judge the operation state of the MHSS.

**Keywords:** Hierarchical fuzzy · State evaluation algorithm · Degradation degree · Analytic hierarchy process · Synchronous moto · Hoist

## 1 Introduction

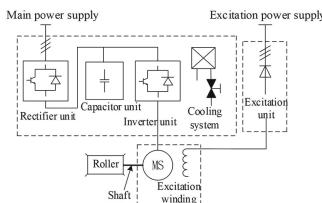
Synchronous motor drive systems have been widely used in large mine hoists. Its operating state is closely related to the safe operation of the mine hoist. Once a fault occurs, it will not only affect safety production of the coal mine, but also endanger the safety of the production personnel [1,2]. In recent years, scholars have studied the fault diagnosis of the hoist, mainly researching the diagnosis technologies and its critical equipment. Using the Internet of Things to diagnose the fault of the hoist and basing on the improved Dezert-Smarandache theory, fault diagnosis reasoning can be performed and remote fault diagnosis can be achieved [3]. The graphical monitoring language of LabVIEW was used to design the condition monitoring and fault diagnosis system of the elevator bearings [4]. Literature [5] Established a three-layer Bayesian intelligent fault inference model (BIFIM) for inverters.

Hierarchical fuzzy evaluation methods are widely used in learning robot control, construction goals and strategies, control systems, wind turbines, and power loads. Literature [6] employed real time state assessment, hierarchical fuzzy evaluation of grid-connected power generation systems to improve operational reliability. A new layered fuzzy evaluation was proposed to realize the forecast of electric power load [7]. Especially in mine hoist system, the study [8] proposed a combination of three-level fuzzy comprehensive evaluation, which has strong practicability. The hierarchical fuzzy comprehensive evaluation method can make an overall evaluation of things or objects that are subject to multiple factors. Therefore, the method is used to comprehensively evaluate the synchronous motor drive system in the paper.

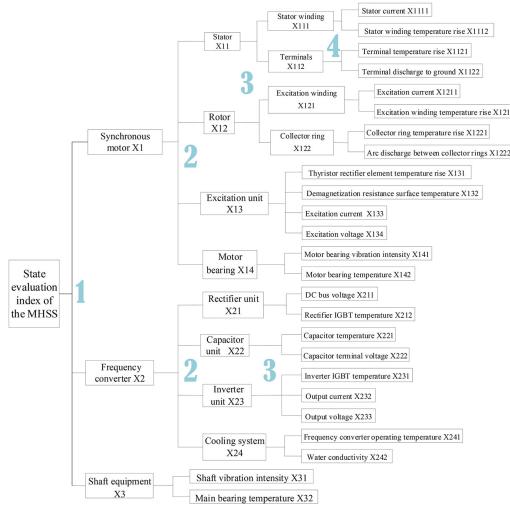
The rest of the paper is organized as follows: Sect. 2 introduces the state evaluation index system of the mine hoist synchronous motor drive system (MHSS). Section 3 analyzes the hierarchical fuzzy comprehensive evaluation model. Section 3.3 researches membership functions and degradation degree. Section 4 demonstrated a case. Section 5 summarizes the paper.

## 2 The State Evaluation Index System of the Mine Hoist Synchronous Motor Drive System (MHSS)

The structure diagram of the MHSS is shown in Fig. 1. It is mainly composed of three links: a synchronous motor, a frequency converter, a shaft and other equipment. The synchronous motor mainly consists of the stator armature winding, the rotor excitation winding, the bearing and excitation system. The cage is lifted and lowered by the main shaft driving drum. The frequency converter comprises a rectifier unit, a filter capacitor unit, an inverter unit and a cooling system to realize motor speed control. The shaft is a component transmitting power. By analyzing the common fault forms of every unit of the MHSS and the relevant provisions of the Coal Mine Safety Engineering for the operation safety of the hoist, the main failures and their characteristics of the MHSS are obtained. The monitoring parameters that can accurately reflect the fault characteristics are selected as the evaluation indexes illustrated in Fig. 2. The index system is divided into four layers from top to bottom. The performance of the upper layer is obtained from the bottom index, and that of each part is obtained step by step. In the end, the state performance of the whole system is obtained.



**Fig. 1.** Structure diagram of the MHSS



**Fig. 2.** The state evaluation index system of the MHSS

### 3 The Hierarchical Fuzzy Comprehensive Evaluation Algorithm

Based on the analysis of the working principle and the causal relationship of faults, the evaluation index system is established. According to the operating conditions and historical data, the weights of evaluation indexes and factors of each layer are allocated. The evaluation factors are standardized for the obtained data. Then, their deterioration degrees (DDs) are calculated. If the values are larger than 0.9, the result will be evaluated as medium or poor. Otherwise, by the corresponding membership functions, the work can calculate the evaluation status with the evaluation algorithm model.

The running state of each system (device or sub-system) can be evaluated by its subordinate indicators through the fuzzy evaluation model. Then the state result can be used as an indicator to judge the state of the upper equipment. The comment collection of the MHSS is divided into four operational status levels:  $V_i = [V_1, V_2, V_3, V_4] \quad i = 1, 2, 3, 4 \quad V_i = [\text{excellent}, \text{good}, \text{medium}, \text{poor}]$ . The description of the four operating status levels is demonstrated in Table 1.

#### 3.1 Weight Collection a of Evaluation Factors

Using analytic hierarchy process (AHP), we can obtain the weight values corresponding to each factor of the MHSS [9, 10]. AHP is a method that combines qualitative and quantitative analysis, which compares the influence degrees of the same level of factors on the previous level of factors. The AHP is applied to calculate the weights of the evaluation indicators of the MHSS, as shown in

**Table 1.** Comments collection of the MHSS.

Year	World population	Components aging degree	Failure possibility	Examination and reparation
Excellent (V1)	Stable	Very low	Very low	No
Good (V2)	Good for a long time	A certain degree	Low	No
Medium (V3)	Normal	No serious damage	Increasing	No
Poor (V4)	Abnormal	High or abnormal	High	Yes

Table 2. The number of evaluation factors which each layer contains decides the one of weight values. The sum of all the weight values of each layer is 1.

**Table 2.** Weight value of evaluation index.

Index	Weight value of the index (Ax)
X1	(0.0837, 0.1385, 0.2328, 0.5450)
X2	(0.3934, 0.1451, 0.3660, 0.0955)
X3	(0.75, 0.25)
X11	(0.4, 0.6)
X12	(0.4, 0.6)
X13	(0.4845, 0.2967, 0.1094, 0.1094)
X14	(0.75, 0.25)
X21	(0.5, 0.5)
X22	(0.6667, 0.3333)
X23	(0.5, 0.25, 0.25)
X24	(0.75, 0.25)
X111	(0.3333, 0.6667)
X112	(0.5, 0.5)
X121	(0.3333, 0.6667)
X122	(0.5, 0.5)

### 3.2 Each Hierarchy Fuzzy Relation Evaluation Algorithm [11, 12]

According to the multi-level idea of AHP, the whole project is divided into several sub projects. The authors must compute the lowest evaluation indicators from the reverse direction. As shown in Fig. 2, the sub indicators for evaluating the operation status of each sub project level are classified into four levels. Especially, we can implement the local comprehensive evaluation of a certain device from the lower level indicators, and gradually get the entire comprehensive evaluation matrix. It can be seen from the above that the bottom fuzzy relation matrix is defined as follows:

$$R_{Xij...k} = \begin{bmatrix} \mu_{Xij...k_{11}} & \cdots & \mu_{Xij...k_{14}} \\ \vdots & \mu_{Xij...k_{lm}} & \vdots \\ \mu_{Xij...k_{n_{ij}...k_1}} & \cdots & \mu_{Xij...k_{n_{ij}...k_4}} \end{bmatrix} \quad (1)$$

Where,  $\mu_{Xij...k_l}$  represents the membership degree of the project,  $X_{ij...k}$ ,  $n_{ij...k}$  represent the number of impact indicators of the item  $X_{ij...k}$ . According to the weight values shown in Table 2, the paper can obtain the fuzzy relation matrixes at the lower level. The comprehensive evaluation model is as follows:

$$B_{Xij} = A_{Xij} \cdot R_{Xij} = A_{Xij} \begin{bmatrix} A_{Xij1} \cdot R_{Xij1} \\ \vdots \\ A_{Xijn_{ij}} \cdot R_{Xijn_{ij}} \end{bmatrix} = \dots \quad (2)$$

Where  $A_{Xij}$  is the weight value of the subproject layer  $X_{ij}$  connected with the parent project. The number of impact indicators of the project  $X_{ij}$  is represented by  $n_{ij}$ .  $R_{Xijk}$  is the next lower fuzzy relation matrix. Obviously, If the fuzzy relation matrix of a certain layer  $R_{Xijk}$  is known, we can evaluate the operation state of the project according to the corresponding weight values  $A_{Xijk}$  in Table 2.

Through the above evaluation process, the final evaluation result set  $B_{xi}(i = 1 \dots)$  can be achieved. Using the maximum membership method, the evaluation element  $V_i$  corresponding to the largest value in the set is taken to obtain the judgment result [13,14]. In other words

$$V = \{V | V_i \rightarrow \max(b_j)\} \quad (3)$$

### 3.3 Degradation Degree (DD) Function

In the evaluation index system established in the paper, the indicators are divided into two categories: the smaller the better type (e.g.. bearing temperature); intermediate type (e.g. excitation current).

(1) For the index of the smaller the better type, the DD function is as follows:

$$g(x) = \begin{cases} 0 & x < x_{\min} \\ \frac{x - x_{\min}}{x_{\max} - x_{\min}} & x_{\min} < x < x_{\max} \\ 1 & x > x_{\max} \end{cases} \quad (4)$$

Where  $x_{\min}$  and  $x_{\max}$  are the threshold values for evaluating the index parameters, respectively.

(2) For the index of the intermediate type, the DD function is as follows:

$$g(x) = \begin{cases} 1, x < x_{\min} \\ \frac{x - x_{\min}}{\alpha - x_{\min}}, x_{\min} \leq x < \alpha \\ 0, \alpha \leq x \leq \beta \\ \frac{x - \beta}{x_{\max} - \beta}, \beta < x \leq x_{\max} \\ 1, x > x_{\max} \end{cases} \quad (5)$$

Where  $x_{\min}$  and  $x_{\max}$  are the upper and lower limit values of the evaluation index parameters;  $\alpha$  and  $\beta$  are the allowable fluctuation ranges for the normal operation of the evaluation index.

## 4 Case Analysis

To verify the effectiveness of hierarchical fuzzy comprehensive evaluation algorithm of the MHSS, the paper took the MHSS with a model number of JKMD-44(Z) in a coal mine in Henan Province, China as an example to analysis. The corresponding technical parameters of the type of hoister are listed in Table 3. Through surveys and measurements, the authors collected the monitoring data of the MHSS in the hoist JKMD-44(Z), where two data simples are demonstrated in Table 4. The evaluation indicators are standardized according to the DD function of each evaluation index, and the DD of each evaluation index is calculated.

**Table 3.** JKMD-44(Z) technical parameters.

Multi-rope friction hoist		AC synchronous motor			
JKMD-4 × 4(Z)		TBP2500-20/3150			
Boost speed	9.4 m/s	Rated power	2500 kW	Rotating speed	45r/min
Maximum static tension	700 kN	Rated voltage	3150 V	Frequency	7.5 Hz
Maximum static tension difference	270 kN	Rated current	484 A	Power factor	$\cos\theta = 1.0$

Taking Sample1 in Table 4 as an example, the running state of the synchronous motor is analyzed according to the calculation process of the hierarchical fuzzy comprehensive evaluation. From Table 4, the stator current 440.3 A and the stator winding temperature rise 75.1 K shown with light blue highlight, the terminal temperature rise and the terminal discharge deterioration value obtained in column 4 in Table 4 are 0.19, 0.72 respectively. The DD values are substituted into the membership function functions under the four operating state levels in Table 5. By Eq. (1), the fuzzy relation matrix can be obtained:

$$R_{X111} = \begin{bmatrix} 0.73 & 0.27 & 0 & 0 \\ 0 & 0.87 & 0.13 & 0 \end{bmatrix}, R_{X112} = \begin{bmatrix} 0 & 0.3 & 0.7 & 0 \\ 0.75 & 0.25 & 0 & 0 \end{bmatrix}$$

From Table 2, the weight sets of  $A_{X111} = (0.3333, 0.6667)$ ,  $A_{X112} = (0.5, 0.5)$  and  $A_{X11} = (0.4, 0.6)$  are got. According to Eq. (2), we can calculate the local comprehensive evaluation state of the stator as  $B_{X11}$ .

$$B_{X11} = (0.4, 0.6) \begin{bmatrix} 0.2433 & 0.6700 & 0.0867 & 0 \\ 0.3750 & 0.2750 & 0.3500 & 0 \end{bmatrix} = (0.3223, 0.4330, 0.2447, 0)$$

**Table 4.** Monitoring data of synchronous motor.

Evaluation index	Sample 1	Sample 2	DD 1	DD 2
Stator current X1111 (A)	440.3	437.5	0.19	0.079
Stator winding temperature rise X1112 (K)	75.1	69.0	0.72	0.66
Terminal temperature rise X1121 (K)	40	35.8	0.57	0.51
Terminal block to ground discharge X1122 (dB)	9	8	0.45	0.4
Excitation current X1211 (X133) (A)	520	518	0.51	0.43
Excitation winding temperature rise X1212 (K)	80	72.4	0.73	0.66
Collector ring temperature rise X1221 (K)	55.8	83	0.62	0.92
Arc discharge between collector rings X1222 (dB)	11	17	0.55	0.85
Thyristor rectifier temperature rise X131 (K)	25.6	21.4	0.56	0.48
Demagnetization resistance surface				
Temperature X132 (	60	52	0.375	0.325
Excitation voltage X134 (V)	115	113.6	0.3	0.037
Motor bearing vibration intensity X141 (mm/s)	1.5	1.3	0.54	0.46
Motor bearing temperature X142 (	0.2	58.7	0.53	0.62

In view of  $B_{X11} = \{b_{111}, b_{112}, b_{113}, b_{114}\}$ ,  $b_{112} = \max(b_{ij})$ , we can judge the stator is in good state by Eq.(3). Likewise, the local comprehensive evaluation state of the rotor can be calculated. The fuzzy relation matrix  $R_{X121}$  and  $R_{X122}$  are:

$$R_{X121} = \begin{bmatrix} 0 & 0.7250 & 0.2250 & 0 \\ 0 & 0.8 & 0.2 & 0 \end{bmatrix}, R_{X122} = \begin{bmatrix} 0 & 0.65 & 0.35 & 0 \\ 0.25 & 0.75 & 0 & 0 \end{bmatrix}$$

$$B_{X12} = A_{X12} \cdot R_{X12} = [0.0750 \ 0.7300 \ 0.1950 \ 0]$$

Since  $b_{122} = \max(b_{ij}) = 0.93$ , we can judge the rotor is in good state by Eq.(3). For the excitation unit, there exit four evaluation factors: temperature rise, demagnetization resistance surface temperature, excitation current and excitation voltage. Therefore, the dimension of the fuzzy relation matrix is  $4 \times 4$ .

$$R_{X13} = \begin{bmatrix} 0 & 0.3000 & 0.7000 & 0 \\ 0.4200 & 0.5800 & 0 & 0 \\ 0.7500 & 0.2500 & 0 & 0 \\ 0 & 0.7250 & 0.2750 & 0 \end{bmatrix}$$

From Table 2, the weight set of  $A_{X13} = (0.4845, 0.2967, 0.1094, 0.1094)$ , we can calculate the local comprehensive evaluation state of the excitation unit as  $B_{X13}$ :

$$B_{X13} = A_{X13} \cdot R_{X13} = (0.2067, 0.4241, 0.3692, 0)$$

By Table 4 and Eqs.(1)–(2), the fuzzy relation matrix  $R_{X14}$  and the local comprehensive evaluation matrix

$$R_{X14} = \begin{bmatrix} 0 & 0.7 & 0.3 & 0 \\ 0 & 0.9 & 0.1 & 0 \end{bmatrix}, B_{X14} = A_{X14} \cdot R_{X14} = (0, 0.75, 0.25, 0)$$

**Table 5.** Membership function of stator current and stator winding temperature.

Comment	Stator current	Stator winding heating
Excellent	$\mu_1(x) = \begin{cases} 1 & x < 0.15 \\ \frac{0.3-x}{0.15} & 0.15 \leq x \leq 0.3 \\ 0 & x > 0.3 \end{cases}$	$\mu_1(x) = \begin{cases} 1 & x < 0.5 \\ \frac{0.7-x}{0.2} & 0.5 \leq x \leq 0.7 \\ 0 & x > 0.7 \end{cases}$
Good	$\mu_2(x) = \begin{cases} \frac{x-0.15}{0.15} & 0.15 \leq x \leq 0.3 \\ \frac{0.65-x}{0.15} & 0.3 \leq x \leq 0.65 \\ 0 & x < 0.15 \text{ or } x > 0.65 \end{cases}$	$\mu_2(x) = \begin{cases} \frac{x-0.15}{0.15} & 0.15 \leq x \leq 0.3 \\ \frac{0.65-x}{0.15} & 0.3 \leq x \leq 0.65 \\ 0 & x < 0.15 \text{ or } x > 0.65 \end{cases}$
Medium	$\mu_3(x) = \begin{cases} \frac{x-0.3}{0.35} & 0.3 \leq x \leq 0.65 \\ \frac{0.9-x}{0.25} & 0.65 \leq x \leq 0.9 \\ 0 & x < 0.3 \text{ or } x > 0.9 \end{cases}$	$\mu_3(x) = \begin{cases} \frac{x-0.7}{0.35} & 0.7 \leq x \leq 0.85 \\ \frac{0.9-x}{0.05} & 0.85 \leq x \leq 0.9 \\ 0 & x < 0.7 \text{ or } x > 0.9 \end{cases}$
Poor	$\mu_4(x) = \begin{cases} 0 & x < 0.65 \\ \frac{x-0.65}{0.25} & 0.65 \leq x \leq 0.9 \\ 1 & x > 0.9 \end{cases}$	$\mu_4(x) = \begin{cases} 0 & x < 0.85 \\ \frac{x-0.85}{0.05} & 0.85 \leq x \leq 0.9 \\ 1 & x > 0.9 \end{cases}$

For the monitoring sample 2 of the synchronous motor in Table 4, the same process is used to evaluate the running state of the synchronous motor. The DD of the temperature rise of the collector ring is  $g = 0.92$ , then it is sure to directly judge the operating state of the device as “poor”. Therefore, the operating state of the synchronous motor corresponding to the sample 2 is “poor”, and it is necessary to immediately stop the inspection and eliminate the fault. The main factor is that the collector ring is overheated, which may be caused by the long time of the carbon brush of the collector ring and the poor contact caused by the wear.

## 5 Conclusion

The hierarchical fuzzy comprehensive evaluation algorithm model including fuzzy matrix and comprehensive evaluation matrix calculation thought is presented and constructed for the MHSS. The authors established the index system of state evaluation composed of four layers and primary indexes, several secondary and tertiary indexes, etc. through the failure analysis of each component of the MHSS. The proposed hierarchical fuzzy comprehensive evaluation algorithm not only evaluate each subsystem according to four operating state levels, but also comprehensively evaluate the global system according to the evaluation results of each system. Moreover, the evaluation process dont acquire too much reliance on experimental data.

**Acknowledgments.** Thank Zhao gu II Mine for providing data support for this research.

**Funding.** This research was funded by the National Key Research and Development Program (2016YFC0600906). Key R & D and Promotion Project of Henan province (Science and Technology) (202102210094).

## References

1. Wang, F.Z., Yao, L., Miao, J.Q.: Fault diagnosis of hoist brake system based on improved particle swarm optimization algorithm. *J. Meas. Control Technol.* **36**, 153–157 (2017). <https://doi.org/10.19708/j.ckjs.2017.07.034>
2. Li, J.L., Yang, Z.J.: Fault diagnosis method of mine hoist based on ontology. *J. Vibr. Test. Diag.* **33**, 993–997 (2013). <https://doi.org/10.16450/j.cnki.issn.1004-6801.2013.06.006>
3. Li, J.L., Xie, J.C., Yang, Z.J., Li, J.J.: Fault diagnosis method for a mine hoist in the internet of things environment. *J. Sensors* **18**, 1920 (2018). <https://doi.org/10.3390/s18061920>
4. Hu, X.L., Xu, J.Y., Zhang, A.C., Sun, Z.: Design of on-line fault diagnosis system for mine hoist based on LabVIEW. *J. Min. Mach.* **38**, 51–54 (2010). <https://doi.org/10.16816/j.cnki.ksjx>
5. Han, S.M., He, Y.S., Zheng, S.Q., Wang, F.Z.: Intelligent fault inference of inverters based on a three-layer Bayesian network. *J. Math. Prob. Eng.* **21**, 1–15 (2019). <https://doi.org/10.1155/2019/3653746>
6. Li, H., Hu, Y.G., Yang, C., Chen, Z., Ji, H.T., Zhao, B.: An improved fuzzy synthetic condition assessment of a wind turbine generator system. *J. Electr. Power Energy Syst.* **45**, 468–C476 (2013). <https://doi.org/10.1016/j.ijepes.2012.09.014>
7. Vellasco, M.M.B.R., Pacheco, M.A.C., Ribeiro Neto, L.S., Souza, F.J.: Electric load forecasting: evaluating the novel hierarchical neuro-fuzzy BSP model. *J. Electr. Power Energy Syst.* **26**, 131–C142 (2004). [https://doi.org/10.1016/S0142-0615\(03\)00060-7](https://doi.org/10.1016/S0142-0615(03)00060-7)
8. Li, J.L., Meng, G.Y., Xie, G.M., Wang, A.M., Ding, J., Zhang, W., Wan, X.W.: Study on health assessment method of a braking system of a mine hoist. *J. Sens.* **19**, 769 (2019). <https://doi.org/10.3390/s19040769>
9. Wang, Y.Y., Chen, M.Y., Jiang, Z.C.: State evaluation of secondary equipment in intelligent substation based on cloud theory. *J. Power Syst. Protect. Control* **46**, 71–77 (2018). <https://doi.org/10.7667/PSPC162115>
10. Leng, H., Tong, Y., Li, X.R.: Research on comprehensive evaluation method of distribution network operation status. *J. Power Syst. Protect. Control* **45**, 53–59 (2017). <https://doi.org/10.7667/PSPC160070>
11. Jin, L.B., Ding, P.J., Zhang, X.L.: Fuzzy Evaluation of Highway Concrete Durability Based on Life Cycle Theory. *J. Xinyang Normal Univ.* **30**, 1003–0972 (2017). <https://doi.org/10.3969/j.issn.1003-0972.2017.03.031>
12. Wang, F.Z., Han, S.M., Cao, B.: A hierarchical fuzzy comprehensive evaluation algorithm for running state of a 6 kV (10 kV) power switch cabinet. *J. Math. Prob. Eng.* **12** (2018). <https://doi.org/10.1155/2018/4517279>
13. Ma, H.Y., Zhou, L., Wang, L.: Evaluation model of equipment health status based on deterioration. *J. Fire Control Command Control* **39**, 66–69 (2014). <https://doi.org/10.3969/j.issn.1002-0640.2014.10.017>
14. Song, R.J., Chen, Y.M.: Fuzzy comprehensive evaluation method for relay protection status based on variable weight coefficient. *Power Syst. Protect. Control* **44**, 46–50 (2016). <https://doi.org/10.7667/PSPC150634>



# Disturbance Observer-Based Design and Analysis of Iterative Learning Control with Nonrepetitive Uncertainties

Zirong Guo<sup>1,2</sup> and Deyuan Meng<sup>1,2(✉)</sup>

<sup>1</sup> The Seventh Research Division, Beihang University (BUAA),  
Beijing 100191, China  
[dymeng@buaa.edu.cn](mailto:dymeng@buaa.edu.cn)

<sup>2</sup> School of Automation Science and Electrical Engineering,  
Beihang University (BUAA), Beijing 100191, China

**Abstract.** This paper considers a general nonsquare multi-input, multi-output iterative learning control (ILC) system with bounded uncertainties, and proposes a disturbance observer-based ILC method. It is shown that ILC with a disturbance observer has a better performance than traditional ILC against nonrepetitive uncertainties. Numerical examples are provided to illustrate the proposed robust ILC conclusions.

**Keywords:** Disturbance observer · Iterative learning control · Nonrepetitive uncertainties

## 1 Introduction

Iterative learning control (ILC) was first proposed to control factory mechanical devices which usually run repetitively. By taking advantage of repetitive operations, ILC can improve the tracking results iteratively, and make the systems have ability to resist disturbance. ILC has a wide range of applications, such as hard disk drives [1] and injection molding processes [2].

Many references have studied different aspects of robust ILC regarding disturbance resistance abilities. Maeda et al. [3] consider a single-input single-output system with near-repetitive disturbances, and combine ILC with a disturbance observer, which mainly focuses on the application of excavation. Hao et al. [4] propose a generalized extended state observer-based indirect-type ILC, and give a sufficient condition to guarantee bounded output tracking results against time- and batch-varying external disturbances. It is summarized that bounded output tracking is a common conclusion, but the bound of tracking error is directly related with the bound of uncertainties. If the uncertainties vary in relatively large scale, the tracking error may not keep in a reasonable range.

---

This work was supported by the National Science Foundation of China under Grant 61873013 and Grant 61922007.

To solve this problem, we use a disturbance observer to give an estimation of tracking disturbance in this paper. A system equivalence transformation method is used to avoid the condition contradiction problem. Theoretical analysis shows that disturbance observer-based ILC can keep not only all the system signals bounded but also the tracking error converging along iteration axis even with non-repetitive uncertainties. Besides, the bound of tracking error is also related to the design parameters, which makes the tracking accuracy more stable than traditional ILC methods.

The rest of this paper is organized as follows. Concerned problems together with the basic assumptions are introduced in Sect. 2. Section 3 contains the ILC design and convergence analysis. Numerical tests and conclusions are provided in Sect. 4 and Sect. 5, respectively.

## 2 Problem Statement

Consider a multi-input, multi-output linear time-invariant system with iteration-varying uncertainties:

$$\begin{cases} x_k(t+1) = Ax_k(t) + Bu_k(t) + w_k(t) \\ y_k(t) = Cx_k(t) + v_k(t) \end{cases} \quad (1)$$

where  $t \in \{0, \dots, T\}$  and  $k \in \mathbb{Z}_+$  are the discrete-time and iteration indexes, respectively,  $x_k(t) \in \mathbb{R}^l$ ,  $u_k(t) \in \mathbb{R}^m$  and  $y_k(t) \in \mathbb{R}^n$  denote the state, input and output variables,  $w_k(t) \in \mathbb{R}^l$  and  $v_k(t) \in \mathbb{R}^n$  denote the input and output disturbances, and  $A \in \mathbb{R}^{l \times l}$ ,  $B \in \mathbb{R}^{l \times m}$ ,  $C \in \mathbb{R}^{n \times l}$  are coefficient matrices.

Let  $y_d(t), t \in \{0, \dots, T\}$  be the desired trajectory, and  $e_k(t) = y_d(t) - y_k(t)$  is defined as the tracking error. Our target is to give a disturbance observer and corresponding control laws to make the tracking error robustly converge along iteration axis, i.e.,  $\limsup_{k \rightarrow \infty} \|e_k(t)\| \leq \beta_e$  with a relatively small bound  $\beta_e \geq 0$ . For further analysis, a boundedness assumption about system uncertainties is put up first.

**Assumption 1:** Uncertainties of the system (1) at any iteration  $k$  and any time  $t$  are bounded, i.e.,

$$\|w_k(t)\| \leq \beta_w, \|v_k(t)\| \leq \beta_v, \|x_k(0)\| \leq \beta_{x0}$$

where  $x_k(0)$  is the initial state value, and  $\beta_w \geq 0, \beta_v \geq 0$ , and  $\beta_{x0} \geq 0$  are some constants.

We leverage a lifting approach on the original system (1) to deduce an augmented system as

$$Y_k = PU_k + N_k \quad (2)$$

where  $N_k = Qx_k(0) + MW_k + V_k$ ,  $Y_k = [y_k^T(1), y_k^T(2), \dots, y_k^T(T)]^T \in \mathbb{R}^{nT}$ ,  $U_k = [u_k^T(0), u_k^T(1), \dots, u_k^T(T-1)]^T \in \mathbb{R}^{mT}$ ,  $W_k = [w_k^T(0), w_k^T(1), \dots, w_k^T(T-1)]^T \in \mathbb{R}^{lT}$ ,  $V_k = [v_k^T(1), v_k^T(2), \dots, v_k^T(T)]^T \in \mathbb{R}^{nT}$ , and  $P \in \mathbb{R}^{nT \times mT}$ ,  $Q \in \mathbb{R}^{nT}$  and  $M \in \mathbb{R}^{nT \times lT}$  are coefficient matrices respectively satisfying

$$P = \begin{bmatrix} CB & 0 & \cdots & 0 \\ CAB & CB & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{T-1}B & CA^{T-2}B & \cdots & CB \end{bmatrix}, Q = \begin{bmatrix} CA \\ CA^2 \\ \vdots \\ CA^T \end{bmatrix}, M = \begin{bmatrix} C & 0 & \cdots & 0 \\ CA & C & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{T-1} & CA^{T-2} & \cdots & C \end{bmatrix}$$

Based on Assumption 1,  $W_k$  and  $V_k$  are bounded, namely,

$$\|W_k\| \leq \beta_W, \|V_k\| \leq \beta_V, \forall k \in \mathbb{Z}_+$$

where  $\beta_W \geq 0$  and  $\beta_V \geq 0$  are some finite unknown bounds. As a result,  $N_k$  is bounded, i.e.,

$$\|N_k\| \leq \|Q\|\beta_{x0} + \|M\|\beta_W + \beta_V \triangleq \beta_N, \forall k \in \mathbb{Z}_+.$$

With the input-output model (2), it is clear to build an observer since there is only one variable which includes all unknown information. In the form of description, the tracking error vector is defined as  $E_k = Y_d - Y_k$ , where  $Y_d = [y_d^T(1), y_d^T(2), \dots, y_d^T(T)]^T \in \mathbb{R}^{nT}$ , and its boundedness is naturally satisfied, namely,  $\|Y_d\| \leq \beta_{Yd}$ . The iteration dynamics of the tracking error can be described as

$$E_{k+1} = E_k - P\bar{U}_k + D_k \quad (3)$$

where  $\bar{U}_k = U_{k+1} - U_k \in \mathbb{R}^{mT}$  and  $D_k = N_k - N_{k+1} \in \mathbb{R}^{nT}$ . The definition of  $\bar{U}_k$  can be regarded as a part of ILC algorithms. The total disturbance  $D_k$  in (3) is guaranteed to be bounded, i.e.,

$$\|D_k\| \leq 2\beta_N \triangleq \beta_D, \forall k \in \mathbb{Z}_+.$$

After figuring out the total disturbance, we can build an observer to give an estimation on it.

### 3 Design and Analysis

Incorporating the ideas in [5], we create the following observer. Consider a disturbance observer for system (3) as:

$$\begin{cases} \hat{D}_k = KE_k - Z_k \\ Z_{k+1} = Z_k + K(-P\bar{U}_k + \hat{D}_k) \end{cases} \quad (4)$$

where  $\hat{D}_k \in \mathbb{R}^{nT}$  is the disturbance estimation,  $Z_k$  is the observer state and  $K$  is the observer gain matrix. This observer uses input and output information

to reconstruct the total disturbance. With the estimation value, we perform a modification on the P-type updating law, i.e.,

$$\bar{U}_k = \Gamma(E_k + \Theta\hat{D}_k) \quad (5)$$

and the ILC algorithm is induced as

$$U_{k+1} = U_k + \Gamma(E_k + \Theta\hat{D}_k) \quad (6)$$

where  $\Gamma \in \mathbb{R}^{mT \times nT}$  and  $\Theta \in \mathbb{R}^{nT \times nT}$  are the learning gain matrices.

To give a complete robust convergence results, a system equivalence transformation approach in [6] will be used. To proceed, a necessary assumption needs to be put forward first.

**Assumption 2:** Matrix  $CB$  is assumed to have full-row rank, i.e.,  $\text{rank}(CB) = n$ .

Assumption 2 actually implies that the matrix  $P$  must have full-row rank, in other words, the number of inputs in system (1) should not be less than that of outputs, namely,  $n \leq m$ . Then, denote  $P = [P_1 \ P_2]$ , where  $P_1 \in \mathbb{R}^{nT \times nT}$  and  $P_2 \in \mathbb{R}^{nT \times (m-n)T}$ . Without loss of generality, let  $P_1$  be a nonsingular matrix. Accordingly, we denote  $\Gamma = [\Gamma_1^T \ \Gamma_2^T]^T$ , where  $\Gamma_1 \in \mathbb{R}^{nT \times nT}$  and  $\Gamma_2 \in \mathbb{R}^{(m-n)T \times nT}$ . Based on Assumption 2, a lemma can be presented to provide the system equivalence transformation results.

**Lemma 1.** Consider the system (2) under the control algorithm (6), if the following condition holds:

$$\rho(I - P\Gamma) < 1, \quad (7)$$

then there exists a nonsingular transformation

$$U_k^* = SU_k \triangleq \begin{bmatrix} U_{1,k}^* \\ U_{2,k}^* \end{bmatrix} \quad (8)$$

where  $U_{1,k}^* \in \mathbb{R}^{nT}$ ,  $U_{2,k}^* \in \mathbb{R}^{(m-n)T}$  and

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} P_1 & P_2 \\ -\Gamma_2(P\Gamma)^{-1}P_1 & I - \Gamma_2(P\Gamma)^{-1}P_2 \end{bmatrix}$$

such that

1)  $U_{2,k}^*$  is iteration-invariant, namely,

$$U_{2,k}^* = U_{2,0}^*, \forall k \in \mathbb{Z}_+. \quad (9)$$

2)  $U_{1,k}^*$  is updated iteratively by

$$U_{1,k+1}^* = U_{1,k}^* + \Gamma^*(E_k + \Theta\hat{D}_k) \quad (10)$$

where the learning gain matrix  $\Gamma^*$  satisfies

$$\Gamma^* = P\Gamma.$$

3) A new ILC system controlled only by  $U_{1,k}^*$  is transformed from the system (2), i.e.,

$$Y_k = P^* U_{1,k}^* + N_k^* \quad (11)$$

where

$$P^* = P \begin{bmatrix} \widehat{S}_{11} \\ \widehat{S}_{21} \end{bmatrix}, N_k^* = N_k + P \begin{bmatrix} \widehat{S}_{12} S_{21} & \widehat{S}_{12} S_{22} \\ \widehat{S}_{22} S_{21} & \widehat{S}_{22} S_{22} \end{bmatrix} U_0$$

and

$$S^{-1} = \begin{bmatrix} \widehat{S}_{11} & \widehat{S}_{12} \\ \widehat{S}_{21} & \widehat{S}_{22} \end{bmatrix} = \begin{bmatrix} \Gamma_1(P\Gamma)^{-1} - P_1^{-1} P_2 \\ \Gamma_2(P\Gamma)^{-1} & I \end{bmatrix}.$$

*Proof.* We can validate the results after applying the transformation (8) into (6) and (2). The detailed proof processes are omitted here due to page limitation.

**Remark 1.** Lemma 1 shows that the input is separated into two parts. The first part  $U_{1,k}^*$  is iteratively updated, and it has the same dimensions with the output. The second part  $U_{2,k}^*$  is iteratively invariant, and its dimensions equal to the dimension difference of input and output. Lemma 1 implies that ILC is a class of one-to-one control method. Besides, it is worth noting that the boundedness of  $N_k^*$  can also be proved by

$$\|N_k^*\| \leq \beta_N + \|P\| \|U_0\| \left\| \begin{bmatrix} \widehat{S}_{12} S_{21} & \widehat{S}_{12} S_{22} \\ \widehat{S}_{22} S_{21} & \widehat{S}_{22} S_{22} \end{bmatrix} \right\| \triangleq \beta_{N^*}, \forall k \in \mathbb{Z}_+.$$

### 3.1 Convergence Analysis of Disturbance Observer

**Theorem 1.** Consider the disturbance observer (4) under the control algorithm (5), and let Assumption 1 and 2 hold. If the condition (7) and the following condition hold simultaneously:

$$\rho(I - K) < 1 \quad (12)$$

then

- 1) the input  $U_k$  and the disturbance estimation  $\widehat{D}_k$  are bounded, i.e.,  $\|U_k\| \leq \beta_U, \|\widehat{D}_k\| \leq \beta_{\widehat{D}}, \forall k \in \mathbb{Z}_+$ , where  $\beta_U \geq 0$  and  $\beta_{\widehat{D}} \geq 0$  are some finite bounds.
- 2) the observer state is bounded, namely,  $\|Z_k\| \leq \beta_Z, \forall k \in \mathbb{Z}_+$  for a finite constant  $\beta_Z \geq 0$ .

*Proof.* 1) From (10) and (11), we can obtain

$$U_{1,k+1}^* = (I - P\Gamma)U_{1,k}^* + \Gamma^* \Theta \widehat{D}_k + \zeta_k^* \quad (13)$$

where  $\zeta_k^* = \Gamma^*(Y_d - N_k^*)$ , and it is bounded, namely,

$$\|\zeta_k^*\| \leq \|P\| \|\Gamma\| (\beta_{Y_d} + \beta_{N^*}) \triangleq \beta_{\zeta^*}, \forall k \in \mathbb{Z}_+. \quad (14)$$

From (3) and (4), we can derive the iterative dynamics of the disturbance estimation

$$\hat{D}_{k+1} = (I - K)\hat{D}_k + KD_k. \quad (15)$$

Rewriting (13) and (15) into one block matrix form, one can get that for any iteration  $k$ :

$$\begin{bmatrix} U_{1,k+1}^* \\ \hat{D}_{k+1} \end{bmatrix} = \begin{bmatrix} I - P\Gamma & \Gamma^*\Theta \\ 0 & I - K \end{bmatrix} \begin{bmatrix} U_{1,k}^* \\ \hat{D}_k \end{bmatrix} + \begin{bmatrix} \zeta_k^* \\ KD_k \end{bmatrix}. \quad (16)$$

If conditions (7) and (12) hold, then we have

$$\rho \left( \begin{bmatrix} I - P\Gamma & \Gamma^*\Theta \\ 0 & I - K \end{bmatrix} \right) < 1. \quad (17)$$

Further, we can obtain that

$$\left\| \begin{bmatrix} \zeta_k^* \\ KD_k \end{bmatrix} \right\| \leq \beta_{\zeta^*} + \|K\|\beta_D, \forall k \in \mathbb{Z}_+.$$

Therefore, both  $U_{1,k}^*$  and  $\hat{D}_k$  are bounded, namely,

$$\|U_{1,k}^*\| \leq \beta_{U_1^*}, \|\hat{D}_k\| \leq \beta_{\hat{D}}, \forall k \in \mathbb{Z}_+ \quad (18)$$

for some finite constants  $\beta_{U_1^*} \geq 0$  and  $\beta_{\hat{D}} \geq 0$ . From previous conclusions, we can derive

$$\|U_{2,k}^*\| \leq \| [S_{21} \ S_{22}] \| \|U_0\| \triangleq \beta_{U_2^*}, \forall k \in \mathbb{Z}_+. \quad (19)$$

Then, it is clear from (8) that

$$\|U_k\| \leq \|S^{-1}\| (\|U_{1,k}^*\| + \|U_{2,k}^*\|) \triangleq \beta_U, \forall k \in \mathbb{Z}_+. \quad (20)$$

- 2) After proving the bounded input result, it can be naturally concluded that all the system signals are bounded, including the system output and the tracking error. Then, if the condition (12) holds, we can derive that the observer state  $Z_k$  is also bounded by following similar steps to the proof of result 1) in this theorem.

**Remark 2.** If we directly analyze the iteration dynamics of input  $U_k$ , we can find a spectral radius condition (see [6]), which can not be fulfilled simultaneously with the condition (7). A system equivalence transformation is used to solve this contradiction problem.

### 3.2 Convergence Analysis of Tracking Error

**Theorem 2.** Consider the system (2) under the control algorithm (6), and let Assumption 1 and 2 hold. If the condition (7) and (12) are satisfied, then

- 1) the tracking error is bounded, i.e.,  $\limsup_{k \rightarrow \infty} \|E_k\| \leq \beta_{E_{sup}}$ .

2) the total disturbance estimation error is bounded as  $k \rightarrow \infty$ , i.e.,  
 $\limsup_{k \rightarrow \infty} \|\tilde{D}_k\| \leq \beta_{\tilde{D}_{sup}}$ .

where  $\beta_{E_{sup}} \geq 0$  and  $\beta_{\tilde{D}_{sup}} \geq 0$  are some finite bounds.

*Proof.* 1) From (3) and (6), we can get the iterative dynamics of the tracking error as

$$E_{k+1} = (I - P\Gamma)E_k + D_k - P\Gamma\Theta\hat{D}_k. \quad (21)$$

If the condition (7) holds, then based on the previous bounded results, we can derive a error convergence result:

$$\limsup_{k \rightarrow \infty} \|E_k\| \leq \frac{\beta_D + \|P\|\|\Gamma\|\|\Theta\|\beta_{\hat{D}}}{1 - \|I - P\Gamma\|} \triangleq \beta_{E_{sup}}. \quad (22)$$

2) The estimation error is defined as  $\tilde{D}_k = D_k - \hat{D}_k$ , and we can get its iteration dynamics as

$$\tilde{D}_{k+1} = (I - K)\tilde{D}_k + (D_{k+1} - D_k). \quad (23)$$

If  $\rho(I - K) < 1$ , similarly, we have

$$\limsup_{k \rightarrow \infty} \|\tilde{D}_k\| \leq \frac{2\beta_D}{1 - \|I - K\|} \triangleq \beta_{\tilde{D}_{sup}}. \quad (24)$$

**Remark 3.** It is worth noting that the bound of the tracking error  $\beta_{E_{sup}}$  can get further reduction through parameter design. If we choose the learning gain matrix  $\Gamma$  as  $P^T(PPT)^{-1}$  and  $\Theta$  as  $I$ , or the matrix  $\Theta$  as  $(P\Gamma)^{-1}$ , then we can get  $\|D_k - P\Gamma\Theta\hat{D}_k\| = \|\tilde{D}_k\|$ . As a result,  $\beta_{E_{sup}}$  evolves to be  $\beta_{\tilde{D}}$ , and it is generally recognized that the bound of the disturbance estimation error is smaller than the bound of disturbances.

**Remark 4.** If we implement a simple transformation to (15), we can derive

$$\hat{D}_{k+1} = \hat{D}_k + K\tilde{D}_k. \quad (25)$$

It is clear the (25) has the same structure of the  $P$ -type updating law. The disturbance observer (4) uses the estimation error of last iteration and a gain matrix  $K$  to update the estimation value iteratively, which is an iterative learning process of the estimation value. (4) gives a reasonable operation structure of the disturbance observer (25).

## 4 Numerical Example

In this section, simulations are implemented to verify the theoretical results. The plant is selected to be a system with 3 inputs and 2 outputs, and system matrices are adopted by

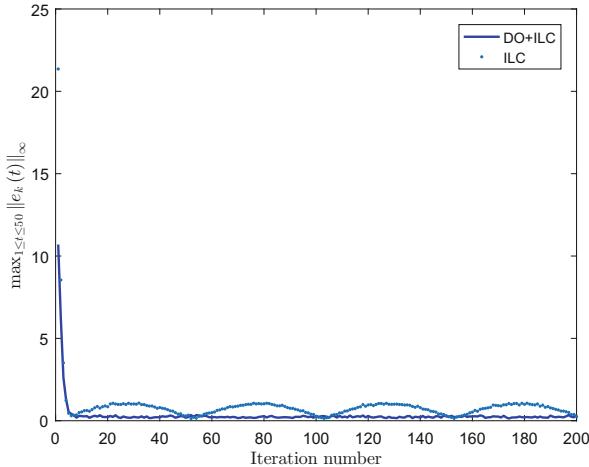
$$A = \begin{bmatrix} 0 & 0.5 & 0 \\ 0.4 & 0.3 & 0.2 \\ 0 & 0.1 & 0.6 \end{bmatrix}, B = \begin{bmatrix} 0 & 1 & 1 \\ 1 & -1 & 1 \\ 0 & 1 & 0 \end{bmatrix}, C = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}.$$

It is clear to see that  $\text{rank}(CB) = 2$ , which satisfies Assumption 2. The desired output trajectory is given by

$$y_d(t) = \begin{bmatrix} 10(0.025t)(1 - 0.025t) \\ -2\sin(0.02\pi t) + \cos(0.01\pi t) - (t/35)^2 \end{bmatrix}$$

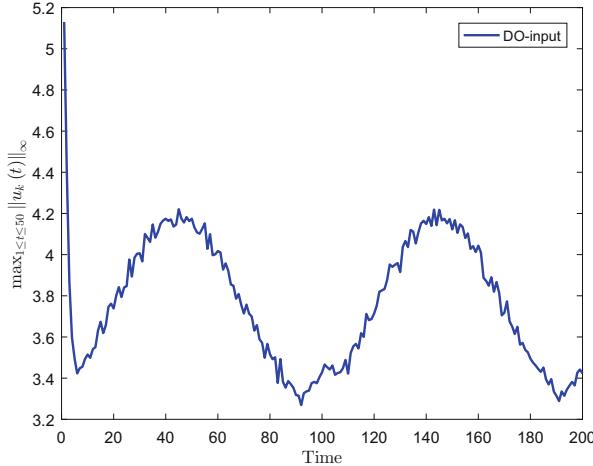
with the operation period  $N = 50$ . The system uncertainties are respectively set to be  $x_k(0) = \delta_{x0}(t, k)$ ,  $v_k(t) = \sin(0.02\pi t) + 2\delta_v(t, k) + 0.5\sin(0.02\pi k)$ ,  $w_k(t) = \sin(0.05\pi t) + 2\delta_w(t, k) + 0.5\cos(0.02\pi k)$ , where iteration-varying uncertainties  $\delta_{x0}$ ,  $\delta_v$  and  $\delta_w$  are appropriately dimensioned, and every element of them varies arbitrarily over interval  $[-0.01, 0.01]$  with respect to the iteration index  $k$  and the time step  $t$ . It can be verified that Assumptions 1 is also guaranteed.

The gain matrices of (4) and (6) are respectively selected as  $K = 0.9 \otimes I_{nN}$ ,  $\Gamma = (PP^T)^{-1}(0.6 \otimes I_{nN})$  and  $\Theta = I$ . It can be verified that conditions (7) and (12) are satisfied. Let the initial input be zero, i.e.,  $u_1(t) = 0, \forall t \in \{0, \dots, N\}$ . Figures 1, 2 and 3 show the simulation results.

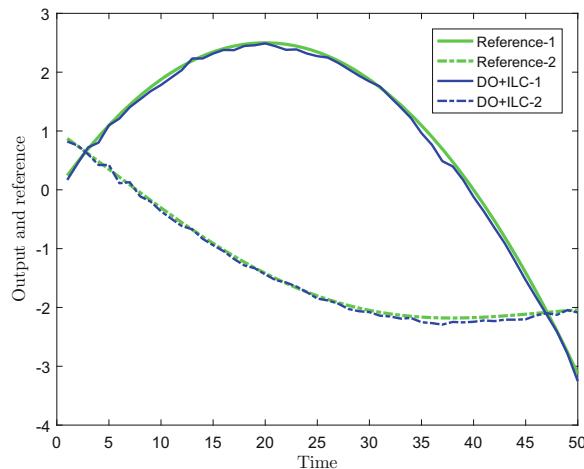


**Fig. 1.** Tracking error evaluated by  $\max_{1 \leq t \leq 50} \|e_k(t)\|_\infty$  for the first 200 iterations.

In Fig. 1, we compare the evolution of the tracking error, which is evaluated by  $\max_{1 \leq t \leq 50} \|e_k(t)\|_\infty$  for the first 200 iterations under the disturbance observer-based ILC algorithm versus that under the traditional ILC algorithm. Figure 1 directly shows that the disturbance observer-based ILC method has a better error resistance ability than traditional ILC methods. Figure 2 depicts the bounded input evaluated by  $\max_{1 \leq t \leq 50} \|u_k(t)\|_\infty$ . From Figs. 1 and 2, theoretical results of Theorems 1 and 2 are demonstrated, respectively. In Fig. 3, the output  $y_k(t)$  and the reference trajectory  $y_d(t)$  operated after 200 iterations are plotted, and it is clear that the output can follow the trajectory within a small bound even with iteration-varying uncertainties such that the disturbance resistance ability is verified.



**Fig. 2.** Bounded control input evaluated by  $\max_{1 \leq t \leq 50} \|u_k(t)\|_\infty$  for the first 200 iterations.



**Fig. 3.** Output after 200 iterations with the desired reference trajectory.

## 5 Conclusions

A disturbance observer-based robust ILC method has been proposed and analyzed in this paper. For general linear time-invariant systems, the disturbance observer-based ILC protocols can guarantee a bounded tracking result even with several nonrepetitive uncertainties. This result shows that this method can handle a more complex situation in real applications, and with adjustable observer parameters, the tracking error can keep in a stable level. Numerical simulations are provided to give a verification on developed theoretical results.

## References

1. He, W., Meng, T., He, X., Ge, S.S.: Unified iterative learning control for flexible structures with input constraints. *Automatica* **96**, 326–336 (2018)
2. Gao, F., Yang, Y., Shao, C.: Robust iterative learning control with applications to injection molding process. *Chem. Eng. Sci.* **56**(24), 7025–7034 (2001)
3. Maeda, G.J., Manchester, I.R., Rye, D.C.: Combined ILC and disturbance observer for the rejection of near-repetitive disturbances, with application to excavation. *IEEE Trans. Control Syst. Technol.* **23**(5), 1754–1769 (2015)
4. Hao, S., Liu, T., Rogers, E.: Extended state observer based indirect-type ILC for single-input single-output batch processes with time- and batch-varying uncertainties. *Automatica* **112**, 1–7 (2020)
5. Kim, K.-S., Rew, K.-H.: Reduced order disturbance observer for discrete-time linear systems. *Automatica* **49**(4), 968–975 (2013)
6. Meng, D., Zhang, J.: System equivalence transformation: Robust convergence of iterative learning control with nonrepetitive uncertainties, [arXiv:1910.10305](https://arxiv.org/abs/1910.10305) (2019)



# On Scaled Consensus, Bipartite Consensus and Scaled Bipartite Consensus: A Unified Viewpoint

Yuxin Wu<sup>1,2</sup> and Deyuan Meng<sup>1,2(✉)</sup>

<sup>1</sup> The Seventh Research Division, Beihang University (BUAA), Beijing 100191, People's Republic of China  
dymeng@buaa.edu.cn

<sup>2</sup> School of Automation Science and Electrical Engineering, Beihang University (BUAA), Beijing 100191, People's Republic of China

**Abstract.** This paper develops a unified framework to solve the scaled consensus, bipartite consensus and scaled bipartite consensus problems for multi-agent systems regardless of them being related to unsigned digraphs or signed digraphs. By linear nonsingular transformations, we turn the convergence analysis of multi-agent systems in the presence of three different control protocols into that under the classical consensus protocol such that necessary and sufficient conditions for the scaled consensus, bipartite consensus and scaled bipartite consensus are obtained, respectively, in a unified manner. Furthermore, we also validate the exponential convergence results.

**Keywords:** Bipartite consensus · Scaled bipartite consensus · Scaled consensus · Exponential convergence

## 1 Introduction

Multi-agent systems, which consist of several collaborative agents, have caught much attention in a lot of fields, ranging from computer science to social science (see, e.g., [1]). As one of the most primary problems for cooperative multi-agent systems, the classical consensus problem that reflects an agreement of agents has been widely studied. To achieve consensus, the Laplacian-type distributed control protocol is generally considered such that the behavior analysis of multi-agent systems can be easily carried out by leveraging the helpful properties of the Laplacian matrix (see, e.g., [2,3]). As a consequence, there have been reported a variety of analysis approaches and results for consensus. It is worth noticing that many other practical problems are closely related to the classical consensus problem, such as the formation control problem [4,5], the synchronization problem [6,7], and the optimization problem under cooperative control [8,9]. However,

---

This work was supported by the National Natural Science Foundation of China under Grant 61922007 and Grant 61873013.

in several real world scenarios like compartmental mass-action systems, water distribution systems and closed queueing networks, the desired control objective for cooperative multi-agent systems is to reach assigned proportions instead of a common value for agents, which renders more general scaled consensus problem, a new class of consensus problems, an attractive topic (see, e.g., [10, 11]).

When considering the multi-agent systems admitting both cooperative interactions and antagonistic interactions among agents, the bipartite consensus problem rather than the classical consensus problem, showing an agreement of agents on modulus, becomes an interesting consensus issue. By making use of the structural balance property, the bipartite consensus problem can easily be settled based on the classical consensus results (see, e.g., [12–15]). Similar to the cooperative multi-agent systems, of more interest may be the specific proportions of the state moduli instead of the common state moduli of agents. Therefore, the scaled bipartite consensus problem, as an extension of the scaled consensus problem to the cooperative-antagonistic multi-agent systems, should also be taken into account, which, as far as we know, has not been explored yet.

The aforementioned several classes of consensus problems are the fundamentally important problems for multi-agent systems. A question naturally arises: can these consensus problems be addressed under a unified framework? The answer is affirmative. In this paper, we deal with three different classes of consensus problems in a unified manner. Through linear nonsingular transformations, we directly turn the scaled consensus, bipartite consensus and scaled bipartite consensus problems for multi-agent systems related to three different control protocols into the same classical consensus problem, under which the convergence analysis of agents can be implemented by employing existing consensus results. The main contributions of this paper are twofold. First, necessary and sufficient conditions are obtained for these three classes of consensus behaviors. Second, we disclose the exponential convergence results of agents, which extend the consensus results in e.g., [10, 13] that only give asymptotic convergence results.

The remainder part of this paper is organized as follows. In Sect. 2, the preliminaries of digraphs are introduced. We provide our concerned problem and the consensus results in Sects. 3 and 4, respectively. Concluding remarks are given in Sect. 5.

*Notations:* Let  $\mathcal{I}_n = \{1, 2, \dots, n\}$ ,  $1_n = [1, 1, \dots, 1]^T \in \mathbb{R}^n$ ,  $\text{diag}\{d_1, d_2, \dots, d_n\} \in \mathbb{R}^{n \times n}$  be a diagonal matrix with diagonal entries as  $d_1, d_2, \dots, d_n$  and zero off-diagonal entries, and  $I$  be the identity matrix with compatible dimensions. For any scalar  $a \in \mathbb{R}$ ,  $|a|$  and  $\text{sign}(a)$  denote its absolute value and sign function, respectively. For any matrix  $A = [a_{ij}] \in \mathbb{R}^{n \times n}$ , if  $a_{ij} \geq 0, \forall i, j \in \mathcal{I}_n$ , then  $A$  is a non-negative matrix, denoted by  $A \geq 0$ .

## 2 Preliminaries of Digraphs

### 2.1 Unsigned Digraph

An *unsigned digraph*, denoted by a triple  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ , is usually adopted to represent a multi-agent system with only cooperations among agents, where

$\mathcal{V} = \{v_i : i \in \mathcal{I}_n\}$  is the node set,  $\mathcal{E} \subseteq \{(v_i, v_j) : v_i, v_j \in \mathcal{V}\}$  is the edge set and  $\mathcal{A} = [a_{ij}] \geq 0$  is the adjacency matrix.  $a_{ij} > 0 \Leftrightarrow (v_j, v_i) \in \mathcal{E}$  and  $a_{ij} = 0$ , otherwise. We assume that there exist no self-loops in  $\mathcal{G}$ , which means  $(v_i, v_i) \notin \mathcal{E}$  and  $a_{ii} = 0$ ,  $\forall i \in \mathcal{I}_n$ . By  $(v_j, v_i) \in \mathcal{E}$ , we call  $v_j$  a neighbor of  $v_i$  and then let the neighbor label set of  $v_i$  be denoted by  $\mathcal{N}_i = \{j : (v_j, v_i) \in \mathcal{E}\}$ . For some distinct nodes  $v_{i_0}, v_{i_1}, \dots, v_{i_l}$ , we call a sequence that consists of  $l$  edges in  $\mathcal{E}$  in the form of  $(v_{i_k}, v_{i_{k+1}}) \in \mathcal{E}$  for  $k = 0, 1, \dots, l - 1$  a path of  $\mathcal{G}$ . If there exists at least one node which can be connected to all the other nodes through paths, then we say that  $\mathcal{G}$  is quasi-strongly connected or has a spanning tree. For the clarity of notations, we employ  $\mathcal{G}(\mathcal{A})$  for the unsigned digraph specified by  $\mathcal{A}$ .

## 2.2 Signed Digraph

A multi-agent system involving both cooperative interactions and antagonistic interactions among agents is generally in the presence of a *signed digraph*. A signed digraph is also denoted by a triple  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ , where  $\mathcal{V}$  and  $\mathcal{E}$  are the same as those of  $\mathcal{G}(\mathcal{A})$  but  $\mathbf{A} = [\mathbf{a}_{ij}] \in \mathbb{R}^{n \times n}$  differs from  $\mathcal{A}$  and is no longer a nonnegative matrix. We denote  $\mathcal{G}(\mathbf{A})$  for the signed digraph corresponding to  $\mathbf{A}$ . For  $\mathcal{G}(\mathbf{A})$ , we have  $\mathbf{a}_{ij} \neq 0 \Leftrightarrow (v_j, v_i) \in \mathcal{E}$  and  $\mathbf{a}_{ij} = 0$ , otherwise. We also assume that  $\mathcal{G}(\mathbf{A})$  has no self-loops, i.e.,  $(v_i, v_i) \notin \mathcal{E}$  and  $\mathbf{a}_{ii} = 0$ ,  $\forall i \in \mathcal{I}_n$ . The other concepts for  $\mathcal{G}(\mathbf{A})$ , such as the neighbor, path and connectivity, which are independent from  $\mathbf{A}$ , can be defined in the same way as those of  $\mathcal{G}(\mathcal{A})$ . As a distinct concept of the signed digraph  $\mathcal{G}(\mathbf{A})$ ,  $\mathcal{G}(\mathbf{A})$  is said to be structurally balanced if there exists a node bipartition  $\{\mathcal{V}_1, \mathcal{V}_2\}$ , where  $\mathcal{V}_1 \cup \mathcal{V}_2 = \mathcal{V}$  and  $\mathcal{V}_1 \cap \mathcal{V}_2 = \emptyset$ , such that  $\mathbf{a}_{ij} \geq 0$  for  $\forall v_i, v_j \in \mathcal{V}_q$  ( $q \in \{1, 2\}$ ) and  $\mathbf{a}_{ij} \leq 0$  for  $\forall v_i \in \mathcal{V}_l, v_j \in \mathcal{V}_q$ ,  $l \neq q$  ( $l, q \in \{1, 2\}$ ). For a structurally balanced signed digraph  $\mathcal{G}(\mathbf{A})$ , there exists some gauge transformation matrix  $D = \text{diag}\{\beta_1, \beta_2, \dots, \beta_n\}$  rendering  $D\mathbf{A}D \geq 0$ , where  $\beta_i \in \{1, -1\}$ ,  $\forall i \in \mathcal{I}_n$ .

## 3 Problem Description

We are interested in multi-agent systems with  $n$  agents under the common state space  $\mathbb{R}$ . Assume that the  $i$ th agent is of the following dynamics:

$$\dot{x}_i(t) = u_i(t), \quad \forall i \in \mathcal{I}_n \quad (1)$$

where  $x_i(t)$  and  $u_i(t)$  are the state and the control protocol of the  $i$ th agent. We aim to design control protocols for the system (1) under unsigned digraphs or signed digraphs to reach some important consensus behaviors, such as scaled consensus, bipartite consensus and scaled bipartite consensus, which will be introduced in the following sections, and exploit a new framework to observe these consensus behaviors in a unified manner.

### 3.1 Scaled Consensus

Under the unsigned digraph  $\mathcal{G}(\mathcal{A})$ , the multi-agent system (1) is said to achieve *scaled consensus* to  $(\alpha_1, \alpha_2, \dots, \alpha_n)$  if for almost all initial state  $x_i(0) \in \mathbb{R}$ , there exists some scalar  $c \neq 0$  such that

$$\lim_{t \rightarrow \infty} \alpha_i x_i(t) = c, \quad \forall i \in \mathcal{I}_n \quad (2)$$

where  $\alpha_1, \alpha_2, \dots, \alpha_n$  are some nonzero scalars. Obviously, the objective (2) can guarantee the scaled consensus considered in [10], i.e.,

$$\lim_{t \rightarrow \infty} [\alpha_1 x_1(t) - \alpha_j x_j(t)] = 0, \quad \forall j = 2, 3, \dots, n.$$

### 3.2 Bipartite Consensus

Under the signed digraph  $\mathcal{G}(\mathbf{A})$ , the multi-agent system (1) is said to achieve *bipartite consensus* if for almost all initial state  $x_i(0) \in \mathbb{R}$ , there exists some scalar  $c \neq 0$  such that

$$\lim_{t \rightarrow \infty} \beta_i x_i(t) = c, \quad \forall i \in \mathcal{I}_n \quad (3)$$

where  $\beta_1, \beta_2, \dots, \beta_n$  are some nonzero scalars taking values from  $\{1, -1\}$ . We can easily validate that the objective (3) guarantees

$$\lim_{t \rightarrow \infty} |x_i(t)| = |c| > 0, \quad \forall i \in \mathcal{I}_n$$

which yields the bipartite consensus problem settled in [12].

Notice that bipartite consensus needs the structural balance property of the signed digraph  $\mathcal{G}(\mathbf{A})$ . If there are no special indications, then we always assume that  $\mathcal{G}(\mathbf{A})$  is structurally balanced. In the following discussions, we denote  $\beta_1, \beta_2, \dots, \beta_n$  as those scalars satisfying  $D\mathbf{A}D \geq 0$  with  $D = \text{diag}\{\beta_1, \beta_2, \dots, \beta_n\}$ . It is worth mentioning that the parameters  $\beta_1, \beta_2, \dots, \beta_n$  are heavily dependent on the topology structure of  $\mathcal{G}(\mathbf{A})$  (see [12] for more detailed discussions).

**Remark 1.** *From the perspective of consensus objectives, scaled consensus views bipartite consensus as a special case, which only restricts  $\alpha_i, \forall i \in \mathcal{I}_n$  to take values from a two-point set  $\{1, -1\}$ . However, scaled consensus deals with multi-agent systems under an unsigned digraph  $\mathcal{G}(\mathcal{A})$ , while bipartite consensus appears in those related to a signed digraph  $\mathcal{G}(\mathbf{A})$ . In fact, the unsigned digraph  $\mathcal{G}(\mathcal{A})$  can be regarded as a special case of the signed digraph  $\mathcal{G}(\mathbf{A})$  when the associated edge weights are all positive (see also [12]). Another difference between scaled consensus and bipartite consensus can be seen from the fact that  $\alpha_i, \forall i \in \mathcal{I}_n$  can be freely selected on any nonzero parameters independent of the topology of  $\mathcal{G}(\mathcal{A})$ , whereas  $\beta_i, \forall i \in \mathcal{I}_n$  heavily depends on the topology of  $\mathcal{G}(\mathbf{A})$ .*

### 3.3 Scaled Bipartite Consensus

Motivated by the discussions in Remark 1, we are also interested in the extended behavior of scaled consensus to the multi-agent systems under signed digraphs, for which we call it scaled bipartite consensus. That is, in the presence of the signed digraph  $\mathcal{G}(\mathbf{A})$ , the multi-agent system (1) is said to achieve *scaled bipartite consensus* to  $(\alpha_1, \alpha_2, \dots, \alpha_n)$  if for almost all initial state  $x_i(0) \in \mathbb{R}$ , there exists some scalar  $c \neq 0$  such that

$$\lim_{t \rightarrow \infty} \alpha_i \beta_i x_i(t) = c, \quad \forall i \in \mathcal{I}_n. \quad (4)$$

By the consensus objective (4), we can guarantee

$$\lim_{t \rightarrow \infty} \left| \frac{x_i(t)}{x_j(t)} \right| = \left| \frac{\alpha_j}{\alpha_i} \right|, \quad \forall i, j \in \mathcal{I}_n.$$

Comparing (4) with (2) and (3), we can see that scaled bipartite consensus obviously contains scaled consensus and bipartite consensus as special cases.

## 4 Unified Consensus Analysis

### 4.1 Consensus Results

To achieve the scaled consensus objective (2) for the multi-agent system (1) in the presence of the unsigned digraph  $\mathcal{G}(\mathcal{A})$ , we propose a distributed protocol that is conducted by the nearest neighbor rule as

$$u_i(t) = \sum_{j \in \mathcal{N}_i} \frac{a_{ij}}{\alpha_i} [\alpha_j x_j(t) - \alpha_i x_i(t)], \quad \forall i \in \mathcal{I}_n. \quad (5)$$

The following theorem states that this protocol ensures scaled consensus to be achieved exponentially fast for the multi-agent system (1) under the quasi-strong connectivity.

**Theorem 1.** *Consider the multi-agent system (1) related to the unsigned digraph  $\mathcal{G}(\mathcal{A})$ . Then it achieves the scaled consensus objective (2) exponentially fast under the control protocol (5) if and only if the unsigned digraph  $\mathcal{G}(\mathcal{A})$  is quasi-strongly connected.*

**Remark 2.** *There has been reported a protocol in [10] for scaled consensus as*

$$u_i(t) = \text{sign}(\alpha_i) \sum_{j \in \mathcal{N}_i} a_{ij} [\alpha_j x_j(t) - \alpha_i x_i(t)], \quad \forall i \in \mathcal{I}_n. \quad (6)$$

*Let  $\bar{\mathcal{A}} = [\bar{a}_{ij}] \in \mathbb{R}^{n \times n}$  with  $\bar{a}_{ij} = |\alpha_i| a_{ij}$ . Due to  $\alpha_i \neq 0$ ,  $\forall i \in \mathcal{I}_n$ , we can rewrite (6) as*

$$u_i(t) = \sum_{j \in \mathcal{N}_i} \frac{|\alpha_i| a_{ij}}{\alpha_i} [\alpha_j x_j(t) - \alpha_i x_i(t)] = \sum_{j \in \mathcal{N}_i} \frac{\bar{a}_{ij}}{\alpha_i} [\alpha_j x_j(t) - \alpha_i x_i(t)], \quad \forall i \in \mathcal{I}_n$$

which falls into the framework of (5) by replacing  $\mathcal{A}$  with  $\bar{\mathcal{A}}$ . Since  $\mathcal{G}(\bar{\mathcal{A}})$  is an unsigned digraph, it is obviously a subgraph of  $\mathcal{G}(\mathcal{A})$  and vice versa. This implies that the consensus analysis of (6) can be covered by following the same processes as those of (5). It is worth mentioning that in Theorem 1, we provide scaled consensus with a necessary and sufficient guarantee. Furthermore, the exponential convergence result of scaled consensus can be confirmed in the presence of the quasi-strongly connected unsigned digraph  $\mathcal{G}(\mathcal{A})$ . By contrast, it concludes only a sufficient condition for the asymptotic convergence of scaled consensus in [10] and requires the strong connectivity property of the associated unsigned digraph  $\mathcal{G}(\mathcal{A})$ . Hence, the scaled consensus result of Theorem 1 can be viewed as a significant extension of that in [10].

Regarding the bipartite consensus objective (3), we design a distributed protocol for the multi-agent system (1) under the signed digraph  $\mathcal{G}(\mathbf{A})$  as

$$u_i(t) = \sum_{j \in \mathcal{N}_i} \beta_j \mathbf{a}_{ij} [\beta_j x_j(t) - \beta_i x_i(t)], \quad \forall i \in \mathcal{I}_n \quad (7)$$

which ensures its exponential convergence under the quasi-strong connectivity property.

**Theorem 2.** Consider the multi-agent system (1) related to the structurally balanced signed digraph  $\mathcal{G}(\mathbf{A})$ . Then it achieves the bipartite consensus objective (3) exponentially fast under the control protocol (7) if and only if the signed digraph  $\mathcal{G}(\mathbf{A})$  is quasi-strongly connected.

**Remark 3.** In [13], the following protocol for bipartite consensus is studied:

$$u_i(t) = - \sum_{j \in \mathcal{N}_i} |\mathbf{a}_{ij}| [x_i(t) - \text{sign}(\mathbf{a}_{ij}) x_j(t)], \quad \forall i \in \mathcal{I}_n. \quad (8)$$

We can verify that  $\beta_i \beta_j \mathbf{a}_{ij} = |\mathbf{a}_{ij}| \geq 0$  owing to  $DAD = [\beta_i \beta_j \mathbf{a}_{ij}] \geq 0$  and further that  $\text{sign}(\mathbf{a}_{ij}) = \beta_i \beta_j$  for  $\mathbf{a}_{ij} \neq 0$ . Thus, we insert  $\beta_i \in \{1, -1\}$ ,  $\forall i \in \mathcal{I}_n$  to transform (8) into

$$u_i(t) = - \sum_{j \in \mathcal{N}_i} \beta_i \beta_j \mathbf{a}_{ij} [x_i(t) - \beta_i \beta_j x_j(t)] = \sum_{j \in \mathcal{N}_i} \beta_j \mathbf{a}_{ij} [\beta_j x_j(t) - \beta_i x_i(t)], \quad \forall i \in \mathcal{I}_n$$

which is equivalent to (7). Theorem 2 establishes a necessary and sufficient exponential convergence condition of bipartite consensus when the multi-agent system is associated with a structurally balanced signed digraph. Obviously, the bipartite consensus result of Theorem 2 can extend that of [13] by refining the stronger exponential convergence result under the quasi-strong connectivity condition.

Next, we consider the scaled bipartite consensus objective (4) under the signed digraph  $\mathcal{G}(\mathbf{A})$  and present the following distributed protocol:

$$u_i(t) = \sum_{j \in \mathcal{N}_i} \frac{\beta_j \mathbf{a}_{ij}}{\alpha_i} [\alpha_j \beta_j x_j(t) - \alpha_i \beta_i x_i(t)], \quad \forall i \in \mathcal{I}_n. \quad (9)$$

Then, we develop the following theorem with a necessary and sufficient result for exponential convergence of agents relevant to the quasi-strongly connected signed digraph.

**Theorem 3.** *Consider the multi-agent system (1) related to the structurally balanced signed digraph  $\mathcal{G}(\mathbf{A})$ . Then it achieves the scaled bipartite consensus objective (4) exponentially fast under the control protocol (9) if and only if the signed digraph  $\mathcal{G}(\mathbf{A})$  is quasi-strongly connected.*

**Remark 4.** *When  $\beta_i = 1$ ,  $\forall i \in \mathcal{I}_n$  holds,  $\mathbf{a}_{ij} \geq 0$ ,  $\forall i, j \in \mathcal{I}_n$  follows immediately. In this case,  $\mathcal{G}(\mathbf{A})$  is corresponding to a nonnegative adjacency matrix  $\mathbf{A} \geq 0$ . This fact obviously yields that the protocol (9) includes the one (5) as a special case where  $\beta_i = 1$ ,  $\forall i \in \mathcal{I}_n$ . When we specially take  $\alpha_i = 1$ ,  $\forall i \in \mathcal{I}_n$ , (9) can be simplified into (7). That is, the protocol (7) is a special case of the one (9). Theorem 3 thus further generalizes the consensus results of Theorems 1 and 2, and consequently those of [10] and [13] under quasi-strongly connected digraphs.*

## 4.2 Consensus Analysis

By noting the discussions of Remark 4, we can develop a unified analysis approach for the considered three consensus problems, for which we only need to give the proof for the scaled bipartite consensus result of Theorem 3.

*Proof of Theorem 3:* From the objective (4), we can easily observe that this proof can be implemented by considering the consensus performance with respect to a modified state  $\alpha_i \beta_i x_i(t)$  of each agent  $v_i$ . Hence, we perform a linear nonsingular transformation of the multi-agent system (1) by taking  $\tilde{x}_i(t) = \alpha_i \beta_i x_i(t)$  due to  $\alpha_i \neq 0$  and  $\beta_i \neq 0$ , and transform the dynamics of the multi-agent system (1) under the protocol (9) into

$$\begin{aligned}\dot{\tilde{x}}_i(t) &= \alpha_i \beta_i \dot{x}_i(t) \\ &= \alpha_i \beta_i \left\{ \sum_{j \in \mathcal{N}_i} \frac{\beta_j \mathbf{a}_{ij}}{\alpha_i} [\alpha_j \beta_j x_j(t) - \alpha_i \beta_i x_i(t)] \right\} \\ &= \sum_{j \in \mathcal{N}_i} \beta_i \beta_j \mathbf{a}_{ij} [\tilde{x}_j(t) - \tilde{x}_i(t)], \quad \forall i \in \mathcal{I}_n.\end{aligned}\tag{10}$$

Denote  $\tilde{x}(t) = [\tilde{x}_1(t), \tilde{x}_2(t), \dots, \tilde{x}_n(t)]^T$  and  $\tilde{\mathcal{A}} = [\tilde{a}_{ij}] \in \mathbb{R}^{n \times n}$  with  $\tilde{a}_{ij} = \beta_i \beta_j \mathbf{a}_{ij}$  for the system (10). It follows  $\tilde{\mathcal{A}} \geq 0$  thanks to  $\tilde{a}_{ij} = |\mathbf{a}_{ij}| \geq 0$  (see Remark 3). Let  $\tilde{\mathcal{L}} = [\tilde{l}_{ij}] \in \mathbb{R}^{n \times n}$  be the Laplacian matrix related to  $\tilde{\mathcal{A}}$ , which satisfies  $\tilde{l}_{ij} = \sum_{k \in \mathcal{N}_i} \tilde{a}_{ik}$  if  $j = i$ ; and  $\tilde{l}_{ij} = -\tilde{a}_{ij}$ , otherwise (see, e.g., [3]). Then, we can rewrite (10) in a vector form as

$$\dot{\tilde{x}}(t) = -\tilde{\mathcal{L}}\tilde{x}(t).\tag{11}$$

A classical consensus result for the system (11) is that for any initial conditions  $\tilde{x}(0)$ ,  $\lim_{t \rightarrow \infty} \tilde{x}(t) = 1_n c$  if and only if the unsigned digraph  $\mathcal{G}(\tilde{\mathcal{A}})$  is quasi-strongly connected, where  $c = \eta^T \tilde{x}(0)$  and  $\eta \geq 0$  is a nonzero constant vector such that  $\eta^T \tilde{L} = 0$  and  $1_n^T \eta = 1$  (see, e.g., [2]). Note that the quasi-strong connectivity of  $\mathcal{G}(\tilde{\mathcal{A}})$  is the same as that of  $\mathcal{G}(\mathbf{A})$  and  $\lim_{t \rightarrow \infty} \tilde{x}(t) = 1_n c$  is equivalent to the scaled bipartite consensus objective (4). We can conclude (4) holds if and only if  $\mathcal{G}(\mathbf{A})$  is quasi-strongly connected.

We are left to prove the exponential convergence of the multi-agent system (1). Denote  $P_1 = [0 \ I]^T \in \mathbb{R}^{n \times (n-1)}$ ,  $Q_1 = [-1_{n-1} \ I] \in \mathbb{R}^{(n-1) \times n}$ , and  $R_1 = [1 \ 0 \ \cdots \ 0] \in \mathbb{R}^{1 \times n}$ . If we take  $\tilde{P} = [1_n \ P_1] \in \mathbb{R}^{n \times n}$ , then  $\tilde{P}$  is nonsingular and  $\tilde{P}^{-1} = [R_1^T \ Q_1^T]^T$ . We perform a nonsingular transformation  $\tilde{P}^{-1} \tilde{x}(t) = [\tilde{x}_1(t) \ z^T(t)]^T$  for the system (11) to obtain

$$\begin{bmatrix} \dot{\tilde{x}}_1(t) \\ \dot{z}(t) \end{bmatrix} = -\tilde{P}^{-1} \tilde{\mathcal{L}} \tilde{P} \begin{bmatrix} \tilde{x}_1(t) \\ z(t) \end{bmatrix} = -\begin{bmatrix} R_1 \tilde{\mathcal{L}} 1_n & R_1 \tilde{\mathcal{L}} P_1 \\ Q_1 \tilde{\mathcal{L}} 1_n & Q_1 \tilde{\mathcal{L}} P_1 \end{bmatrix} \begin{bmatrix} \tilde{x}_1(t) \\ z(t) \end{bmatrix} = \begin{bmatrix} 0 & -R_1 \tilde{\mathcal{L}} P_1 \\ 0 & -Q_1 \tilde{\mathcal{L}} P_1 \end{bmatrix} \begin{bmatrix} \tilde{x}_1(t) \\ z(t) \end{bmatrix} \quad (12)$$

where we also use  $\tilde{\mathcal{L}} 1_n = 0$ . Since  $\mathcal{G}(\tilde{\mathcal{A}})$  is quasi-strongly connected,  $\tilde{\mathcal{L}}$  has exactly one zero eigenvalue and the other eigenvalues all have positive real parts (see, e.g., [2]). This together with (12) implies that  $-Q_1 \tilde{\mathcal{L}} P_1$  is a Hurwitz stable matrix, i.e., all of its eigenvalues have negative real parts. Hence, (12) contains a Hurwitz stable linear system  $\dot{z}(t) = -Q_1 \tilde{\mathcal{L}} P_1 z(t)$ , the solution of which is given as

$$z(t) = e^{-Q_1 \tilde{\mathcal{L}} P_1 t} z(0) = e^{-Q_1 \tilde{\mathcal{L}} P_1 t} Q_1 \tilde{x}(0). \quad (13)$$

According to (12), we also have another linear system with respect to  $\tilde{x}_1(t)$  under dynamics as  $\dot{\tilde{x}}_1(t) = -R_1 \tilde{\mathcal{L}} P_1 z(t)$ . For this system, we can insert (13) to arrive at

$$\begin{aligned} \tilde{x}_1(t) &= \tilde{x}_1(0) + \int_0^t (-R_1 \tilde{\mathcal{L}} P_1) z(\tau) d\tau \\ &= R_1 \tilde{x}(0) + R_1 \tilde{\mathcal{L}} P_1 \left( Q_1 \tilde{\mathcal{L}} P_1 \right)^{-1} \int_0^t de^{-Q_1 \tilde{\mathcal{L}} P_1 \tau} Q_1 \tilde{x}(0) \\ &= \left[ R_1 - R_1 \tilde{\mathcal{L}} P_1 \left( Q_1 \tilde{\mathcal{L}} P_1 \right)^{-1} Q_1 \right] \tilde{x}(0) \\ &\quad + R_1 \tilde{\mathcal{L}} P_1 \left( Q_1 \tilde{\mathcal{L}} P_1 \right)^{-1} e^{-Q_1 \tilde{\mathcal{L}} P_1 t} Q_1 \tilde{x}(0). \end{aligned} \quad (14)$$

Let  $\tilde{\eta}^T = R_1 - R_1 \tilde{\mathcal{L}} P_1 \left( Q_1 \tilde{\mathcal{L}} P_1 \right)^{-1} Q_1$ . With the properties that  $1_n R_1 + P_1 Q_1 = \tilde{P} \tilde{P}^{-1} = I$ ,  $Q_1 1_n = 0$ ,  $\tilde{\mathcal{L}} 1_n = 0$  and  $R_1 1_n = 1$ , we can easily validate that

$$\begin{aligned} \tilde{\eta}^T \tilde{\mathcal{L}} &= R_1 \tilde{\mathcal{L}} - R_1 \tilde{\mathcal{L}} P_1 \left( Q_1 \tilde{\mathcal{L}} P_1 \right)^{-1} Q_1 \tilde{\mathcal{L}} \\ &= R_1 \tilde{\mathcal{L}} - R_1 \tilde{\mathcal{L}} P_1 \left( Q_1 \tilde{\mathcal{L}} P_1 \right)^{-1} Q_1 \tilde{\mathcal{L}} (1_n R_1 + P_1 Q_1) \\ &= R_1 \tilde{\mathcal{L}} - R_1 \tilde{\mathcal{L}} (I - 1_n R_1) \\ &= 0 \end{aligned}$$

and that

$$1_n^T \tilde{\eta} = \tilde{\eta}^T 1_n = R_1 1_n - R_1 \tilde{\mathcal{L}} P_1 \left( Q_1 \tilde{\mathcal{L}} P_1 \right)^{-1} Q_1 1_n = 1.$$

Since  $\tilde{\mathcal{L}}$  has only one zero eigenvalue, it has a unique left eigenvector that satisfies both  $\eta^T \tilde{\mathcal{L}} = 0$  and  $1_n^T \eta = 1$ . Thus,  $\tilde{\eta} = \eta$  holds. By noticing  $c = \eta^T \tilde{x}(0)$  and  $\tilde{x}(t) = \tilde{P}[\tilde{x}_1(t) \ z^T(t)]^T$ , we can deduce that

$$\tilde{x}(t) - 1_n c = 1_n (\tilde{x}_1(t) - c) + P_1 z(t) = \left( 1_n R_1 \tilde{\mathcal{L}} P_1 (Q_1 \tilde{\mathcal{L}} P_1)^{-1} + P_1 \right) e^{-Q_1 \tilde{\mathcal{L}} P_1 t} Q_1 \tilde{x}(0). \quad (15)$$

Owing to the Hurwitz stability of  $-Q_1 \tilde{\mathcal{L}} P_1$ , it follows from (15) that  $\lim_{t \rightarrow \infty} \tilde{x}(t) = 1_n c$  exponentially fast, which directly leads to the exponential convergence result of  $\lim_{t \rightarrow \infty} \alpha_i \beta_i x_i(t) = c$ ,  $\forall i \in \mathcal{I}_n$ . The proof of Theorem 3 is completed. ■

**Remark 5.** *The proof of Theorem 3 obviously yields that we can implement the convergence analysis of the newly observed scaled consensus [10], bipartite consensus [12] and scaled bipartite consensus problems in a unified framework based on the classical consensus results (see, e.g., [2]). It benefits from the typical linear nonsingular transformation of linear systems, which can be incorporated to further exploit a new analysis approach for exponential convergence of multi-agent systems. It is clear that the convergence speed of the multi-agent system (1) heavily depends on the nonzero eigenvalues of its Laplacian matrix, especially the minimum nonzero eigenvalue, despite it being related to unsigned digraphs or signed digraphs.*

**Remark 6.** *In addition to the above discussions, we again consider the consensus protocols. We can leverage  $\beta_i \in \{1, -1\}$  and  $\beta_i \beta_j \mathbf{a}_{ij} = \tilde{a}_{ij}$  to rewrite the protocol (7) as*

$$u_i(t) = \sum_{j \in \mathcal{N}_i} \frac{\beta_i \beta_j \mathbf{a}_{ij}}{\beta_i} [\beta_j x_j(t) - \beta_i x_i(t)] = \sum_{j \in \mathcal{N}_i} \frac{\tilde{a}_{ij}}{\beta_i} [\beta_j x_j(t) - \beta_i x_i(t)], \quad \forall i \in \mathcal{I}_n.$$

For the same reason, the protocol (9) can be rewritten as

$$u_i(t) = \sum_{j \in \mathcal{N}_i} \frac{\beta_i \beta_j \mathbf{a}_{ij}}{\alpha_i \beta_i} [\alpha_j \beta_j x_j(t) - \alpha_i \beta_i x_i(t)] = \sum_{j \in \mathcal{N}_i} \frac{\tilde{a}_{ij}}{\alpha_i \beta_i} [\alpha_j \beta_j x_j(t) - \alpha_i \beta_i x_i(t)], \quad \forall i \in \mathcal{I}_n.$$

Via comparing them with (5), we can find that the protocols (5), (7) and (9) essentially share a unified scaled consensus protocol form related to unsigned digraphs with nonnegative adjacency weights. This, as well as Theorems 1, 2 and 3, guarantees that not only the scaled consensus, bipartite consensus, and scaled bipartite consensus results but also the protocols designed for them can be treated in a unified viewpoint.

**Remark 7.** When considering the multi-agent system (1) under signed digraphs, we only analyze the structural balance case. For the counterpart case related to structurally unbalanced signed digraphs, we can easily obtain that  $\tilde{\mathcal{L}}$  in (11) is still a Laplacian matrix with respect to a structurally unbalanced signed digraph  $\mathcal{G}(\tilde{\mathcal{A}})$ . By following the results developed in [13], the dynamic behaviors of the system (1) can be determined. Specifically, the scaled interval bipartite consensus or stability can be achieved for the multi-agent system (1) in the presence of structurally unbalanced signed digraphs under the protocol (9), which heavily depends on the structural balance properties of root subgraphs (see [13] for more details).

## 5 Conclusions

In this paper, we have discussed three classes of consensus problems, namely, the scaled consensus, bipartite consensus, and scaled bipartite consensus problems, for multi-agent systems in the presence of unsigned digraphs or signed digraphs under a unified viewpoint. By employing some linear nonsingular transformations, these three classes of consensus problems fall into the same classical consensus problem. Based on the well developed classical consensus results, we have obtained the necessary and sufficient conditions for scaled consensus, bipartite consensus, and scaled bipartite consensus, respectively, in a unified manner. In addition, the exponential convergence results for these consensus behaviors have also been validated.

## References

1. Cao, Y., Yu, W., Ren, W., Chen, G.: An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Trans. Industr. Inf.* **9**(1), 427–438 (2013)
2. Ren, W., Beard, R.W., McLain, T.W.: Coordination variables and consensus building in multiple vehicle systems. In: Cooperative Control, pp. 171–188. Springer, Berlin (2004)
3. Ren, W., Beard, R.W., Atkins, E.M.: Information consensus in multivehicle cooperative control. *IEEE Control Syst. Mag.* **27**(2), 71–82 (2007)
4. Xiao, F., Wang, L., Chen, J., Gao, Y.: Finite-time formation control for multi-agent systems. *Automatica* **45**(11), 2605–2611 (2009)
5. Oh, K.-K., Ahn, H.-S.: Formation control of mobile agents based on inter-agent distance dynamics. *Automatica* **47**(10), 2306–2312 (2011)
6. Yu, W., Chen, G., Lü, J.: On pinning synchronization of complex dynamical networks. *Automatica* **45**(2), 429–435 (2009)
7. Liu, S., Xie, L., Lewis, F.L.: Synchronization of multi-agent systems with delayed control input information from neighbors. *Automatica* **47**(10), 2152–2164 (2011)
8. Zargham, M., Ribeiro, A., Ozdaglar, A.E., Jadbabaie, A.: Accelerated dual descent for network flow optimization. *IEEE Trans. Autom. Control* **59**(4), 905–920 (2014)
9. Olshevsky, A.: Linear time average consensus and distributed optimization on fixed graphs. *SIAM J. Control Optim.* **55**(6), 3990–4014 (2017)

10. Roy, S.: Scaled consensus. *Automatica* **51**, 259–262 (2015)
11. Meng, D., Jia, Y.: Scaled consensus problems on switching networks. *IEEE Trans. Autom. Control* **61**(6), 1664–1669 (2016)
12. Altafini, C.: Consensus problems on networks with antagonistic interactions. *IEEE Trans. Autom. Control* **58**(4), 935–946 (2013)
13. Meng, D., Du, M., Jia, Y.: Interval bipartite consensus of networked agents associated with signed digraphs. *IEEE Trans. Autom. Control* **61**(12), 3755–3770 (2016)
14. Meng, D., Du, M., Wu, Y.: Extended structural balance theory and method for cooperative-antagonistic networks. *IEEE Trans. Autom. Control* **65**(5), 2147–2154 (2020)
15. Meng, D., Liang, J., Wu, Y., Meng, Z.: Connection of signed and unsigned networks based on solving linear dynamic systems. *IEEE Trans. Syst. Man Cybern. Syst.* (to appear). <https://doi.org/10.1109/TSMC.2019.2945538>



# Voltage Balancing of Modular Multilevel Converter Based on Cerebellar Model Articulation Controller

Xiangsheng Liu<sup>1</sup>(✉), Yuanyuan Yang<sup>1</sup>, Lin Ren<sup>1</sup>, Zhengxin Zhou<sup>1</sup>, Yunxia Jiang<sup>1</sup>, Lailong Song<sup>2</sup>, and ZhengLin Jiang<sup>1</sup>

<sup>1</sup> Sanda University, Shanghai, China

[lxssit@163.com](mailto:lxssit@163.com)

<sup>2</sup> Shandong Jinzhong Science and Technology Group Limited Company, Jinan, China

**Abstract.** This paper proposes a closed loop control strategy based on cerebellar model articulation controller the MMC (Modular Multilevel Converter) sub-modules of capacitor voltage balance. MMC circuit based on the multiple child modules cascade system, each module of MMC (SM) is set independent equalizing capacitor voltage circuit. By the intelligent control mode of cerebellar model to control the MMC of cutting module, so as to achieve MMC capacitor voltage balance. The simulation results show that the proposed voltage sharing control strategy based on CMAC effectively achieves the capacitor voltage balance, and improves the stability and anti-interference ability of the system.

**Keywords:** Modular multilevel converter · Voltage sharing control · Cerebellar model articulation controller · Carrier phase shifting modulation · Voltage balance

## 1 Introduction

The modular multilevel converter has the advantages of low operating loss and good output waveform quality, and it solves the problem of dynamic voltage equalization of general converters. It is widely used in the fields of high-voltage direct current transmission and high-voltage motor drive.

The balance of the MMC capacitor voltage is the key to the stable operation of the MMC. The MMC voltage equalization control includes bridge arm voltage equalization control and sub-module voltage equalization control, of which the sub-module voltage equalization control is a hotspot at home and abroad. At present, there are mainly two algorithms for controlling the voltage of the submodule capacitance. One is based on the PI controller's submodule voltage equalization control strategy. This algorithm is simple and has a good voltage equalization effect. However, in order to reduce the harmonic component of the output current, the number of MMC submodules is usually large. It is easy to cause the control circuit to be too complicated [1]. The other is a voltage

equalization algorithm for sequencing control based on the voltage of one-phase sub-module. When there are many sub-modules, the sequencing operation of the voltage is easy to cause unnecessary actions of the switching device [2]. In view of this situation, literature [3] improved the traditional pressure equalization method and proposed a control strategy for layered pressure equalization.

During the operation of the MMC control system, the parameter setting of the traditional PI controller is difficult, and the dynamic performance is poor due to the failure to adjust the parameters in real time when a fault occurs. Using intelligent controller can better overcome the problem of real-time parameter adjustment, improve the response speed of the system, and reduce the harmonics of the system.

This paper uses a closed-loop control strategy based on the cerebellar model neural network to control the balance of MMC capacitor voltage.

## 2 Model Analysis Principle of Modular Multilevel Converter

Figure 1 is the equivalent circuit schematic diagram of MMC. MMC is composed of three phases and six bridge arms, each bridge arm SM equivalent module is formed by cascading N sub-modules. Two reactors are connected in series between the upper and lower bridge arms, and the output resistive load is connected in a star connection. The neutral point of the output load is connected with the neutral point of the two power sources to keep the potential of the neutral point of the load consistent with the neutral point of the power source and zero potential. In Fig. 1, the three phases A, B, and C are symmetrically analyzed with phase A (a). The correlation of the upper arm is denoted by p, and the correlation of the lower arm is denoted by n. The mathematical model of MMC can be obtained:

$$\begin{cases} u_{ap} = \frac{U_{dc}}{2} - u_a - u_{sp} \\ u_{an} = \frac{U_{dc}}{2} + u_a - u_{sn} \end{cases} \quad (1)$$

$u_{ap}$  and  $u_{an}$  are the voltages of the upper and lower bridge arms,  $U_{dc}$  is the DC side voltage,  $u_{sp}$  and  $u_{sn}$  are the upper and lower bridge arm inductance voltages,  $u_a$  is the A-phase output voltage, and  $i_{ap}$  is the upper bridge arm current,  $i_{an}$  is the current of the lower arm and  $i_{cir.a}$  is the current of the arm.

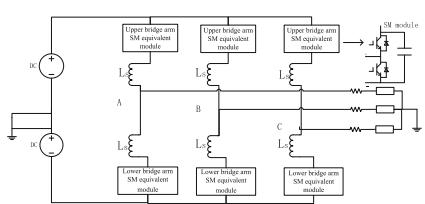
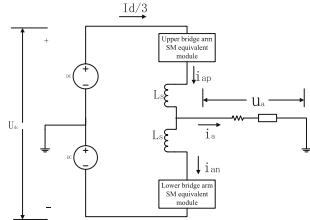
Figure 2 shows the bridge arm current in the inverter state, and the current of the upper and lower bridge arms is known from Kirchhoff's basic theorem:

$$i_a = i_{ap} - i_{an} \quad (2)$$

$$i_{cir.a} = \frac{i_{ap} + i_{an}}{2} \quad (3)$$

Synchronous (2), (3) formula can get the bridge arm current:

$$\begin{cases} i_{ap} = \frac{i_a}{2} + i_{cir.a} \\ i_{an} = -\frac{i_a}{2} + i_{cir.a} \end{cases} \quad (4)$$

**Fig. 1.** MRAS schematic structure**Fig. 2.** Single-phase inverter state equivalent circuit

Through analysis, the sum of the currents of the upper and lower bridge arms is the common mode current, So half the sum of the bridge arm currents is the circulating current of the phase, That is, part of the current flowing into the lower arm has the same amplitude and phase, and only flows between the phases, and does not pass through the AC side.

Synonyms (1), (2), (3) formula can be obtained:

$$\begin{cases} u_{ap} = \frac{U_{dc}}{2} - u_a - \frac{L_s}{2} \frac{di_a}{dt} - \frac{L_s}{2} \frac{di_{cir.a}}{dt} \\ u_{an} = \frac{U_{dc}}{2} + u_a + \frac{L_s}{2} \frac{di_a}{dt} - \frac{L_s}{2} \frac{di_{cir.a}}{dt} \end{cases} \quad (5)$$

Without considering the influence of circulating current fluctuations, the voltages of the upper and lower bridge arms are just the DC side voltage, and the upper and lower bridge arms each bear half. This is the first part of the voltage. The second part is the output AC side voltage provided by the bridge arm. The third part is the part of circulating current [4].

According to the above analysis, the voltage of the submodule is:

$$u_{co} = \frac{U_{dc}}{N} \quad (6)$$

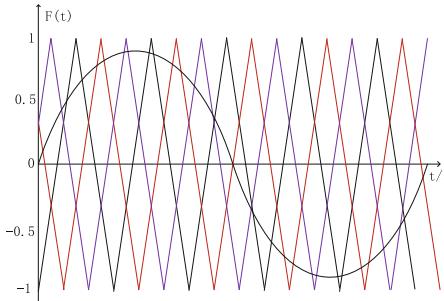
It is not difficult to see that the submodule voltage is used to maintain the balance of the DC side voltage. Therefore, the balance of the sub-module voltage is the key to the stable operation of the MMC.

### 3 MMC Carrier Phase Shift Control

#### 3.1 Principle of Carrier Phase Shift Modulation

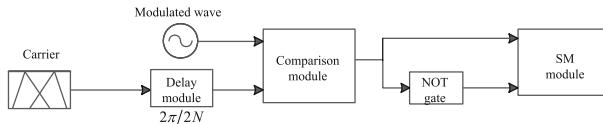
Figure 3 is a schematic diagram of the modulation wave and carrier wave of the three sub-modules of the upper half-bridge arm in the carrier phase shift modulation method. As shown in the figure, the period of the carrier is divided into  $2N$  parts to obtain the phase of the carrier phase shift. The  $2N$  phase-shifted carriers are sorted and allocated alternately to the upper and lower bridge arms, the upper bridge arm is assigned an odd number of carriers and the lower bridge arm is assigned an even number of carriers. The sinusoidal signal is selected as

the modulated wave of the lower bridge arm and the upper bridge arm, and the phase difference is  $180^\circ$ . The phases of the modulated waves of the three phases differ from each other by  $120^\circ$ .



**Fig. 3.** Schematic diagram of upper half-bridge carrier and modulated wave

Figure 4 is the principle block diagram of the open loop control of a single sub-module of the upper arm of the MMC. In the figure, the phase-shifted carrier and the modulated wave are used to control the switching device on and off. The lower bridge arm sub-module needs to lag  $180^\circ$  to control the modulated wave.

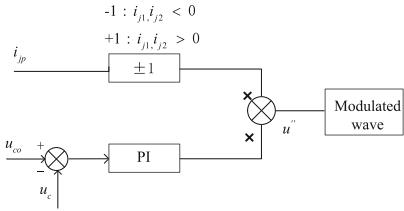


**Fig. 4.** Carrier phase shift control of single sub-module of upper bridge arm

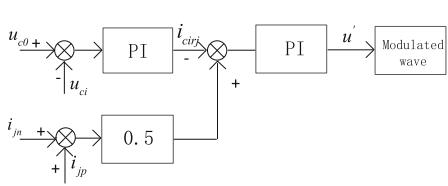
### 3.2 Carrier Phase Shift Control of Single Sub-module of Upper Bridge Arm

The closed-loop control of carrier phase shift is divided into two parts: submodule voltage equalization control and bridge arm voltage equalization control. The output through the control circuit is superimposed on the modulated wave. Further, shifting the modulation wave up and down is equivalent to changing the turn-on and turn-off time of the IGBT, thereby changing the size of the capacitor voltage and realizing the control of the capacitor voltage equalization.

Figure 5 is a block diagram of the submodule voltage equalization control. The submodule voltage equalization control compares the capacitor voltage  $U_c$  in the SM with its ideal value  $U_0$ . Charge the capacitor, the output value is



**Fig. 5.** Schematic diagram of submodule voltage equalization control



**Fig. 6.** Schematic diagram of bridge arm voltage equalization control

+1; when the bridge arm current is less than zero, the circuit discharges the capacitor, and the output value is  $-1$  [5]. The opposite is also true.

Figure 6 is a block diagram of the bridge arm voltage equalization control. If the average value of the bridge arm voltage is less than the ideal value of the capacitor voltage, the circulating current value and the reference value are both positive. The circuit charges the capacitor. When the circulating current value is greater than the circulating current reference value, the output is Positive, increase the capacitor charging time, when the circulating current value is less than the circulating current reference value, the output is negative, reduce the capacitor charging time, vice versa.

These two kinds of control are obtained by superimposing the compensation amount and the modulation wave to obtain the modulation waveform, but the principle of use is different. The submodule voltage equalization control maintains independent capacitor voltage equalization, and the bridge arm voltage equalization control is the control of a certain phase. It constitutes the pressure balance control of the MMC system [6–9].

### 3.3 CMAC Controller Design

The cerebellar model articulation controller can complete the control of the forward transmission of the system, and shorten the response time of the system, and at the same time realize the real-time adjustment of PI parameters. Compared with the traditional PID, it can complete the real-time control of the system, and use this method to improve the stability and anti-interference ability of the system [10].

The control algorithm of the system is

$$E(k) = \frac{1}{2}(u(k) - u_n(k))^2 \frac{\alpha_i}{c} \quad (7)$$

$$u(k) = u_n(k) + u_p(k) \quad (8)$$

In the formula,  $i$  is the binary selection vector,  $c$  is the generalization parameter in the network,  $u_n(k)$  is the corresponding output generated under the CMAC network, and  $u_p(k)$  is the output generated by the conventional controller PID.

The adjustment index of CMAC is

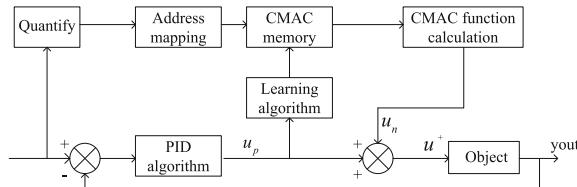
$$E(k) = \frac{1}{2}(u(k) - u_n(k))^2 \frac{\alpha_i}{c} \quad (9)$$

$$\Delta\omega(k) = \eta \frac{u(k) - u_n(k)}{c} a_i = \eta \frac{u_p(k)}{c} a_i \quad (10)$$

$$\omega(k) = \omega(k-1) + \Delta\omega(k) + \alpha(\omega(k) - \omega(k-1)) \quad (11)$$

Where  $\eta$  is the learning rate of the network;  $\alpha$  is the amount of inertia and  $\omega_k$  is the weight.

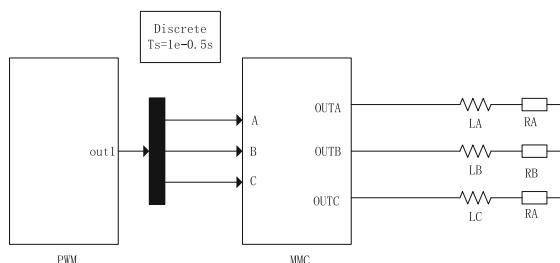
The principle block diagram of the cerebellar model articulation controller is shown in Fig. 7. At  $t=0$ , the PID algorithm controls the system. At this time,  $\omega = 0$ ,  $u = u_p$ . Under this system, with the gradual learning of the cerebellar model, the output of the traditional PID will gradually be reduced to 0, and at this time increase the output value of CMAC, the result can be regarded as the overall output [11].



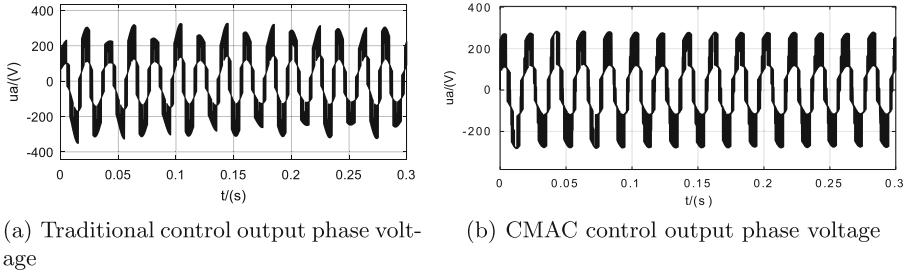
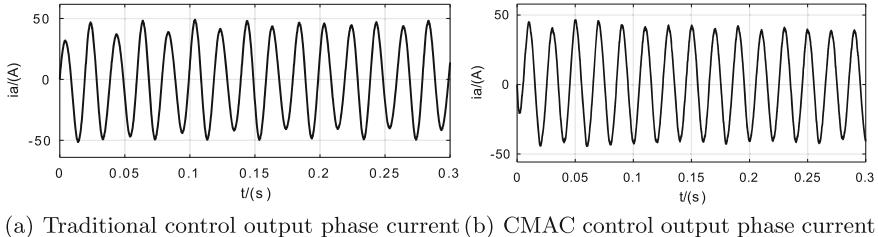
**Fig. 7.** Schematic diagram of bridge arm voltage equalization control

#### 4 MMC Capacitor Voltage Equalization Control Simulation

The simulation parameters of the modular multilevel converter are: the DC side voltage is 600 V, the sub-module (SM) capacitance is 6mF, the inductance of



**Fig. 8.** MMC simulation model based on MATLAB

**Fig. 9.** Output phase voltage under disturbance**Fig. 10.** Output phase current under disturbance

each phase arm is 5 mH, the independent sub-module capacitor voltage precharge value is 150 V, set each phase The number of submodules of the bridge arm is 3, and the simulation time is set to 0.3s.

Figure 8 is a simulation structure diagram of four-level MMC based on MATLAB. In the figure, the PWM module is a part of the control circuit based on the cerebellar model. Its internal CMAC and sub-module PI controller form an adaptive cerebellar model articulation controller. The MMC module contains the main circuit of the MMC.

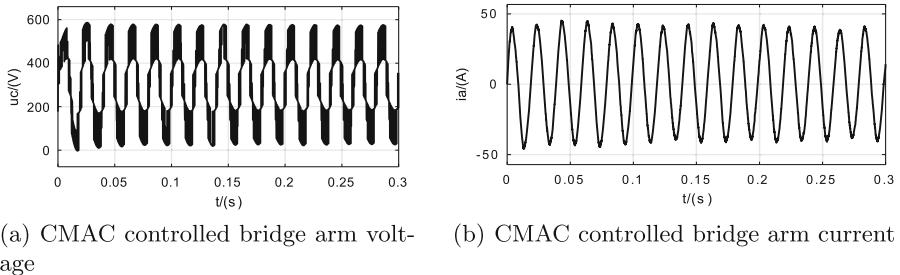
Figure 9 shows that both the traditional control and CMAC control voltage waveforms exhibit four levels, which is consistent with the operating principle of MMC. However, the traditional control voltage waveform is obviously distorted, and its voltage waveform harmonics are relatively large. Compared with the traditional control, the waveform distortion rate of the CMAC control is smaller, and the CMAC voltage equalization control method has stronger anti-interference ability (Fig. 10).

**Table 1.** AC side waveform distortion rate

	Phase voltage distortion rate THD%	Phase current distortion rate THD%
CMAC control	20.36	1.38
Traditional control	39.91	11.15

The output current waveform also shows that CMAC control is superior to traditional control. In practice, the circuit cannot maintain absolute symmetry, and the precharge is not necessarily the ideal value of 200 V. In addition, there are some uncertainties in actual operation, so set the submodule capacitor precharge value in a non-ideal state to simulate, Table 1 is A Waveform distortion rate of phase voltage and current.

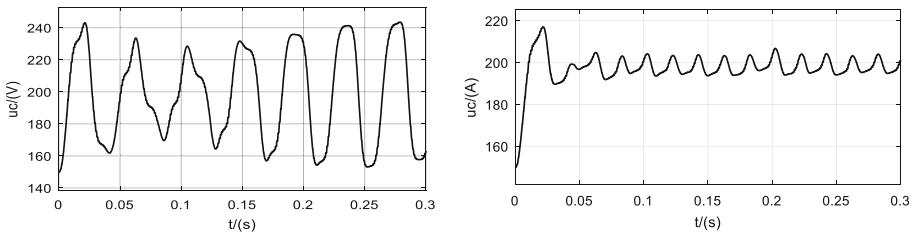
The bridge arm voltage in Fig. 11 is the same as the AC side output voltage in Fig. 9. Compared with the CMAC control waveform, the traditional control generates more harmonics. With the improved control method, the bridge arm voltage is more stable when the MMC operates in steady state. The bridge arm current controlled by CMAC in Fig. 11 (b) is exactly half of the output phase current, which is consistent with the previous analysis. The distortion rate of the traditional control bridge arm current is greater than the current controlled by the CMAC. Combined with Eq. (4), it can be seen that the MMC voltage equalization control also has a certain suppression effect on the circulating current.



(a) CMAC controlled bridge arm voltage

(b) CMAC controlled bridge arm current

**Fig. 11.** CMAC controlled bridge arm voltage and current



(a) Traditional control capacitor voltage

(b) CMAC control capacitor voltage

**Fig. 12.** CMAC control and traditional control capacitor voltage

Figure 12 shows that the deviation of the voltage of the traditional control submodule even exceeds 80 V, while the deviation of the voltage of the CMAC control submodule is 16 V. The voltage fluctuation is small, which further verifies

the effectiveness of the voltage-sharing control strategy based on the cerebellar model.

## 5 Conclusions

The traditional MMC voltage equalization control, because its PI parameters cannot be adjusted in real time in the case of a fault or a given disturbance, causes the MMC phase-to-phase circulating current to increase, which makes the capacitor voltage unstable and has a greater impact on the MMC's steady-state operation.

In this paper, a voltage equalization control strategy based on the cerebellar model is used to control the capacitor voltage balancing scheme. Based on the traditional pressure equalization control, by combining the traditional PI controller with the cerebellar model articulation controller, the pressure equalization control of the sub-module is formed. While ensuring that the capacitor voltage of the MMC sub-module is stable, the anti-interference ability of the MMC control system is improved, which also has a certain suppression effect on the circulating current of the MMC.

## References

1. Zhao, X., Zhao, C., Li, G.: Capacitor voltage balance control of modular multilevel converter using carrier phase shift technology. Proc. CSEE **21**(31), 48–55 (2011)
2. Liao, W.: Research on some key technologies of modular multilevel converter (MMC) operation and control. Hunan University (2016)
3. Lin, Z., Liu, C., et al.: Sub-module capacitor voltage hierarchical voltage equalization control method for modular multilevel converters. Autom. Electr. Power Syst. **39**(7), 175–181 (2015)
4. Minyuan, G., Zheng, X.: Optimal balance control of capacitor voltage of MMC type VSC-HVDC system. Chin. J. Electr. Eng. **31**(12), 9–14 (2011)
5. Zhao, C.: Research on modular multilevel inverter. Anhui University of Science and Technology (2012)
6. Zhou, Y., Jiang, D., Guo, J., et al.: Analysis of the voltage fluctuation and internal circulating current of the submodule of the modular multilevel converter. Chin. J. Electr. Eng. **32**(24), 8–14 (2012)
7. Wang, S., Zhou, X., Tang, G., et al.: Selection and calculation of capacitance values of submodules of modular multilevel HVDC transmission system. Power Syst. Technol. **35**(1), 26–32 (2011)
8. Tu, Q., Xu, Z., Zheng, X., et al.: An optimized voltage balance control method for modular multilevel converter. J. Electr. Eng. Technol. **26**(5), 15–20 (2011)
9. Lu, Y., Wang, Z., Peng, M., et al.: A submodule optimized voltage equalization method for modular multilevel converters. Autom. Electr. Power Syst. **38**(3), 52–58 (2014)
10. Liu, X.: Position sensorless of permanent magnet synchronous motor based on cerebellar model articulation controller. Shanghai University of Applied Sciences (2018)
11. Xin, Y., Wang, Z., Li, G., et al.: Modified multi-level converter sub-module capacitor voltage balance improved control method. Power Grid Technol. **38**(5), 1291–1296 (2014)



# Consistency of Continuous Multi-agent Systems with Privacy Protection

Meiyan Yu, Hongyong Yang<sup>(✉)</sup>, Yujiao Sun, and Fei Liu

School of Information and Electrical Engineering, Ludong University,  
Yantai 264025, China  
[hyyang@yeah.net](mailto:hyyang@yeah.net)

**Abstract.** In order to solve the privacy protection problem of a group of continuous multi-agent with communication delay, a consistency algorithm with privacy function is designed, which can make the agent converge to the common value accurately by using event triggering mechanism. Firstly, the condition of privacy protection function is given, and a function which can hide the real state of agents is constructed. Then, the consistency algorithm is proposed, and the convergence analysis is given. Ultimately, the effectiveness and feasibility of the designed formation algorithm are confirmed by simulation experiments.

**Keywords:** Privacy protection · Event triggering · Multi-agent formation control · Consistency convergence

## 1 Introduction

In the wake of developments in science and technology, the multi-agent system is becoming more and more important in control theory and control science. A multi-agent system is a new distributed computing technology. It does not rely on independent individuals to achieve the goal, but through the cooperation of multiple agents to complete the task together. Multi-robot system has the characteristics of high efficiency and high self-adaptive, which makes up for the limitation of single robot.

In the past few years, multi-agent formation control has become a hotspot. Because the consistency problem is the basis of multi-agent formation control, it has attracted wide attention of the researchers. The problem of the maximum consistency of the first-order multi-agent group is studied in reference [1]. Aiming at the problem of high-order consistency of a heterogeneous multi-agent system which has unknown correspondence delay is studied in reference [2], a self-adaption adjustment algorithm is designed, which can realize the online adjustment of self-delay.

With the increase of task complexity, the number of multi-agent in a system is also increasing rapidly, which makes the communication easily exceed the bearing range of the system. Because the continuous communication between

agents will cause great network load, therefore, the multi-agent formation control based on intermittent communication technology has been favored by people. The event triggering mechanism realizes that agents in multi-agent system can communicate with neighboring nodes only when the triggering condition is met, it reduces the network load pressure [3]. In reference [4,5], the problem of event-triggered consistency of the linear multi-agent system is studied. An event triggered consistency algorithm based on predictor is designed in [5], which can achieve consistency without constant data communication of agents.

Generally speaking, the basis of multi-agent formation control is that agents can communicate with neighboring agents, which will lead to privacy disclosure. It maybe cause losses. Therefore, how to prevent agents privacy disclosure has been the focus of academic circles [6]. In reference [7], the average consistency of discrete privacy protection is studied. By adding random noise, the agents finally converge to the average value of the initial value. Because the emerging privacy attacks make simple privacy protection have great risks. The concept of differential privacy with a rigorous statistical model has attracted a lot of attention. Differential privacy mechanism prevents sensitive information from being disclosed by randomizing the output. Reference [8] studies the discrete form of differential privacy consistency algorithm, which combines event triggering with differential privacy, it realizes privacy protection and reduces network load.

Because the differential privacy method is only suitable for the discrete multi-agent algorithm, and in the theoretical analysis, the verification of the algorithm needs a lot of computation. In addition, the existing literature rarely considers the communication delay. So based on the above kinds of literature, we study the consistency of privacy protection of continuous multi-agent with communication delay. This paper mainly focuses on the following two aspects:

- (1) For the continuous multi-agent with communication delay, a function is proposed to conceal the real state of the agent. It can ensure that the continuous multi-agents converge to the same value accurately.
- (2) A privacy preserving consistency control algorithm based on event triggering mechanism is proposed, and the convergence of the algorithm is analyzed theoretically to reduce the network load.

## 2 Preliminaries

The list of agents is regarded as an undirected graph  $G$ ,  $G = (V, E)$ , where  $V = \{1, \dots, n\}$  is the vertex collection and  $G$  is the edge collection,  $A = [a_{ij}] \in R^n (i, j \in V)$  is the adjacent matrix,  $(i, j) \in E$  iff node  $i$  and node  $j$  are able to send messages to each other, and in this case,  $a_{ij} = 1$ , or else  $a_{ij} = 0$ .

Consider the matrix  $L$  as the Laplacian matrix of graph  $G$ ,  $L = D - A$ , where  $D = diag\{d_1, \dots, d_n\}$  is the degree matrix of graph  $G$ , and  $d_i = \sum_{j \in N_i} a_{ij} \in R^n (i = 1, \dots, n)$ ,  $N_i = \{j \in V | (i, j) \in E\}$ .

For the privacy protection of agents, we hope that the agent-real information cannot be obtained by other agents when it communicates with other agents. According to the privacy protection requirements of the system state value,

we introduce an output function to mask the internal state of the agent, which is defined as below:

**Definition 1.** [9]  $h_i(t, x_i, \pi_i)$  is a function that can mask the internal state performance of the agent if the following conditions are met:

- C1 :  $h_i(t, x_i, \pi_i) \neq x_i, \forall x_i \in R^n$ ;
- C2 :  $h_i(t, x_i, \pi_i)$  ensures the unrecognizability of the initial conditions;
- C3 : To  $\forall x_i \in R^n$ ,  $h_i(t, x_i, \pi_i)$  cannot protect the nearby value of agent status;
- C4 : Given  $t, \pi_i, i = 1, 2, \dots, n$ ,  $h_i(t, x_i, \pi_i)$  increases monotonically with  $x_i$ ;
- C5 : Given  $t$  and  $\pi_i$ ,  $\lim_{t \rightarrow \infty} h_i(t, x_i, \pi_i) = x_i, i = 1, 2, \dots, n$ .

Where  $\pi_i$  is the set of parameter variables in  $h_i$ , it will be given in the  $h_i$  design.

Note. Definition 1 gives the requirements of the function to protect the sensitive information of agents. Condition C1 can make the value of the function of masking state not equal to the real value. The condition C2 and C3 are to protect the initial state of the agent from being stolen. The definition of condition C4 is similar to the function  $K_\infty$ , but more generally,  $h_i(t, x_i, \pi_i)$  is not nonnegative with respect to  $x$ . The condition C5 is necessary to make the function converge to the true value. Definition 1 shows that the function satisfying condition C1 – C5 can ensure that the initial state of the agent is not leaked.

Based on the above description, we construct the following privacy protection functions:

$$\begin{aligned} y_i(t) &= h_i(t, x_i, \pi_i) \\ h_i(t, x_i, \pi_i) &= (m_i + \phi_i e^{-\zeta_i t}) f_i x_i(t) + \alpha_i e^{-\mu_i t} \end{aligned} \quad (1)$$

where  $y_i(t)$  is the output of function  $h_i(t, x_i, \pi_i)$ ,  $\pi_i = \{\phi_i, \zeta_i, \alpha_i, m_i, f_i, \mu_i\}$ ,  $\phi_i, \zeta_i, \alpha_i, m_i, f_i, \mu_i > 0$  and satisfies  $m_i f_i = 1$ .

**Lemma 1.**  $h_i(t, x_i, \pi_i)$  in (1) is a function that conceals the real internal state.

*Proof.* Obviously, when  $\phi_i, \zeta_i, \alpha_i, m_i, f_i, \mu_i > 0$ ,  $(m_i + \phi_i e^{-\zeta_i t}) f_i x_i(t) + \alpha_i e^{-\mu_i t} \neq x_i(t)$ , C1 is satisfied.

In (1), The thieves cannot get all information of  $\pi_i = \{\phi_i, \zeta_i, \alpha_i, m_i, f_i\}$ , so they can't recognize the initial conditions. Condition C2 is satisfied.

If  $x^* \in R^n$  and  $x^*$  meets  $\|x(0) - x^*\| < \xi$ , so:

$$\begin{aligned} &\|h(0, x, \pi) - x^*\| \\ &= \| (M + \Phi) F x + \alpha - x^* \| \leq \| (M + \Phi) F x - x^* \| + \| \alpha \| \end{aligned} \quad (2)$$

$$\Phi = diag(\phi_1, \dots, \phi_n), M = diag(m_1, \dots, m_n), \alpha = [\alpha_1, \dots, \alpha_n]^T, F = diag(f_1, \dots, f_n).$$

According to the definition of neighborhood, inequality is not in the  $\xi$  neighborhood of  $x^*$ . So privacy function satisfies condition C3.

From the function of  $h_i(t, x_i, \pi_i)$ , and the function satisfies  $m_i f_i = 1$ , so  $\frac{\partial h_i(t, x_i, \pi_i)}{\partial x_i} = (m_i + \phi_i e^{-\zeta_i t}) f_i > 0$  is obtained, which means  $h_i(t, x_i, \pi_i)$  is strictly monotonically increasing about  $x_i(t)$ . Condition C4 is satisfied.

Because  $|h(t, x_i, \pi_i) - x_i| = \phi_i e^{-\zeta_i t} f_i x_i + \alpha_i e^{-\mu_i t}$  holds,  $x_i(t) > 0$ ,  $\phi_i, \zeta_i, \alpha_i, f_i > 0$ , so  $|h_i(0, x_i, \pi_i) - x_i(0)| = \phi_i f_i x_i(0) + \alpha_i > 0$ . In addition,  $\lim_{t \rightarrow \infty} |h_i(t, x_i, \pi_i) - x_i| = 0$ , from the beginning of  $t$  and the time when  $t$  tends to infinity,  $|h(t, x_i, \pi_i) - x_i| - x_i$  function is decreasing. C5 is satisfied.

To sum up, according to Definition 1,  $h_i(t, x_i, \pi_i) = (m_i + \phi_i e^{-\zeta_i t}) f_i x_i(t) + \alpha_i e^{-\mu_i t}$  is a function with privacy protection performance.

Next, we use the proposed privacy protection function to construct a privacy protection consistent control algorithm to protect the initial state privacy of the agent.

### 3 Algorithm Description

In this section, we build the following privacy consistent control algorithm based on the above privacy protection function:

$$\begin{aligned} \dot{x}_i(t) &= Ax_i(t) + Bu_i(t) \\ u_i(t) &= c \sum_{j \in N_i} a_{ij} (y_j(t_k^j - \tau) - y_i(t_k^i)) \\ y_i(t) &= h_i(t, x_i, \pi_i) \\ h_i(t, x_i, \pi_i) &= (m_i + \phi_i e^{-\zeta_i t}) f_i x_i(t) + \alpha_i e^{-\mu_i t} \end{aligned} \quad (3)$$

where  $x_i(t)$  is the internal state of agent  $i$ . The output of the  $h_i(t, x_i, \pi_i)$  is  $y_i(t)$ ,  $m_i f_i = 1$ .

$y_i(t_k^i)$  is the output value of  $h_i(t, x_i, \pi_i)$  at the event triggering time of  $t_k^i$ .  $u_i(t)$  is the input function of the system,  $h_i(t, x_i, \pi_i)$  is the privacy protection function. The system's communication delay is  $\tau$ ,  $c > 0$  is the control gain, matrix A and matrix B are constant matrices. Here, we think  $u_i(t)$  and  $y_i(t)$  are public values,  $x_i(t)$  and  $h_i(t, x_i, \pi_i)$  are agent's privacy values.

Because frequent communication between agents will cause a lot of network load, event triggering strategy is used to reduce the communication frequency between agents. We set an event for the agents to decide whether to send the status information  $y_i(t)$  to other agents. The event trigger function is as follows:

$$t_{k+1}^i = \inf\{t > t_k^i \mid g_i(t) > 0\} \quad (4)$$

The event trigger function  $g_i(t)$  is designed as  $g_i(t) = \|e_i(t)\| - \gamma_i e^{-\eta_i t}$ . Where  $e_i(t)$  is the measurement error, which is defined as  $e_i(t) = y_i(t_k^i) - y_i(t)$ , and  $\gamma_i, \eta_i > 0$  is the parameters to be set.

### 4 Consistency Analysis

In the first section, we prove that  $h_i(t, x_i, \pi_i)$  is a function with privacy to protect the initial state of agent. Next we analyze the consistency of system (3).

From  $e_i(t) = y_i(t_k^i) - y_i(t)$ ,  $MF = I$ ,  $I$  is the identity matrix. In light of algorithm (3), its vector form is as follows:

$$\begin{aligned} \dot{x}(t) &= [I_N \otimes A - c(D \otimes B)(M + \phi e^{-\zeta t})] Fx(t) \\ &\quad + c(L \otimes B + D \otimes B)(M + \phi e^{-\zeta(t-\tau)}) Fx(t - \tau) \\ &\quad + c(L \otimes B + D \otimes B)(M + \phi e^{-\zeta(t-\tau)}) e(t - \tau) - c(D \otimes B)e(t) + \varphi(t) \end{aligned} \quad (5)$$

Where  $x(t) = [x_1(t), \dots, x_n(t)]^T$ ,  $e(t) = [e_1(t), \dots, e_n(t)]^T$

$$\phi = \text{diag}(\phi_1, \dots, \phi_n), \zeta = \text{diag}(\zeta_1, \dots, \zeta_n), \mu = \text{diag}(\mu_1, \dots, \mu_n)$$

$$\varphi(t) = c(D \otimes B)\alpha e^{-\mu t} + c(L \otimes B + D \otimes B)\alpha e^{-\mu(t-\tau)}, D = \text{diag}(d_1, \dots, d_n)$$

**Lemma 2.** If all eigenvalues of matrix  $K$  have negative real parts, then there exists constants  $\beta \geq 1, \epsilon > 0$  that make the following inequality be true:

$$\| e^{K(t-t_0)} \| \leq \beta e^{-\epsilon(t-t_0)} \quad (6)$$

Let  $\underline{\eta} = \min\{\eta_1, \dots, \eta_n\}$ ,  $\underline{\zeta} = \min\{\zeta_1, \dots, \zeta_n\}$ ,  $\underline{\mu} = \min\{\mu_1, \dots, \mu_n\}$ ,  $\kappa_1 = L \otimes B + D \otimes B$ . The following Theorem 1 holds.

**Theorem 1.** If there are constants  $0 < \varrho < \epsilon < v = \min\{\underline{\eta}, \underline{\zeta}, \underline{\mu}\}$ , communication delay  $\tau \in [0, \tau_0]$ ,  $\tau_0 = \frac{\ln \frac{\epsilon - \varrho}{\beta c \|\kappa_1\|}}{\varrho}$  and matrix satisfies  $(I_N \otimes A - cM(D \otimes B))F$  that it's all eigenvalues have negative real parts, then the consistency algorithm we proposed can guarantee the asymptotic consistency of the multi-agent system.

*Proof.* Integral (5) to get:

$$\begin{aligned} x(t) &= e^{F \int_{t_0}^t [I_N \otimes A - c(D \otimes B)(M + \phi e^{-\zeta s})] ds} x(t_0) \\ &\quad + c \int_{t_0}^t e^{F \int_{t_0}^\theta [I_N \otimes A - c(D \otimes B)(M + \phi e^{-\zeta s})] ds} \Gamma(\theta) d\theta \end{aligned} \quad (7)$$

where  $\Gamma(\theta) = \kappa_1(H + \phi e^{-\zeta(\theta-\tau)})Fx(\theta - \tau) - \kappa_2 e(\theta) + \kappa_1(H + \phi e^{-\zeta(\theta-\tau)})e(\theta - \tau) + \varphi(\theta)$ ,  $\kappa_2 = D \otimes B$

Owing to:

$$\begin{aligned} &e^{\int_{t_0}^t [I_N \otimes A - c(D \otimes B)(M + \phi e^{-\zeta s})] F ds} \\ &= e^{[(I_N \otimes A - cM(D \otimes B))F(t-t_0)]} e^{-\frac{c(D \otimes B)\phi F}{\zeta} (e^{-\zeta t_0} - e^{-\zeta t})} \\ &\leq e^{[(I_N \otimes A - cM(D \otimes B))F(t-t_0)]} \end{aligned} \quad (8)$$

Because all eigenvalues of  $[(I_N \otimes A - cM(D \otimes B))F]$  have negative real parts, according to Lemma 2, we can get:

$$e^{\int_{t_0}^t [I_N \otimes A - c(D \otimes B)(M + \phi e^{-\zeta s})] F ds} \leq \beta e^{-\epsilon(t-t_0)} \quad (9)$$

According to the definition of  $\underline{\eta}, \underline{\zeta}, \underline{\mu}$ , we have:

$$\| x(t) \| \leq \beta e^{-\epsilon(t-t_0)} \| x(t_0) \| + \beta c \int_{t_0}^t e^{-\epsilon(t-\theta)} p(\theta) d\theta \quad (10)$$

where  $p(\theta) = \| \kappa_1 \| (\| M \| + \| \phi \| e^{-\underline{\zeta}(\theta-\tau)}) \| F \| \| x(\theta - \tau) \| + \| \kappa_2 \| \| e(\theta) \| + \| \kappa_1 \| (\| M \| + \| \phi \| e^{-\underline{\zeta}(\theta-\tau)}) \| e(\theta - \tau) \| + \| \varphi(\theta) \|$

At the same time:

$$\| \varphi(t) \| \leq c \| \kappa_2 \| \| \alpha \| e^{-\underline{\mu}t} + c \| \kappa_1 \| \| \alpha \| e^{-\underline{\mu}(t-\tau)} \quad (11)$$

And from the function (4), we can get  $\| e(t) \| \leq \gamma \| e^{-\eta t} \|. Take it into (10) and combine it with formula (11), we get the following formula:$

$$\begin{aligned}
\|x(t)\| \leq & \beta c \|\kappa_2\| \|\gamma\| \int_{t_0}^t e^{-\epsilon(t-\theta)} e^{-\underline{\eta}\theta} d\theta \\
& + \beta c \|\kappa_1\| \int_{t_0}^t e^{-\epsilon(t-\theta)} (\|M\| + \|\phi\| e^{-\underline{\zeta}(\theta-\tau)}) \|F\| \|x(\theta-\tau)\| d\theta \\
& + \beta e^{-\epsilon(t-t_0)} \|x(t_0)\| + \beta c \|\kappa_1\| \|M\| \|\gamma\| \int_{t_0}^t e^{-\epsilon(t-\theta)} e^{-\underline{\eta}(\theta-\tau)} d\theta \\
& + \beta c \|\kappa_1\| \|\gamma\| \|\phi\| \int_{t_0}^t e^{-\epsilon(t-\theta)} e^{-(\underline{\zeta}+\underline{\eta})(\theta-\tau)} d\theta \\
& + \beta c^2 \|\kappa_2\| \|\alpha\| \int_{t_0}^t e^{-\epsilon(t-\theta)} e^{-\underline{\mu}\theta} d\theta \\
& + \beta c^2 \|\kappa_1\| \|\alpha\| \int_{t_0}^t e^{-\epsilon(t-\theta)} e^{-\underline{\mu}(\theta-\tau)} d\theta
\end{aligned} \tag{12}$$

We integrate some subexpressions in (12) and remove the negative subformulas from the right side of the inequality, inequality (13) is obtained.

$$\begin{aligned}
\|x(t)\| \leq & \beta e^{-\epsilon(t-t_0)} \|x(t_0)\| \\
& + \beta c \|\kappa_1\| \int_{t_0}^t e^{-\epsilon(t-\theta)} (\|M\| + \|\phi\| e^{-\underline{\zeta}(\theta-\tau)}) \|F\| \|x(\theta-\tau)\| d\theta \\
& + \frac{\beta c \|\kappa_2\| \|\gamma\| e^{-\underline{\eta}t_0} e^{-\epsilon(t-t_0)}}{\underline{\eta}-\epsilon} + \frac{\beta c \|\kappa_1\| \|M\| \|\gamma\| e^{\underline{\eta}(\tau-t_0)} e^{-\epsilon(t-t_0)}}{\underline{\eta}-\epsilon} \\
& + \frac{\beta c \|\kappa_1\| \|\gamma\| \|\phi\| e^{(\underline{\zeta}+\underline{\eta})(\tau-t_0)} e^{-\epsilon(t-t_0)}}{(\underline{\zeta}+\underline{\eta})-\epsilon} + \frac{\beta c^2 \|\kappa_2\| \|\alpha\| e^{-\underline{\mu}t_0} e^{-\epsilon(t-t_0)}}{\underline{\mu}-\epsilon} \\
& + \frac{\beta c^2 \|\kappa_1\| \|\alpha\| e^{\underline{\mu}(\tau-t_0)} e^{-\epsilon(t-t_0)}}{\underline{\mu}-\epsilon}
\end{aligned} \tag{13}$$

Next, we only need to prove that formula (14) holds when  $\sigma > 0$ .

$$\|x(t)\| < \sigma Z e^{-\varrho(t-t_0)} \tag{14}$$

Where  $0 < \varrho < \epsilon < \nu = \min\{\eta, \zeta, \mu\}$ .  $Z$  is a nonnegative matrix.

Let  $\|z(t)\| = \sigma Z e^{-\varrho(t-t_0)}$ , if (14) does not hold, then there must be a  $t^* > t_0$  to make  $\|x(t^*)\| = \|z(t^*)\|$  and  $\|x(t)\| < z(t)$  hold for  $t < t^*$ . Replace  $\|x(\theta-\tau)\| d\theta$  and integrate  $\beta c \|\kappa_1\| \int_{t_0}^t e^{-\epsilon(t-\theta)} (\|M\| + \|\phi\| e^{-\underline{\zeta}(\theta-\tau)}) \|F\| \|x(\theta-\tau)\| d\theta$ , the results are as follows:

$$\begin{aligned}
& \beta c \|\kappa_1\| \int_{t_0}^t e^{-\epsilon(t-\theta)} (\|M\| + \|\phi\| e^{-\underline{\zeta}(\theta-\tau)}) \|F\| \|x(\theta-\tau)\| d\theta \\
& \leq \frac{\beta c \|\kappa_1\| \|\sigma Z e^{\varrho\tau} (e^{-\varrho(t^*-t_0)} - e^{-\epsilon(t^*-t_0)})}{\epsilon-\varrho} + \frac{\beta c \|\kappa_1\| \|F\| \|\phi\| \sigma Z e^{-\underline{\zeta}t_0 + (\underline{\zeta}+\varrho)\tau} e^{-\epsilon(t^*-t_0)}}{\underline{\zeta}+\varrho-\epsilon}
\end{aligned} \tag{15}$$

$$\text{Let } A_1 = \frac{\beta c \|\kappa_1\| \|F\| \|\phi\| \sigma e^{-\underline{\zeta}t_0}}{\underline{\zeta}+\varrho-\epsilon}, B = \frac{\beta c \|\gamma\| e^{-\underline{\eta}t_0}}{\underline{\eta}-\epsilon} (\|\kappa_2\| + \|\kappa_1\| e^{-\underline{\eta}t_0}) + \frac{\beta c \|\kappa_1\| \|\gamma\| \|\phi\| e^{-(\underline{\zeta}+\underline{\eta})t_0}}{(\underline{\zeta}+\underline{\eta})-\epsilon} e^{-(\underline{\eta}+\underline{\zeta})\tau} + \frac{\beta c^2 \|\alpha\| e^{-\underline{\mu}t_0}}{\underline{\mu}-\epsilon} (\|\kappa_2\| + \|\kappa_1\| e^{\underline{\mu}\tau}), A_1, B > 0.$$

According to (13)–(15) and the relevant definitions above, we can get:

$$\begin{aligned}
\|x_i(t)\| \leq & (\beta \|x(t_0)\| + A_1 Z e^{(\underline{\zeta}+\varrho)\tau} + B) e^{-\epsilon(t-t_0)} \\
& + \frac{\beta c \|\kappa_1\| \|\sigma Z e^{\varrho\tau} (e^{-\varrho(t-t_0)} - e^{-\epsilon(t-t_0)})}{\epsilon-\varrho}
\end{aligned} \tag{16}$$

Next, we consider  $\beta \|x(t_0)\| + A_1 Z e^{(\underline{\zeta}+\varrho)\tau} + B \geq \frac{\beta c \|\kappa_1\| \|\sigma Z e^{\varrho\tau}\|}{\epsilon-\varrho}$  and  $\beta \|x(t_0)\| + A_1 Z e^{(\underline{\zeta}+\varrho)\tau} + B < \frac{\beta c \|\kappa_1\| \|\sigma Z e^{\varrho\tau}\|}{\epsilon-\varrho}$  two situations, by using the zero point theorem, it is proved that the existence of  $\varrho$  makes  $\frac{\beta c \|\kappa_1\| \|\sigma Z e^{\varrho\tau}\|}{\epsilon-\varrho} \leq 1$  and  $\tau_0 = \frac{\ln \frac{\epsilon-\varrho}{\beta c \|\kappa_1\|}}{\varrho}$  holds. Then we can get  $\|x(t)\| < \sigma Z e^{-\varrho(t-t_0)}$  holds in both cases. Due to the space problem, the proof is omitted here.

In the following chapters, we prove that when  $t > t_k^i$ , the Zeno effect of event triggering mechanism is excluded.

First, according to  $e_i(t) = y_i(t_k^i) - y_i(t)$ , we take the derivative of  $e_i(t)$  and make it into the following formula:

$$\begin{aligned} \frac{d(\|e_i(t)\|)}{dt} &\leq \|e_i(t)\| \\ &\leq \|A\| \|e_i(t)\| + \|B\| \|u_i\| f_i \phi_i e^{-\zeta_i t} \\ &\quad + \|A\| \alpha_i \mu_i e^{-\mu_i t} + y_i(t_k^i) + \|B\| \|u_i\| \end{aligned} \quad (17)$$

Integrate the above formula and let  $t = t_{k+1}^i$ :

$$\begin{aligned} \|e_i(t_{k+1}^i)\| &\leq \frac{\|B\| \|u_i\| f_i \phi_i e^{-\zeta_i t_{k+1}^i} e^{(\|A\| + \zeta_i)(t_{k+1}^i - t_k^i)}}{\|A\| + \zeta_i} \\ &\quad + \frac{\|A\| \alpha_i \mu_i e^{-\mu_i t_{k+1}^i} e^{(\|A\| + \mu_i)(t_{k+1}^i - t_k^i)}}{\|A\| + \mu_i} + \frac{(\|B\| \|u_i\| + y_i(t_k^i)) e^{\|A\|(t_{k+1}^i - t_k^i)}}{\|A\|} - \Delta_i \end{aligned} \quad (18)$$

$$\text{Where } \Delta_i = \frac{\|B\| \|u_i\| f_i \phi_i e^{-\zeta_i t_{k+1}^i}}{\|A\| + \zeta_i} + \frac{\|A\| \alpha_i \mu_i e^{-\mu_i t_{k+1}^i}}{\|A\| + \mu_i} + \frac{(\|B\| \|u_i\| + y_i(t_k^i))}{\|A\|}.$$

According to (4), We can get  $\|e_i(t_{k+1}^i)\| \leq \gamma_i e^{-\eta_i t_k^i}$ , combined with formula (18), we work out the following formula:

$$t_{k+1}^i - t_k^i \geq \frac{\ln \frac{\gamma_i e^{-\eta_i t_k^i} + \Delta_i}{\Delta_i}}{\|A\| + \|\zeta_i\| + \mu_i} \quad (19)$$

Because  $\frac{\gamma_i e^{-\eta_i t_k^i} + \Delta_i}{\Delta_i} > 1$ , that is,  $\ln \frac{\gamma_i e^{-\eta_i t_k^i} + \Delta_i}{\Delta_i} > 0$ , we get  $t_{k+1}^i - t_k^i > 0$ . The Zeno effect was excluded.

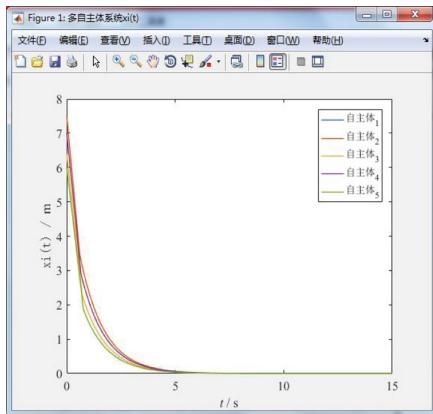
## 5 Experimental Verification

In this section, we test the validity of the consistency algorithm by experimental simulation.

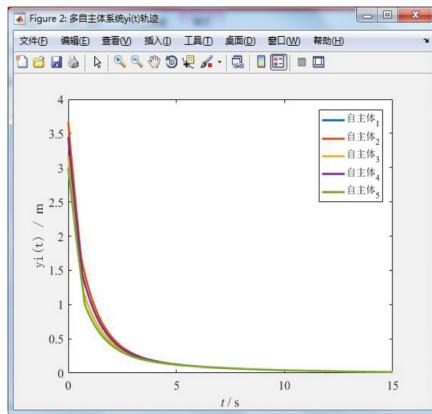
For convenience, an undirected network with five agents is considered. In the whole simulation, only one internal state of the agent is taken into account, so the system matrix  $A$  and the input matrix  $B$  are taken as constants,  $A = -0.9, B = 0.05, \gamma_i = 0.7, \phi_i = 0.7, m_i = 2, \mu_i = 0.8, \alpha_i = 0.2, f_i = 5, \zeta_i = 0.3, i = 1, \dots, 5, c = 0.1, \tau = 0.1$  is taken, and the adjacent matrix of agents is taken as:

$$\begin{pmatrix} 0.1 & 0.3 & 0 & 0.1 & 0.2 \\ 0.3 & -0.2 & 0.1 & 0.2 & 0.1 \\ 0 & 0.1 & -0.2 & 0.1 & -0.1 \\ 0.1 & 0.2 & 0.1 & 0.3 & 0.1 \\ 0.2 & 0.1 & -0.1 & 0.1 & -0.1 \end{pmatrix}, \text{ The experimental results are shown in Fig. 1, 2, 3 and 4:}$$

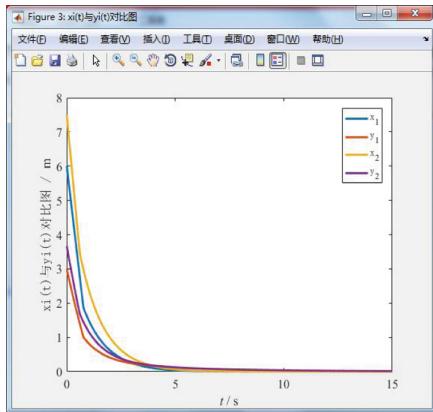
Figure 1 shows the trajectory of the internal state  $x_i(t)$  of the agent. It can be seen from the figure that the agents finally converges to the same value. Figure 2 shows the trace of the function  $y_i(t)$  which masks the agents' internal state of  $x_i(t)$ ; it can be seen from Fig. 3 that  $x_i(t)$  and  $y_i(t)$  tend to the same value. In this paper, the trigger mechanism is adopted, and Fig. 4 marks the event trigger time of each agent. As we can see from the figure, agent  $i$  is triggered once every interval.



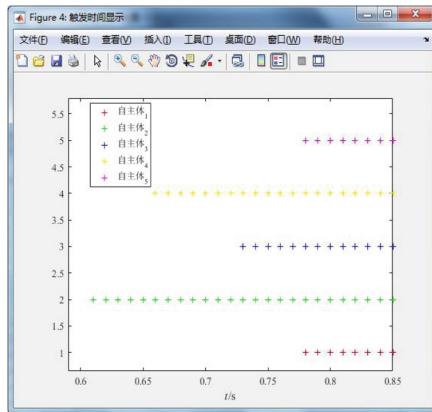
**Fig. 1.** The trace of agent's internal state  $x_i(t)$



**Fig. 2.**  $y_i(t)$  state track of agent



**Fig. 3.**  $x_i(t)$ ,  $y_i(t)$  state trajectories of agents 1 and 2



**Fig. 4.** Event trigger time of each agent

## 6 Conclusion

Based on the event triggering mechanism, this paper studies the privacy protection consistency of continuous multi-agent with time delay, a function that can hide the real state of the agent is constructed. We also propose a consistency algorithm based on the event triggering mechanism and make a detailed theoretical analysis of the convergence of the algorithm. In the next step, the privacy protection of multi-agent will be further studied combined with quantitative communication.

**Acknowledgements.** The work is supported by the National Natural Science Foundation of China (61673200, 61771231), the Major Basic Research Project of Natural Science Foundation of Shandong Province of China (ZR2018ZC0438) and the Key Research and Development Program of Yantai City of China(2019XDHZ085).

## References

1. Wang, F.Y., Yang, H.Y., Weng, S.: Maximum consistency of complex multi-agent system. *Comput. Simul.* **06**, 409–412 (2015)
2. Tian, Y.P., Zhang, Y.: High-order consensus of heterogeneous multi-agent systems with unknown communication delays. *Automatica* **48**(6), 1205–1212 (2012)
3. Dimarogonas, D.V., Frazzoli, E., Johansson, K.H.: Distributed event-triggered control for multi-agent systems. *IEEE Trans. Autom. Control* **57**(5), 1291–1297 (2012)
4. Cheng, B., Li, Z.: Fully distributed event-triggered protocols for linear multiagent networks. *IEEE Trans. Autom. Control* **64**(4), 1655–1662 (2019)
5. Sun, J., Yang, Q., Liu, X.: Event-triggered consensus for linear continuous-time multi-agent systems based on a predictor. *Inf. Sci.* **2018**, S0020025518302056 (2018)
6. Yin, C., Xi, J., Sun, R.: Location privacy protection based on differential privacy strategy for big data in industrial Internet-of-Things. *IEEE Trans. Ind. Inf.* **14**(8), 3628–3636 (2017)
7. Manitara, N.E., Hadjicostis, C.N.: Privacy-preserving asymptotic average consensus. In: 2013 European Control Conference (ECC), pp. 760–765. IEEE, 17 July 2013
8. Gao, L.: Research on Multi-agent Consistency and Differential Privacy Protection Based on Event Triggered Control and Quantitative Communication (2018)
9. Wang, A.J.: Research on Multi-agent Consistency and Privacy Protection. Southwest University (2019)



# Short-Term Prediction of Photovoltaic Power Generation Based on Deep Belief Network with Momentum Factor

Lai Lei<sup>1</sup>, Jiangzhen Guo<sup>2</sup>, Fuzhong Wang<sup>3(✉)</sup>, and Li Zhang<sup>3</sup>

<sup>1</sup> Zhengzhou Electric Power College, Zhengzhou 450000, China

<sup>2</sup> State Grid of China Technology College, Jinan 250002, China

<sup>3</sup> School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454000, China  
Wangfzh@hpu.edu.cn

**Abstract.** Accurate prediction of short-term photovoltaic power generation will help the grid dispatching department to make reasonable arrangements to ensure the safe and stable operation of the power system, thereby increasing the proportion of new energy in the power system. In order to improve the accuracy of photovoltaic power generation prediction model, a short-term photovoltaic power generation prediction model based on deep belief network with momentum factor is proposed. It focuses on the selection of training samples for short-term prediction models of photovoltaic power generation, short-term prediction models of photovoltaic power generation based on deep belief networks, and optimization of the model by adding dynamically adjusted momentum factors. The actual measured data of the Australian Desert Solar Research Center were used to verify the short-term prediction model of photovoltaic power generation. The simulation experiment results show that the short-term prediction model of photovoltaic power generation established in this paper can effectively characterize the various factors that affect photovoltaic power generation and the measured the complex non-linear relationship between powers has high prediction accuracy.

**Keywords:** PV power generation · Forecasting method · Similar day · Deep belief network · Momentum factor

## 1 Introduction

Photovoltaic power generation plays an important part in improving the energy structure and protecting the environment. However, due to the impact of weather changes, photovoltaic power generation has a certain degree of randomness and fluctuation, and large-scale access to the power system will cause a certain impact on it [1]. Accurate prediction of photovoltaic power generation will help the grid dispatching department to make reasonable arrangements and ensure the stability of the power system. At the same time, it can reduce the waste of

© The Editor(s) (if applicable) and The Author(s), under exclusive license

to Springer Nature Singapore Pte Ltd. 2021

Y. Jia et al. (Eds.): CISC 2020, LNEE 705, pp. 778–791, 2021.

[https://doi.org/10.1007/978-981-15-8450-3\\_81](https://doi.org/10.1007/978-981-15-8450-3_81)

photovoltaic power and increase the proportion of new energy in the power system.

At present, photovoltaic power generation prediction methods can be divided into two types: direct prediction method and indirect prediction method [2]. The modeling process of indirect prediction method [3] is complicated, so it is rarely applied at present. Direct prediction methods mainly include time series, Artificial Neural Network (ANN), Support Vector Machine (SVM), and combined model prediction methods.

Time series method is mainly based on a linear model. The modeling process is simple, but it has certain limitations when dealing with the nonlinear part of photovoltaic power generation [4,5].

ANN method overcomes the limitations of time series method in predicting the non-linear part. References [6–8] use Back Propagation Neural Network (BPNN), Elman neural network, and Radial Basis Function neural network (RBFNN) establishes prediction models and optimizes them to a certain degree. These models have added meteorological characteristics features to improve the accuracy of predictions. However, neural networks tend to fall into local optimal solutions. and most neural networks currently use a three-layer shallow structure, poor characterization in the face of large, complex, and multidimensional data.

The learning process of SVM is a convex optimization process, which can find the global optimal solution of the objective function [9,10], thus solving the problem that ANN easily falls into the local optimal solution. However, SVM has slow convergence speed and low accuracy when processing large, complex, multi-dimensional data, and different parameters such as the kernel function will have a greater impact on its prediction results.

The combined prediction method uses the weighted calculation of the prediction results of each single model to obtain the final result [11]. References [12] proposes a combination of Auto-Regressive Moving Average Model (ARMA) and ANN, so that both prediction accuracy and speed are taken into account; Reference [13] optimized the combination of four gray prediction models, and automatically adjusted the combination weights according to the deviation between the actual output and the expected output, so as to obtain a more accurate prediction effect. However, the combination prediction method still belongs to the preferred weighted combination of the above models in essence, and does not solve the problems existing in the above shallow model.

Deep Belief Network (DBN) [14] has a strong fitting ability, which can quickly analyze large, complex and multi dimensional data. The short-term prediction model of photovoltaic power generation is established by using DBN, and the complex non-linear relationship between various factors affecting photovoltaic power generation and the measured power is hierarchically represented by the multilayer network structure of DBN. However, the gradient descent method used in the fine-tuning stage of the traditional DBN algorithm itself has problems such as slow convergence speed and easy fall into local optimum. For this reason,

a dynamically adjusted momentum factor [15] is introduced to solve the above problems.

## 2 Prediction Model Training Sample Selection

Improper selection of prediction model training samples will cause the parameters obtained by model training to not reach the optimal solution, and increase the amount of calculation of the model, thereby affecting the speed and accuracy of prediction model prediction. This paper uses the principle of similar days to determine the training samples, and determines the training sample data required for the prediction model based on the similarity between the historical days and the days to be tested, reducing the amount of data used by the training model [16]. Based on the improvement in references [16], the weather feature vector is set to hourly temperature, humidity, and light intensity, which further improves the selection accuracy of training samples.

The selection process of prediction model training samples is as follows:

The first step is to analyze the meteorological factors affecting photovoltaic power generation. The meteorological data and the actual power generation data are calculated by PEARSON coefficient, and the results are shown in Table 1 below. Therefore, the temperature, humidity, and light intensity data with the highest correlation are selected as the basis for the selection of training samples.

**Table 1.** Pearson coefficient between meteorological factors and photovoltaic power

Influence factor	Wind speed	Temperature	Humidity	Light	Intensity
Correlation	Coefficient	0.152	0.427	-0.378	0.928

The second step is to normalize the original data in order to unify the dimensions.

The third step is to determine the weight of different meteorological feature vectors by using the weighted method. There are sample data, each sample has  $k$  meteorological factor parameters, forming an  $n \times k$  order matrix  $[\varphi_{\lambda\varepsilon}]_{n \times k}$ ,  $\varphi_{\lambda\varepsilon}$  is the value of the  $\lambda - th$  meteorological factor parameter of the  $\varepsilon - th$  sample, and its value is:

$$E_\varepsilon = -\frac{1}{\ln n} \cdot \sum_{i=1}^n \theta_{\lambda\varepsilon} \ln \theta_{\lambda\varepsilon} \quad (1)$$

In Eq. (1),  $E_\varepsilon$  is the sinter of the  $\varepsilon - th$  meteorological factor parameter of the  $\lambda - th$  sample;  $\theta_{\lambda\varepsilon} = \frac{\varphi_{\lambda\varepsilon}}{\sum_{\lambda=1}^n \varphi_{\lambda\varepsilon}}$ ;  $\varepsilon = 1, 2, \dots, k$ .

At  $\theta_{\lambda\varepsilon} = 0$ , let  $\theta_{\lambda\varepsilon} \ln \theta_{\lambda\varepsilon} = 0$ , there are:

$$\omega_\varepsilon = \frac{1 - E_\varepsilon}{\sum_{\varepsilon=1}^k (1 - E_\varepsilon)} \quad (2)$$

In Eq. (2),  $\omega_\varepsilon$  is the weight of the  $\varepsilon - th$  meteorological factor parameter;  $\omega_\varepsilon \in [0, 1]$  and  $\sum_{\varepsilon=1}^k (\omega_\varepsilon) = 1$ .

The fourth step is to calculate the weighted euclidean distance of the meteorological parameter feature vector between the  $\lambda - th$  sample and the sample to be measured. Let the normalized meteorological parameter feature vector of the  $\lambda - th$  sample after normalization be  $\mathbf{x}_\lambda = [x_\lambda(1), x_\lambda(2), \dots, x_\lambda(k)]^T$ , the sample to be tested is  $\mathbf{x}_0 = [x_0(1), x_0(2), \dots, x_0(k)]^T$ , then the weighted Euclidean distance between them is:

$$D_\lambda = 1 - \sqrt{\sum_{\varepsilon=1}^k \omega_\varepsilon [x_\lambda(\varepsilon) - x_0(\varepsilon)]^2} \quad (3)$$

The fifth step is to calculate the correlation coefficient between the  $\lambda - th$  sample and the  $\varepsilon - th$  meteorological feature component of the sample to be tested:

$$\mu_\lambda(\varepsilon) = \frac{\Delta_{\min} + 0.5\Delta_{\max}}{\Delta_\lambda(\varepsilon) + 0.5\Delta_{\max}} \quad (4)$$

In Eq. (4),  $\Delta_\lambda(\varepsilon) = |x_0(\varepsilon) - x_\lambda(\varepsilon)|$ ; 0.5 is the resolution coefficient;  $\Delta_{\max} = \max_\lambda \max_\varepsilon |x_0(\varepsilon) - x_\lambda(\varepsilon)|$  is the maximum difference between the two levels;  $\Delta_{\min} = \min_\lambda \min_\varepsilon |x_0(\varepsilon) - x_\lambda(\varepsilon)|$  is the minimum difference between the two levels.

Then the weighted correlation of the meteorological feature vector between the  $\lambda - th$  sample and the sample to be tested is [16]:

$$\tau_\lambda = \sum_{\varepsilon=1}^k \omega_\varepsilon \mu_\lambda(\varepsilon) \quad (5)$$

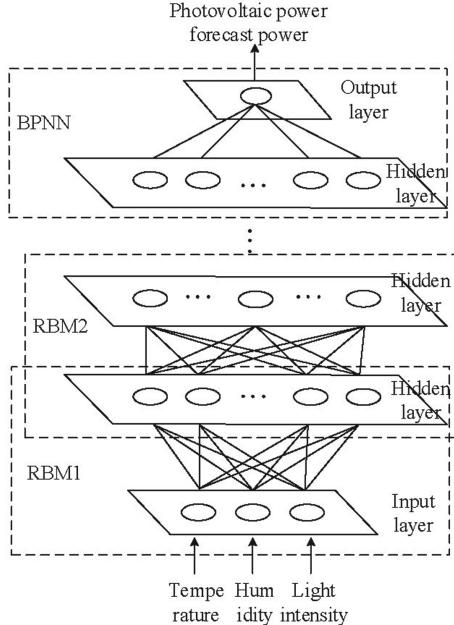
The sixth step is to set the objective function as  $f = 0.5D_\lambda + 0.5\tau_\lambda$  [16], according to the actual situation, set  $f$  greater than a certain value to filter the training sample set.

### 3 Construction of Short-Term Prediction Model of Photovoltaic Power

The multi-layer network structure of DBN can hierarchically characterize the nonlinear relationship between various factors affecting photovoltaic power generation and the photovoltaic power to be predicted, and improve the accuracy of photovoltaic power prediction, so using DBN to establish short-term prediction model of photovoltaic power. In order to accurately determine the relevant parameters of the DBN, an unsupervised training method is used to pre-train the photovoltaic power prediction model to determine the initial solutions of the weight and bias of the photovoltaic power prediction model. Then the supervised global fine-tuning is used to obtain the optimal solution of the weight and bias of the photovoltaic power prediction model. In the fine-tuning process, a dynamically adjusted momentum factor is added to optimize the fine-tuning process.

### 3.1 Structure of Short-Term Prediction Model of Photovoltaic Power Generation Based on DBN

The short-term prediction model of photovoltaic power based on DBN is shown in Fig. 1:



**Fig. 1.** Short term photovoltaic power prediction model based on DBN

The model is composed of multiple Restricted Boltzmann Machines (RBM) and a BP neural network. Each RBM consists of a visible layer and a hidden layer that are bidirectionally connected. The layer acts as the visible layer of the next layer of RBM. The bottom layer RBM, namely the visible layer of RBM1, is used as the input layer of the model. The input layer and multiple hidden layers form a DBN for extracting the features of the sample data; The last hidden layer and the output layer form a neural network, take the extracted feature vectors as input, and generate photovoltaic power prediction results through regression fitting.

The inputs to the model are temperature, humidity, and light intensity vectors, denoted by  $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ , respectively. The output of the model is the predicted power of photovoltaic power generation, denoted by  $\mathbf{y}_p$ . Use the input vector  $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]$  selected in Sect. 2 and the corresponding actual power to form the training sample set  $\{\mathbf{x}, \mathbf{y}\}$  to train the model.

### 3.2 Pre-training of Short-Term Prediction Model Based on DBN

DBN is formed by stacking a series of RBMs, so the RBM training method can be used for layer-by-layer training to obtain the initial solution of the network parameters of the photovoltaic short-term prediction model, Contrastive Divergence (CD) [17, 18] is optimized for the problem of inefficiency of Gibbs sampling method in the face of high-dimensional data, The efficiency of the algorithm is improved, so this paper uses the CD method to pre-train the model layer by layer.

Suppose the number of visible layer and hidden layer neurons are respectively  $n, m$ .  $v_i, h_j$  is the state of the  $i$ th neuron in the visible layer and the  $j$ th neuron in the hidden layer, their corresponding offsets are  $a_i, b_j$ ;  $w_{ij}$  is the connection weight between the  $i$ th neuron in the explicit layer and the  $j$ th neuron in the hidden layer;  $\beta = (w_{ij}, a_i, b_j)$  is the parameter of RBM. Then the energy function of RBM in state  $(\mathbf{v}, \mathbf{h})$  can be expressed as:

$$E(\mathbf{v}, \mathbf{h} | \beta) = - \sum_{i=1}^n a_i v_i - \sum_{j=1}^m b_j h_j - \sum_{i=1}^n \sum_{j=1}^m v_i w_{ij} h_j \quad (6)$$

When the visible layer unit vector  $\mathbf{v} = (v_1, \dots, v_i, \dots, v_n)$  is randomly given, the probability that the  $j$ -th unit  $h_j$  of the hidden layer is activated (assigned a value of 1) is:

$$p(h_j = 1 | \mathbf{v}) = \text{Sigmoid} \left( b_j + \sum_{i=1}^n v_i w_{ij} \right) \quad (7)$$

In Eq. (7),  $\text{Sigmoid}(x) = (1 + e^{-x})^{-1}$  is the activation function.

When the hidden layer unit vector  $\mathbf{h} = (h_1, \dots, h_i, \dots, h_n)$  is randomly given, the probability that the first unit of the explicit layer is activated (assigned a value of 1) is:

$$p(v_i = 1 | \mathbf{h}) = \text{Sigmoid} \left( a_i + \sum_{j=1}^m w_{ij} h_j \right) \quad (8)$$

Contrastive Divergence (CD) is used to solve the initial parameters of the RBM parameter  $\beta = (w_{ij}, a_i, b_j)$ , First select a sample as the initial state of the manifest layer  $\mathbf{v}$ ; then calculate  $\mathbf{h}$  according to Eq. (7), Then calculate the updated state  $\mathbf{v}' = (v'_i)$  of the hidden layer neuron according to formula (8); then calculate the updated  $\mathbf{h}' = (h'_j)$  of the hidden layer neuron according to formula (7), the update formula is formula (9):

$$\begin{aligned} \Delta w_{ij} &= \sigma_{CD} (\langle v_i h_j \rangle - \langle v'_i h'_j \rangle) \\ \Delta a_i &= \sigma_{CD} (\langle v_i \rangle - \langle v'_i \rangle) \\ \Delta b_j &= \sigma_{CD} (\langle h_j \rangle - \langle h'_j \rangle) \end{aligned} \quad (9)$$

In Eq. (9),  $\sigma_{CD}$  is the learning rate of the CD method;  $\langle \cdot \rangle$  is the mathematical expectation of the variable.

Then select another sample as the initial state of the display layer and repeat the above steps until all the samples in the training sample have been selected once.

### 3.3 Optimization of Photovoltaic Power Generation Short-Term Prediction Model Based on Dynamically Adjusted Momentum Factor DBN

After the pre-training is completed, the overall parameters of the photovoltaic power short-term prediction model need to be fine-tuned in reverse to make the model converge to the global best. The traditional DBN fine-tuning stage generally uses the BP neural network algorithm to fine-tune the parameters of the model [19]. At the same time, the extracted feature vectors are used as input to fit the photovoltaic power generation to output the prediction results. However, the traditional gradient descent method used in the process itself has problems such as slow convergence speed, easy oscillation, and easy localization.

The momentum gradient descent method [20] adds a momentum factor (equivalent to adding a damping term), which can reduce the oscillation of the training process. At the same time, the increased momentum factor can make the network ignore the smaller surface features, which can avoid falling into the local minimum [20]. In addition, this method can accelerate the adjustment in the direction of convergence and has a faster running speed. Therefore, this paper introduces the momentum factor to optimize when updating the network weights. However, when the value of the momentum factor is constant, it may slow down the convergence rate to a certain extent, so this article uses the method used in [15] to continuously adjust the momentum factor during the training process. The fine-tuning process of the model is as follows:

- (1) Get the corresponding output according to the input of the model. According to the initial values of the parameters obtained by pre-training, the state of the hidden layer neural unit is judged, and its activation value is calculated. Then it propagates backwards layer by layer, and calculates the activation values of the neural units in the hidden layers of each layer. Finally, calculate the output of the output layer (top layer) [19].
- (2) The error back propagation algorithm is used to recalculate the weight and offset of the DBN network. The cost function of the network is as follows [19](10):

$$\mathbf{E} = \frac{1}{m^{(q)}} \sum_{j=1}^{m^{(q)}} \left[ Y'_j \left( \mathbf{w}^{(q)}, \mathbf{b}^{(q)} \right) - Y_j \right]^2 \quad (10)$$

In Eq. (10),  $m^{(q)}$  is the number of hidden layer neurons in the  $q - th$  RBM;  $\mathbf{E}$  is the error square vector of the network,  $Y'_j, Y_j$  are the actual and ideal output of the first hidden neuron in the output layer (top layer);  $\left( \mathbf{w}^{(q)}, \mathbf{b}^{(q)} \right)$  is the weight and bias vector to be trained in the  $q - th$  RBM.

- (3) The momentum gradient descent method is used to modify the weight and bias of the DBN network. The correction formula is [15]:

$$\Delta(w_{ij}, b_j)(t+1) = \alpha\Delta(w_{ij}, b_j)(t) - \lambda \frac{\partial E}{\partial(w_{ij}, b_j)}(t) \quad (11)$$

$$(w_{ij}, b_j)(t+1) = (w_j, b_j)(t) + \Delta(w_{ij}, b_j)(t+1) \quad (12)$$

In Eq. (11),  $\alpha$  is the momentum factor and is the learning rate. As can be seen from Eq. (11), this method adds  $\alpha\Delta(w_{ij}, b_j)(t)$  terms to the traditional algorithm. Each adjustment to the network weights and offsets will take into account the current network weights and offset adjustments and the last network weight. The result of adjustment. If the amplitude of the last correction is large, the current correction amount will be reduced, otherwise, the current correction amount will be increased. Since this method can always adjust in the direction of convergence and accelerate the adjustment amount in this direction, this method has a faster running speed while avoiding the divergence of the network.

At the same time, compared with the method where the momentum factor is a fixed constant, the momentum factor of the method can be continuously adjusted during the training process to speed up the convergence speed.

### 3.4 Solving Process of Short-Term Prediction Algorithm for Photovoltaic Power

The steps to solve the short-term prediction algorithm of photovoltaic power generation based on momentum factor in deep belief network are as follows:

- (1) Selection and normalization of training samples for prediction models. Select the hourly temperature, humidity, and light intensity of the historical date as the training samples, and use Eqs. (1) to (5) to normalize and filter the samples to obtain the training sample set.
- (2) Formula (6) to formula (9) are used to pre-train the DBN-based photovoltaic short-term prediction model.
- (3) Use formula (10)–formula (12) to optimize the short-term prediction model of photovoltaic power based on DBN after pre-training.
- (4) After the training is completed, the hourly temperature, humidity, light intensity and other relevant data of the day to be tested are input into the trained prediction model to obtain the photovoltaic power prediction value of the day to be tested.

## 4 Example Analysis

In order to fully verify the scientificity and effectiveness of the short-term prediction model of photovoltaic power generation established in this article, the operating system is Windows10 64, the processor is Intel (R) Core (TM) i5-8300 CPU2.30 GHz, and the RAM is 8G platform. In the above, MATLAB R2018a was used to perform the simulation experiment according to the solution flow of the short-term photovoltaic power generation prediction algorithm introduced in Sect. 3.4.

#### 4.1 Data Sources and Evaluation Indicators

The data in this article comes from a photovoltaic power generation system in the Australian Desert Solar Research Center [21]. The data is collected every 5 min, and the hourly data is averaged and converted into 1 data point per hour. The research period is 08: 00–18: 00 per day, that is, 11 data points per day. The historical data of photovoltaic power and meteorological factors in the region from March 2017 to May 2017 and March 2018 to May 2018 for the same season are selected as the original data. The weather is stable on May 29, 2018 and non-stationary May 2, 2015 as the days to be tested. The measured data on the day to be tested is shown in Table 2 below (rounded to two decimal places):

**Table 2.** Measured data of photovoltaic power generation on the day to be tested

2018-5-2 time/h	Actual data/kW	2018-5-29 time/h	Actual data/kW
8	4.64	8	2.396
9	19.01	9	21.04
10	41.16	10	42.87
11	58.06	11	59.38
12	72.06	12	67.83
13	73.97	13	70.33
14	68.86	14	66.55
15	50.35	15	57.19
16	44.34	16	42.07
17	25.07	17	22.23
18	5.56	18	4.55

This article uses the Mean Absolute Percentage Error (MAPE) indicator [2] commonly used in photovoltaic power generation prediction to evaluate the prediction model, The smaller the  $e_{MAPE}$ , the higher the prediction accuracy of the model, and its definition is as in formula (13):

$$e_{MAPE} = \frac{1}{11} \sum_{i=1}^{11} \frac{|y_i - y_{i0}|}{y_{i0}} \times 100\% \quad (13)$$

In Eq. (13),  $y_i$  is the predicted power value;  $y_{i0}$  is the measured power value.

#### 4.2 Model Structure and Parameter Settings

Select the training samples from the 180d data as the training set according to the method in Sect. 1, and use the data of the two test days in Table 2 as the test set. Perform the training test on the model according to the method described in Sect. 2, and obtain the DBN The prediction accuracy for different hidden layers

and different units is shown in Table 3 below (retain two decimal places). It can be seen from Table 3 that when the number of hidden layers is 2 and the number of units is 10 and 20 respectively, the minimum  $e_{MAPE}$  of the model is 14.22, that is, the optimal parameters at this time.

**Table 3.** Prediction performance comparison of DBN prediction models with different structures

Number of hidden layers	Number of hidden layer units	$e_{MAPE}\%$
1	10	14.45
	15	17.45
	20	21.56
	25	21.66
2	10	16.59
	15	19.83
	20	14.22
	25	27.47
3	10	44.29
	15	32.54
	20	34.26
	25	31.55

The relevant parameters of DBN are shown in Table 4 below:

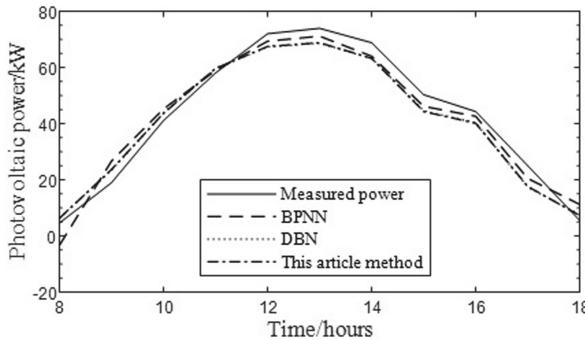
**Table 4.** Relevant parameter settings

Parameter name	Numerical value
Learning rate	0.01
Momentum	0.5
Batch size	10
Maximum number of pre-training iterations	250
Maximum number of iterations during fine-tuning	250
Error function target value	0.001

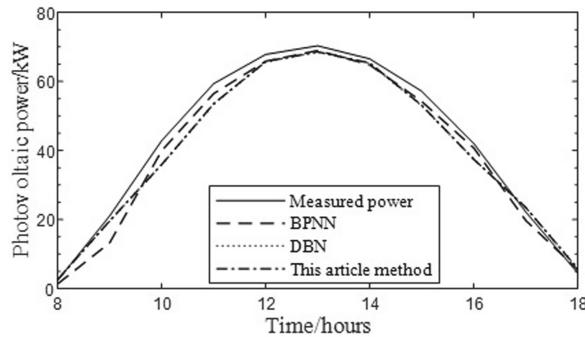
### 4.3 Calculation Result Analysis

Taking the measured data in Table 2 as an example, the model proposed in this paper, the BPNN photovoltaic power generation prediction model, and the traditional DBN photovoltaic power generation short-term prediction model are

used for prediction analysis. The parameter settings of the traditional DBN are consistent with this article except for momentum. The BPNN is set to a single hidden layer, the number of hidden units is 20, the number of iterations is 250, the learning rate is 0.01, and the target accuracy is 0.001. The prediction results are shown in Figs. 2 and 3 below.



**Fig. 2.** Forecast results on May 2, 2018



**Fig. 3.** Forecast results on May 29, 2018

Substituting the predicted and measured values of the three methods into Eq. (13), the  $e_{MAPE}$  of the three models is obtained as shown in Table 5 below:

Comparing Figs. 2, 3 and Table 5, we can see that the prediction results of May 29, 2018 are significantly better than other days to be tested. This is because the weather on May 29, 2018 is more ideal, and the trend of photovoltaic power generation is also more stable. The power curve fluctuated on May 2, 2018, and the prediction accuracy of several methods also decreased to a large extent. This is because the data collection area belongs to the desert area, there are more

**Table 5.** Prediction performance comparison of DBN prediction models with different structures

Date	Model	$e_{MAPE}\%$
2018-5-2	BPNN	33.50
	BPNN	16.34
	This article method	15.95
	25	21.66
2018-5-29	BPNN	12.54
	DBN	9.71
	This article method	9.07

sunny days, and fewer samples are not sunny days, and Under the influence of cloudiness and sand and dust in non-stationary weather, the actual power of photovoltaic power generation is greatly deviated from the predicted power.

As can be seen from Table 5, the  $e_{MAPE}$  value of this method is reduced by 0.64 and 3.47 on May 29, 2018 compared to the traditional DBN method and BPNN method; on May 2, 2018, when the weather condition is abrupt, this method is compared. The  $e_{MAPE}$  values of the traditional DBN method and BPNN method are reduced by 0.39 and 17.55, respectively, indicating that the method in this paper effectively improves the short-term prediction accuracy of photovoltaic power generation.

## 5 Conclusion

- (1) Propose a short-term prediction model of photovoltaic power generation based on a deep belief network of momentum factors. The model uses DBN's multi-layer network structure to hierarchically characterize the complex non-linear relationship between factors such as temperature, humidity, and light intensity vectors that affect photovoltaic power and the power to be measured. By introducing a dynamically adjusted momentum factor to optimize the model, the problems of slow convergence, easy oscillation and falling into local optimality of the gradient descent method are solved.
- (2) Using the measured data of the Australian Desert Solar Research Center, the method proposed in this paper is compared with the traditional DBN method and BPNN method to verify the simulation experiments. The short-term power prediction model has high prediction accuracy.

**Funding.** This research was funded by the National Natural Science Foundation of China (U1804143). Technology Developing Program of Henan Electric Power Company5217S0200001. The Key Scientific Research projects of Henan Higher Education Institutions20B470008.

## References

1. Ding, M., Wang, W.S., Wang, X.L., Song, Y.T., Chen, D.Z., Sun, M.: A review on the effect of large-scale PV generation on power systems. *J. Proc. CSEE* **34**(1), 1–14 (2014). <https://doi.org/10.13334/j.0258-8013.pcsee.2014.01.001>
2. Gong, Y.F., Lu, Z.X., Qiao, Y., Wang, Q.: An overview of photovoltaic energy system output forecasting technology. *J. Autom. Electric. Pow. Syst.* **40**(4), 140–151 (2016). <https://doi.org/10.7500/AEPS20150711003>
3. Zhu, X., Ju, R.R., Cheng, X., Ding, Y.Y., Zhou, H.: A very short-term prediction model for photovoltaic power based on numerical weather prediction and ground-based cloud images. *J. Autom. Electr. Pow. Syst.* (6) (2015). <https://doi.org/10.7500/AEPS20140409004>
4. Li, Y., Si, Y., Shu, L.: An ARMAX model for forecasting the power output of a grid connected photovoltaic system. *J. Renew. Energy* **66**, 78–89 (2014). <https://doi.org/10.1016/j.renene.2013.11.067>
5. Wang, G.C., Su, Y., Shu, L.J.: One-day-ahead daily power forecasting of photovoltaic systems based on partial functional linear regression models. *J. Renew. Energy* **96**, 469–478 (2016). <https://doi.org/10.1016/j.renene.2016.04.089>
6. Li, F., Song, Q.J., Cai, T., Zhao, J.B., Yan, Q.Q., Chen, Z.H.: Based on principal component analysis and the BP neural network in the application of grid-connected photovoltaic power energy prediction. *J. Renew. Energy Resour.* **05**, 61–67 (2017). <https://doi.org/10.3969/j.issn.1671-5292.2017.05.009>
7. Zhao, W.Q., Guo, B.X., Li, G., Li, Z. : Output power forecast of PV plant based on Elman neural network optimized by intelligent water drop algorithm. *J. Acta Energiae Solaris Sin.* **38**(006), 1553–1559 (2017)
8. Ye, L., Chen, Z., Zhao, Y.N., Zhu, Q.W.: Photovoltaic power forecasting model based on genetic algorithm and fuzzy radial basis function neural network. *J. Autom. Electr. Pow. Syst.* **39**(16), 16–22 (2015). <https://doi.org/10.7500/AEPS20140903004>
9. Eseye, A.T., Zhang, J., Zheng, D.: Short-term photovoltaic solar power forecasting using a hybrid Wavelet-PSO-SVM model based on SCADA and Meteorological information. *J. Renew. energy* **118**, 357–367 (2018). <https://doi.org/10.1016/j.renene.2017.11.011>
10. Zhang, Y.J., Yang, L.F., Ge, S.Y., Zhou, H.X.: Short-term photovoltaic power forecasting based on K-means algorithm and support vector machine. *J. Pow. Syst. Protect. Control* **46**(21), 118–124 (2018). <https://doi.org/10.7667/PSPC171595>
11. Yang, X.Y., Liu, H., Zhang, B., Chen, S.: A combination method for photovoltaic power forecasting based on entropy weight method. *J. Acta Energiae Solaris Sinica* **5**, 744–749 (2014). <https://doi.org/10.3969/j.issn.0254-0096.2014.05.002>
12. Gao, Y., Zhang, B.L., Mao, J.I., Liu, Y.: Machine learning-based adaptive very-short-term forecast model for photovoltaic power. *J. Pow. Syst. Technol.* **39**(2), 307–311 (2015). <https://doi.org/10.13335/j.1000-3673.pst.2015.02.002>
13. Wang, X.P., Zhou, X.L., Xin, J., Yang, J.: A prediction method of PV output power based on the combination of improved grey back propagation neural network. *J. Pow. Syst. Protect. Control* **44**(18), 81–87 (2016). <https://doi.org/10.7667/PSPC151675>
14. Hinton, G.E., Osindero, S., Yee-Whye, T.: A fast learning algorithm for deep belief nets. *J. Neural Comput.* **18**(7), 1527–1554 (2006). <https://doi.org/10.1162/neco.2006.18.7.1527>

15. Shao, H., Zheng, G.: Convergence analysis of a back-propagation algorithm with adaptive momentum. *J. Neurocomput.* **74**(5), 749–752 (2011). <https://doi.org/10.1016/j.neucom.2010.10.008>
16. Ge, L., Lu, W.W., Yuan, X.D., Zhou, Q.: Power forecasting of photovoltaic plant based on improved similar dayand ABC-SVM. *J. Acta Energiae Solaris Sinica* **3**, 28 (2018)
17. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. *J. Neural Comput.* **14**(8), 1771–1800 (2002). <https://doi.org/10.1162/089976602760128018>
18. Hinton, G.E.: A practical guide to training restricted Boltzmann machines. *J. Momentum* **9**(1), 926–947 (2010). [https://doi.org/10.1007/978-3-642-35289-8\\_32](https://doi.org/10.1007/978-3-642-35289-8_32)
19. Liang, R., Yang, B., Ma, R.Z., Wu, J., Wu, K.H., Lin, Z.Z., Wen, F.S. : Spatial electric load forecasting for distribution systems using multi-source information and deep belief network-deep neural network. *J. Electr. Power Constr.* **39**(10) (2018). <https://doi.org/10.3969/j.issn.1000-7229.2018.10.002>
20. Tian, Y.: A forecasting model for current load of neural network group based upon momentum factor. *J. Pow. Syst. Protect. Control* **44**(17), 31–38 (2016). <https://doi.org/10.7667/PSPC161099>
21. <http://dkasolarcentre.com.au/historical-data/download>



# Improved Adaptive Filter with Unknown Process and Measurement Noise Covariance

Jirong Ma<sup>1</sup>(✉), Yumei Hu<sup>2,3</sup>, Qinghua Ma<sup>1</sup>, Shujun Yang<sup>1</sup>, Jianqiang Zheng<sup>1</sup>, and Shuaiwei Wang<sup>1</sup>

<sup>1</sup> Xi'an Modern Control Technology Research Institute, Xi'an, China  
mjr994@sina.com

<sup>2</sup> School of Automation, Northwestern Polytechnical University, Xi'an, China

<sup>3</sup> Key Laboratory of Information Fusion Technology, Ministry of Education, Xi'an, China

**Abstract.** This paper considers the problem of state estimation with unknown process and measurement noise covariance for a linear Gaussian system. Under the assumption of conjugate priors, inverse Wishart distribution is chosen for both process and measurement noise covariance by introducing a latent variable. Then, the joint state estimation and unknown parameters identification is derived in variational Bayesian framework, and the system state, latent variable, process and measurement noise covariance are updated iteratively. The performance of the proposed algorithm is demonstrated by comparing with the other VB based adaptive filter in a target tracking simulation system.

**Keywords:** Adaptive Kalman filter · Variational Bayesian · Unknown process and measurement noise covariance · Conjugate priors

## 1 Introduction

State estimation involved state-space models are widely used in many applications, such as navigation, control and surveillance. With full knowledge of model parameters, Kalman filter (KF) is the optimal filter for linear Gaussian systems. However, our knowledge about the parameters are usually inaccurate or unknown, especially the noise statistics like covariance, which demands the ability for matching model parameters [1]. The classic method to solve this problem are adaptive filters, which can achieve a joint estimation of system state and parameters, and mainly including robust filtering, multiple model (MM) filtering, Monte Carlo (MC) and variational Bayesian (VB) method. The purpose of robust filtering is to minimize the gain of transfer matrix from parameters to estimation error, so as to achieve a compromise between state estimation accuracy and robustness. However, these parameters are often assumed to be constant with constraints, which makes robust filter relatively conservative. For example,

the estimation accuracy of time-varying disturbance parameters is often unsatisfactory, even can not be identified. In MM filtering, parameters are modeled by a random switching Markov chain. Li *et al.* [2] gave a review of MM methods and analyzes their advantages in maneuvering target tracking system. Nevertheless, the estimation ability of MM methods is limited to the available “model dictionary” and computational cost increases rapidly with the number of models. In terms of MC method, Markov Chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) are used in [3] to solve the problem of nonlinear state space parameter estimation. Schön *et al.* [4] reviewed the application of SMC methods and specific implementations (such as metropolis Hastings sampling, maximum likelihood and Gibbs sampling) in system identification. But MC methods are also faces with the problem of computational cost.

In recent years, an iterative joint optimization method with unified framework has become the mainstream in solving the problems of system identification and state estimation, including expectation maximization (EM) and VB [5]. EM realizes the joint optimization by establishing a feedback loop which is divided into E-step and M-step. Compared with EM algorithm, VB is more suitable for dealing with complex inference problems by means of complex graph model, especially for dealing with unknown multi-parameter problems. Särkka *et al.* [6, 7] present two VB based recursive adaptive KF for linear and nonlinear state space model, inverse Gamma distribution and inverse Wishart distribution are chosen as the conjugate prior for unknown diagonal measurement noise covariance matrix(each element) and general measurement noise covariance matrix, and thus solve the problem of state estimation with unknown measurement noise in a closed form manner. Related work has many applications in target tracking and information fusion. Gao *et al.* [8] extended the VB based adaptive KF to centralized multi-sensor fusion, and proposed an augment strategy and a sequential strategy for centralized fusion. Li *et al.* [9] consider the unknown measurement noise problem in maneuvering target tracking situation, and proposed a novel VB-IMM algorithm. Hosseini *et al.* [10] combined the VB approach with Rao-Blackwellized Monte Carlo data association method to deal with the multiple target tracking problem with unknown observation noise variance.

Although the research on VB based adaptive filter for unknown measurement noise is readily available, extension algorithm for unknown process noise is not straightforward since conjugate prior for process noise covariance is not easy to find[7]. Since process noise doesn't affect the performance of filter directly, Huang *et al.* [11] estimated predict error covariance instead, and proposed a novel joint recursive adaptive filter. Ma *et al.* [12] chose an inverse Wishart distribution as the conjugate prior model for process noise covariance by introducing a latent variable, and update the joint posterior probability density function(PDF) iteratively. Ardestiri *et al.* [13] considered the adaptive smoother, which only need the prior model for initial state and initial noise covariance, and presented a VB based smoothing algorithm for joint estimation of dynamic system state, process and measurement noise covariance.

In this paper, the work in [12] is extended to both unknown process and measurement noise covariance situation, and an improved adaptive KF filter based on VB (VBAKF-QR) is developed. The approximate posterior PDFs are derived in VB framework and updated iteratively. The performance of VBAKF-QR is compared with the method proposed in [11] by simulation.

## 2 Problem Formulation

Consider the following discrete linear Gaussian system

$$x_k = F_k x_{k-1} + w_k, \quad (1)$$

$$y_k = H_k x_k + v_k, \quad (2)$$

where  $x_k$  is the state vector with the dimension of  $n_x$  at time  $k$ ,  $y_k$  is the sensor measurement vector with the dimension of  $n_y$ .  $F_k$  and  $H_k$  are the known state transition matrix and measurement matrix. The process noise  $w_k$  and the measurement noise  $v_k$  are zero mean Gaussian distribution with unknown covariance matrix  $Q_k$  and  $R_k$  respectively.

Solving the above problem in Bayesian framework is equal to compute the joint PDF  $p(x_k, Q_k, R_k | y_{1:k})$ , which requires the following two steps,

- Predict step:

$$p(x_k, Q_k, R_k | y_{1:k-1}) = p(x_k | y_{1:k-1}, Q_k, R_k) p(Q_k | y_{1:k-1}) p(R_k | y_{1:k-1}). \quad (3)$$

- Update step:

$$p(x_k, Q_k, R_k | y_{1:k}) \propto p(y_k | x_k, Q_k, R_k) p(x_k, Q_k, R_k | y_{1:k-1}). \quad (4)$$

In Kalman filter framework, we have,

$$p(x_k | y_{1:k-1}, Q_k, R_k) = p(x_k | y_{1:k-1}, Q_k) = \mathcal{N}(x_k | x_{k|k-1}, P_{k|k-1}), \quad (5)$$

$$p(y_k | x_k, Q_k, R_k) = p(y_k | x_k, R_k) = \mathcal{N}(y_k | H_k x_k, R_k), \quad (6)$$

where

$$x_{k|k-1} = F_k x_{k-1|k-1}, \quad (7)$$

$$P_{k|k-1} = F_k P_{k-1|k-1} F_k^T + Q_k. \quad (8)$$

As  $Q_k$  and  $R_k$  are unknown, the predict step in Eq. (3) requires intractable integration. VB, which can approximate an intractable PDF with a tractable PDF, are introduced for state estimation with unknown process and measurement noise covariance.

In order to get a closed-form analytical solution, conjugate prior models are chosen for  $Q_k$  and  $R_k$  to guarantee the same functional form of the posterior distribution and the prior distribution, which can make it easy calculated. Based on Eq. (5), the likelihood PDF of  $Q_k$  and  $R_k$  are both Gaussian distribution,

the difference is that  $Q_k$  is only a component of the covariance  $P_{k|k-1}$  and  $R_k$  is the covariance. A latent variable  $m_k$  is introduced to decompose  $P_{k|k-1}$ ,

$$\mathcal{N}(x_k|x_{k|k-1}, P_{k|k-1}) = \int \mathcal{N}(x_k|m_k, Q_k) \mathcal{N}(m_k|x_{k|k-1}, P_{m,k}) dm_k, \quad (9)$$

where  $P_{m,k} = F_k P_{k-1|k-1} F_k^T$ .

Since inverse Wishart distribution is the conjugate prior for the covariance of a Gaussian distribution [12], the prior model of  $Q_k$  and  $R_k$  are chosen as,

$$p(Q_k|y_{1:k-1}) = \text{IW}(Q_k|t_{q,k|k-1}, T_{q,k|k-1}), \quad (10)$$

$$p(R_k|y_{1:k-1}) = \text{IW}(R_k|t_{r,k|k-1}, T_{r,k|k-1}). \quad (11)$$

The initial process and measurement noise covariance  $Q_0$  and  $R_0$  are also chosen to follow the inverse Wishart distribution with the mean  $\bar{Q}_0$  and  $\bar{R}_0$ , which are called as initial nominal process and measurement noise covariance. The initial parameter  $t_{q,k|k-1}$ ,  $T_{q,k|k-1}$ ,  $t_{r,k|k-1}$ , and  $T_{r,k|k-1}$ , the dynamic model  $p(Q_k|Q_{k-1})$  and  $p(R_k|R_{k-1})$  are chosen Similar to [7, 11].

### 3 Solutions by Variational Bayesian

To solve the problem of state estimation with unknown process and measurement noise covariance, we need to calculate the joint posterior PDF  $p(x_k, Q_k, R_k|y_{1:k})$ . It turns to calculate  $p(x_k, m_k, Q_k, R_k|y_{1:k})$  since a latent variable  $m_k$  is introduced in Sect. 2. By using mean-field approximation [5], the posterior PDF can be approximated as,

$$p(x_k, m_k, Q_k, R_k|y_{1:k}) \approx q(x_k, m_k, Q_k, R_k) \approx q(x_k)q(m_k)q(Q_k)q(R_k). \quad (12)$$

Kullback-leibler (KL) divergence, which has non-negative property, is used to measure the difference between these two PDFs,

$$\begin{aligned} & \text{KL}(q(x_k, m_k, Q_k, R_k)||p(x_k, m_k, Q_k, R_k|y_{1:k})) \\ &= \mathbb{E} \left[ \log \frac{q(x_k, m_k, Q_k, R_k)}{p(x_k, m_k, Q_k, R_k|y_{1:k})} \right] \\ &= \mathbb{E} \left[ \log \frac{q(x_k, m_k, Q_k, R_k)p(y_k)}{p(x_k, m_k, Q_k, R_k, y_k|y_{1:k-1})} \right] \\ &= -\underbrace{\mathbb{E} \left[ \log \frac{p(x_k, m_k, Q_k, R_k, y_k|y_{1:k-1})}{q(x_k, m_k, Q_k, R_k)} \right]}_{\mathcal{B}(q)} + \log p(y_k), \end{aligned} \quad (13)$$

where the exception is taken on  $q(x_k, m_k, Q_k, R_k)$ .  $\mathcal{B}(q)$  is called evidence lower bound (ELBO). The aim of VB is to maximize  $\mathcal{B}(q)$  instead of minimizing the

KL divergence. By using coordinate ascent method, we have

$$q(x_k) \propto \exp \left\{ \mathbb{E}_{q(m_k, Q_k, R_k)} [\log p(x_k, m_k, Q_k, R_k, y_k | y_{1:k-1})] \right\}, \quad (14)$$

$$q(m_k) \propto \exp \left\{ \mathbb{E}_{q(x_k, Q_k, R_k)} [\log p(x_k, m_k, Q_k, R_k, y_k | y_{1:k-1})] \right\}, \quad (15)$$

$$q(Q_k) \propto \exp \left\{ \mathbb{E}_{q(x_k, m_k, R_k)} [\log p(x_k, m_k, Q_k, R_k, y_k | y_{1:k-1})] \right\}, \quad (16)$$

$$q(R_k) \propto \exp \left\{ \mathbb{E}_{q(x_k, m_k, Q_k)} [\log p(x_k, m_k, Q_k, R_k, y_k | y_{1:k-1})] \right\}. \quad (17)$$

### 3.1 Derivation of the Posterior $q(x_k)$

Based on the prior model in Sect. 2, the approximate PDF is derived as

$$\begin{aligned} & \exp \left\{ \mathbb{E}_{q(m_k, Q_k, R_k)} [\log p(x_k, m_k, Q_k, R_k, y_k | y_{1:k-1})] \right\} \\ & \propto \exp \left\{ \mathbb{E}_{q(m_k, Q_k, R_k)} [\log \mathcal{N}(x_k | m_k, Q_k) \mathcal{N}(y_k | H_k x_k, R_k)] \right\} \\ & \propto \exp \left\{ -\frac{1}{2} \text{tr} (\mathbb{E}[(x_k - m_k)^T Q_k^{-1} (x_k - m_k)]) \right\} \\ & \quad \times \exp \left\{ -\frac{1}{2} \text{tr} (\mathbb{E}[(y_k - H_k x_k)^T R_k^{-1} (y_k - H_k x_k)]) \right\} \\ & \propto \exp \left\{ -\frac{1}{2} \text{tr} ((x_k - \mathbb{E}[m_k])^T \mathbb{E}[Q_k^{-1}] (x_k - \mathbb{E}[m_k])) \right\} \\ & \quad \times \exp \left\{ -\frac{1}{2} \text{tr} ((y_k - H_k x_k)^T \mathbb{E}[R_k^{-1}] (y_k - H_k x_k)) \right\} \\ & \propto \mathcal{N}(x_k | \mathbb{E}[m_k], \{\mathbb{E}[Q_k^{-1}]\}^{-1}) \mathcal{N}(y_k | H_k x_k, \{\mathbb{E}[R_k^{-1}]\}^{-1}). \end{aligned} \quad (18)$$

Since the product of two Gaussian distribution is also Gaussian, assume the posterior mean is  $x_{k|k}$  and the covariance is  $P_{k|k}$ . Let  $A_k = \{\mathbb{E}[Q_k^{-1}]\}^{-1}$ ,  $B_k = \{\mathbb{E}[R_k^{-1}]\}^{-1}$ , it can be calculated in KF framework as follows

$$K_{x,k} = A_k H_k^T (H_k A_k H_k^T + B_k)^{-1}, \quad (19)$$

$$x_{k|k} = \mathbb{E}[m_k] + K_{x,k} (y_k - H_k \mathbb{E}[m_k]), \quad (20)$$

$$P_{k|k} = A_k - K_{x,k} H_k A_k. \quad (21)$$

### 3.2 Derivation of the Posterior $q(m_k)$

Based on the prior model in Sect. 2, the approximate PDF is derived as

$$\begin{aligned} & \exp \left\{ \mathbb{E}_{q(x_k, Q_k, R_k)} [\log p(x_k, m_k, Q_k, R_k, y_k | y_{1:k-1})] \right\} \\ & \propto \exp \left\{ \mathbb{E}_{q(x_k, Q_k, R_k)} [\log \mathcal{N}(x_k | m_k, Q_k) \mathcal{N}(m_k | x_{k|k-1}, P_{m,k})] \right\} \\ & \propto \exp \left\{ -\frac{1}{2} \text{tr} (\mathbb{E}[(x_k - m_k)^T Q_k^{-1} (x_k - m_k)]) \right\} \mathcal{N}(m_k | x_{k|k-1}, P_{m,k}) \\ & \propto \exp \left\{ -\frac{1}{2} \text{tr} (\mathbb{E}[Q_k^{-1}] (\mathbb{E}[x_k] - m_k) (\mathbb{E}[x_k] - m_k)^T) \right\} \mathcal{N}(m_k | x_{k|k-1}, P_{m,k}) \\ & \propto \mathcal{N}(\mathbb{E}[x_k] | m_k, A_k) \mathcal{N}(m_k | x_{k|k-1}, P_{m,k}). \end{aligned} \quad (22)$$

The same as  $x_{k|k}$  and  $P_{k|k}$ , the posterior mean  $m_k$  and covariance  $P_{m,k|k}$  can be calculated as:

$$K_{m,k} = P_{m,k} (A_k + P_{m,k})^{-1}, \quad (23)$$

$$m_{k|k} = x_{k|k-1} + K_{m,k}(x_{k|k} - x_{k|k-1}), \quad (24)$$

$$P_{m,k|k} = P_{m,k} - K_{m,k} P_{m,k}. \quad (25)$$

### 3.3 Derivation of the Posterior $q(Q_k)$

The form of the inverse Wishart distribution used in this paper is given in [14]. Thus, the approximate posterior PDF is derived as

$$\begin{aligned} & \exp \left\{ \mathbb{E}_{q(x_k, m_k, R_k)} [\log p(x_k, m_k, Q_k, R_k, y_k | y_{1:k-1})] \right\} \\ & \propto \exp \left\{ \mathbb{E}_{q(x_k, m_k)} [\log \mathcal{N}(x_k | m_k, Q_k) \text{IW}(Q_k | t_{q,k|k-1}, T_{q,k|k-1})] \right\} \\ & \propto |Q_k|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} (\mathbb{E}[(x_k - m_k)^T] Q_k^{-1} (x_k - m_k)) \right\} \\ & \quad \times |Q_k|^{-(t_{q,k|k-1} + n_x + 1)/2} \exp \left\{ -\frac{1}{2} \text{tr} (T_{q,k|k-1} Q_k^{-1}) \right\} \\ & \propto |Q_k|^{-(t_{q,k|k-1} + n_x + 2)/2} \exp \left\{ -\frac{1}{2} \text{tr} ((T_{q,k|k-1} + C_k) Q_k^{-1}) \right\}, \end{aligned} \quad (26)$$

where

$$\begin{aligned} C_k &= \mathbb{E}[(x_k - m_k)(x_k - m_k)^T] \\ &= (\mathbb{E}[x_k] - \mathbb{E}[m_k])(\mathbb{E}[x_k] - \mathbb{E}[m_k])^T + \text{cov}(x_k x_k^T) + \text{cov}(m_k m_k^T) \\ &= (x_{k|k} - m_{k|k})(x_{k|k} - m_{k|k})^T + P_{k|k} + P_{m,k|k}. \end{aligned} \quad (27)$$

Thus, the posterior of  $q(Q_k)$  also follows inverse Wishart distribution, and can be updated as

$$q(Q_k) \propto \text{IW}(Q_k | t_{q,k|k}, T_{q,k|k}), \quad (28)$$

where

$$\begin{aligned} t_{q,k|k} &= t_{q,k|k-1} + 1, \\ T_{q,k|k} &= T_{q,k|k-1} + C_k. \end{aligned} \quad (29)$$

Based on the property of inverse Wishart distribution,  $Q_k^{-1}$  follows a Wishart distribution, and the mean value is [15]

$$\mathbb{E}[Q_k^{-1}] = t_{q,k|k} T_{q,k|k}^{-1}. \quad (30)$$

### 3.4 Derivation of the Posterior $q(R_k)$

In a similar way, the approximate posterior PDF is derived as

$$\begin{aligned} & \exp \left\{ \mathbb{E}_{q(x_k, m_k, Q_k)} [\log p(x_k, m_k, Q_k, R_k, y_k | y_{1:k-1})] \right\} \\ & \propto \exp \left\{ \mathbb{E}_{q(x_k)} [\log \mathcal{N}(y_k | x_k, R_k) \text{IW}(R_k | t_{r,k|k-1}, T_{r,k|k-1})] \right\} \\ & \propto |R_k|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} (\mathbb{E}[(y_k - H_k x_k)^T] R_k^{-1} (y_k - H_k x_k)) \right\} \\ & \quad \times |R_k|^{-(t_{r,k|k-1} + n_x + 1)/2} \exp \left\{ -\frac{1}{2} \text{tr} (T_{r,k|k-1} R_k^{-1}) \right\} \\ & \propto |R_k|^{-(t_{r,k|k-1} + n_x + 2)/2} \exp \left\{ -\frac{1}{2} \text{tr} ((T_{r,k|k-1} + D_k) R_k^{-1}) \right\}, \end{aligned} \quad (31)$$

where

$$\begin{aligned} D_k &= \mathbb{E}[(y_k - H_k x_k)(y_k - H_k x_k)^T] \\ &= (y_k - \mathbb{E}[x_k])(y_k - \mathbb{E}[x_k])^T + H_k \text{cov}(x_k x_k^T) H_k^T \\ &= (y_k - x_{k|k})(y_k - x_{k|k})^T + H_k P_{k|k} H_k^T. \end{aligned} \quad (32)$$

Thus, the posterior of  $q(R_k)$  also follows inverse Wishart distribution,

$$q(R_k) \propto \text{IW}(R_k | u_{r,k|k}, U_{r,k|k}), \quad (33)$$

where

$$\begin{aligned} u_{r,k|k} &= t_{r,k|k-1} + 1, \\ U_{r,k|k} &= T_{r,k|k-1} + D_k. \end{aligned} \quad (34)$$

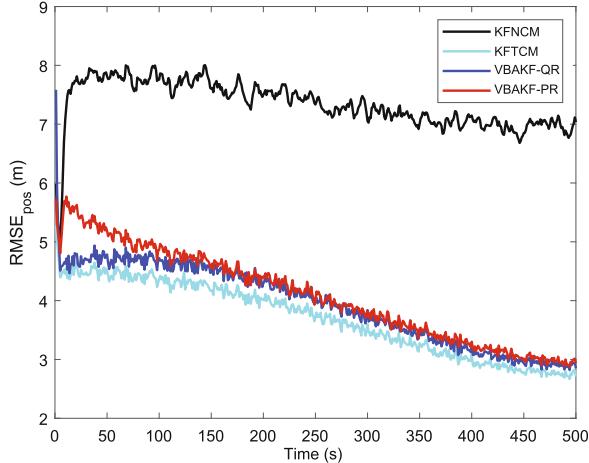
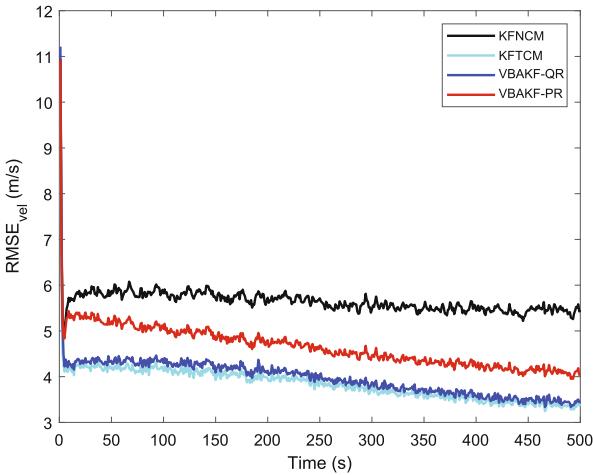
The same as  $Q_k^{-1}$ , the inverse  $R_k^{-1}$  follows a Wishart distribution with

$$\mathbb{E}[R_k^{-1}] = t_{r,k|k} T_{r,k|k}^{-1}. \quad (35)$$

## 4 Simulations

Consider a 2D target tracking simulation scenario, kinematic parameters and algorithm parameters are same in [11]. The proposed VBAKF-QR algorithm is compared with other three algorithms, including the VB based adaptive KF proposed in [11], and we call it VBAKF-PR, the KF with a nominal process and measurement noise covariance (KFNCM), and the KF with true process and measurement noise covariance (KFTCM). The nominal process and measurement noise covariance are selected as  $\tilde{Q}_k = \alpha \mathbf{I}_4$  and  $\tilde{R}_k = \beta \mathbf{I}_2$ , and take  $\alpha = 1$ ,  $\beta = 50$ . The root mean square error (RMSE) is used to evaluate the algorithm performance, and defined in [12].

The RMSE results in position and velocity of 1000 Monte carlo runs are shown in Fig. 1 and Fig. 2. On the whole, VBAKF-QR and VBAKF-PR get

**Fig. 1.** RMSE in position**Fig. 2.** RMSE in velocity

better estimation accuracy than KFNCM, which directly use the nominal process and measurement noise covariance as the filter parameters. For target position, the performance of the proposed VBAKF-QR filter is better than VBAKF-PQ in first 100s, and finally get the same estimation accuracy as VBAKF-PR. For target velocity, VBAKF-QR gets a better performance than VBAKF-PR, and can nearly get the same performance as KFTCM. At the same time, VBAKF-QR converge much more quickly than VBAKF-PR. This is because, the VBAKF-QR estimates the unknown process and measurement noise covariance directly.

## 5 Conclusion

This paper studied the joint estimation of system state, unknown process and measurement noise covariance, and provided a derivation of VB based adaptive KF. Inverse Wishart distribution is the conjugate prior for both process and measurement noise covariance, and provided an intermediate latent variable. In order to maximize the ELBO, the joint posterior PDFs for these variables can be updated iteratively. Simulation results show that the proposed VBAKF-QR method outperforms the other VB based adaptive KF.

## References

1. Mehra, R.: Approaches to adaptive filtering. *IEEE Trans. Autom. Control* **17**(5), 693–698 (1972)
2. Li, X.R., Jilkov, V.P.: Survey of maneuvering target tracking. Part V. multiple-model methods. *IEEE Trans. Aerosp. Electron. Syst.* **41**(4), 1255–1321 (2005)
3. Schon, T.B., Wills, A., Ninnness, B.: System identification of nonlinear state-space models. *Automatica* **47**(1), 39–49 (2011)
4. Schon, T.B., Dahlin, J., Lindsten, F.: Sequential Monte Carlo methods for system identification. *IFAC Papers On Line* **48**(28), 775–786 (2015)
5. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* **112**(518), 859–877 (2017)
6. Särkka, S., Nummenmaa, A.: Recursive noise adaptive Kalman filtering by variational Bayesian approximations. *IEEE Trans. Autom. Control* **54**(3), 596–600 (2009)
7. Särkka, S., Hartikainen, J.: Non-linear noise adaptive Kalman filtering via variational Bayes. In: Proceedings of 2013 IEEE International Workshop on Machine Learning for Signal Processing, pp. 1–6 (2013)
8. Gao, X., Chen, J., Tao, D., Li, X.: Multi-sensor centralized fusion without measurement noise covariance by variational Bayesian approximation. *IEEE Trans. Aerosp. Electron. Syst.* **47**(1), 718–272 (2011)
9. Li, W., Jia, Y.: State estimation for jump Markov linear systems by variational Bayesian approximation. *IET Control Theory Appl.* **6**(2), 319–326 (2012)
10. Sadat Hosseini, S., Jamali, M.M., Särkka, S.: Variational Bayesian adaptation of noise covariances in multiple target tracking problems. *Measurement* S0263224118301544 (2018)
11. Huang, Y., Zhang, Y., Wu, Z., Li, N., Chambers, J.: A novel adaptive Kalman filter with inaccurate process and measurement noise covariance matrices. *IEEE Trans. Autom. Control* **63**(2), 594–601 (2018)
12. Ma, J., Lan, H., Wang, Z., Wang, X., Moran, B.: Improved adaptive Kalman filter with unknown process noise covariance. In: Proceedings of 2018 International Conference on Information Fusion (2018)
13. Ardeshiri, T., Özkan, E., Orguner, U., Gustafsson, F.: Approximate Bayesian smoothing with unknown process and measurement noise covariances. *IEEE Signal Process. Lett.* **22**(12), 2450–2454 (2015)
14. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: *Bayesian Data Analysis*, vol. 2. CRC Press, Boca Raton (2014)
15. Mardia, K.V., Kent, J.T., Bibby, J.M.: *Bayesian Data Analysis*. Academic Press, Cambridge (1979)



# Extended State Observer-Based Sliding Mode Control for Epilepsy

Wei Wei<sup>1,2,3</sup>, Ping Li<sup>1,2,3</sup>, and Min Zuo<sup>1,2,3(✉)</sup>

<sup>1</sup> School of Artificial Intelligence, Beijing Technology and Business University,  
Beijing 100048, China  
[zuomin@btbu.edu.cn](mailto:zuomin@btbu.edu.cn)

<sup>2</sup> National Engineering Laboratory for Agri-product Quality Traceability,  
Beijing Technology and Business University, Beijing 100048, China

<sup>3</sup> Beijing Key Laboratory of Big Data Technology for Food Safety,  
Beijing 100048, China

**Abstract.** Closed-loop neuromodulation has great potentials in epilepsy and other neurological diseases. Given the complexity and particularity of epilepsy, the control strategy should be robust enough to those uncertainties and disturbances. To reduce the dependence on model information and ensure the closed-loop regulation of epilepsy, an extended state observer-based sliding mode control scheme is designed to suppress epileptic spikes produced by the Jansens neural mass model. Simulation results confirm the extended state observer-based sliding mode control in suppressing epilepsy.

**Keywords:** Epilepsy · Neural mass model · Closed-loop neuromodulation · Extended state observer · Sliding mode control

## 1 Introduction

Epilepsy is one of the most common neurological disorders, and over 50 million people have been suffering great pain and discomfort resulting from epilepsy [1]. Epilepsy is characterized by recurrent seizures resulting from excessive excitation or inadequate inhibition of neurons [2]. Treatment of epilepsy is always a hot topic of doctors and scholars, and various approaches have been proposed. At present, there are three kinds of treatment methods, including drug therapy, surgical treatment, and neuromodulation therapy.

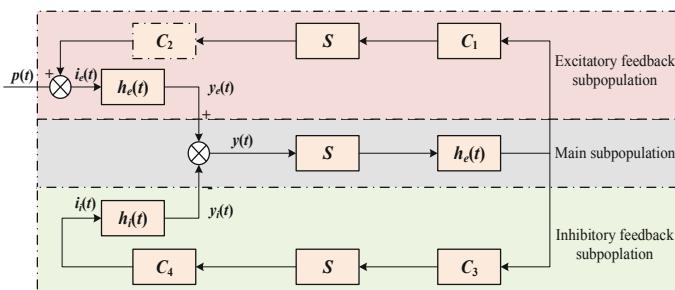
For drug therapy, antiepileptic drugs (AEDs) may be helpful. However, long-term drug treatment will lead to drug-resistant epilepsy (DRE). Moreover, AEDs can also bring many side effects [3]. Surgical is not suited for all DRE patients, identifying the epileptogenic zone (EZ) is also a challenge. Comparatively, neuromodulation has little damage to nerve tissues and it is effective in suppressing epilepsy [4].

For the neuromodulation, there are open-loop mode and closed-loop mode in clinic. The open-loop stimulation is generated by a device and sent to the brain.

Additionally, stimulus parameters are programmed in advance and cannot adapt to the patients clinical symptoms. Moreover, the open-loop mode depends largely on the clinical experience [5]. One the contrary, the closed-loop mode is not a scheduled stimulation. It adjusts the stimuli in real-time according to patients reactions.

Modulations can be studied on animals or computational models. Before an animal model is utilized, a computational model is always employed to accumulate experience. Thus, most epilepsy control is currently studied on computational models. Neural mass model (NMM) is a commonly utilized model. NMM includes parameters directly related to mechanisms of epilepsy, such as excitation, inhibition, or connection. It reveals the physiological mechanism of seizures and provides a platform for checking the closed-loop neuromodulation of epilepsy. Scholars have paid attention to the closed-loop mode and made some achievements [6–11]. PID control has been designed to suppress epileptiform waves generated by an NMM [6, 7]. However, PID is a passive control method, which is weak in dealing with nonlinear, uncertainties, and disturbances. Therefore, it is difficult to obtain satisfactory performance in suppressing the epilepsy. To ensure satisfied performance, fuzzy PID control [8] was applied to the NMM with multiple coupled neural populations. However, fuzzy control rules are largely dependent on experience. Feedback linearization control [9] depends on an accurate model. However, epilepsy is extremely complex and it is impossible to obtain an accurate model. Unscented Kalman filter (UKF) based closed-loop iterative learning control (ILC) [10], and particle swarm optimization (PSO) [11] were utilized to estimate key parameters on a computational model, satisfied numerical results are obtained. However, it is still obscure that which parameter(s) induce(s) epilepsy in clinic.

Due to the uncertainties and inevitable disturbances, clinical treatment requires a control method that depends less on a model of epilepsy. Then, it is robust to various disturbances. Active disturbance rejection control (ADRC) is the very approach that suits the control of epilepsy. Extended state observer (ESO) estimates the total disturbance, and a control law compensates the total disturbance in real-time. For improving the performance furtherly, sliding mode control (SMC) is designed to suppress those uncompensated disturbances.



**Fig. 1.** Structure of the neural mass model

## 2 Neural Mass Model

A neural mass model (NMM) describes average activity level of neuron clusters. A single Jansens neural mass model [12] is taken to simulate the epileptiform firing pattern, and Fig. 1 shows its structure.

The NMM is composed of main, excitatory, and inhibitory feedback subpopulations. Each subgroup consists of two parts. A second-order linear transformation function and an asymmetric S-shaped function. The linear conversion function converts the average presynaptic discharge rate to the average postsynaptic membrane potential. When  $t > 0$ , for the excitatory and inhibitory subgroups, their impulse responses are

$$\begin{cases} h_e(t) = H_e t e^{-t/\tau_e} / \tau_e \\ h_i(t) = H_i t e^{-t/\tau_i} / \tau_i \end{cases} \quad (1)$$

where  $H_e$  is excitatory synaptic gain, and  $H_i$  is inhibitory synaptic gain.  $\tau_e$  and  $\tau_i$  represent the time delays in the dendritic transmission.  $H_e$  and  $H_i$  can be adjusted to get different outputs.

$S(v)$  converts the average membrane potential to the average density of the action potential, which is the average ignition rate. Static nonlinear transfer function can be expressed as

$$S(v) = \frac{2e_0}{1 + e^{r(v_0 - v)}} \quad (2)$$

where  $2e_0$  represents the maximum discharge rate of the cluster,  $v_0$  is the postsynaptic potential with a 50% firing rate, and  $r$  represents the slope of  $S(v)$  at  $v_0$ .

In the model shown in Fig. 1, parameters  $C_1$  to  $C_4$  are connection constants, representing the average number of synaptic connections between a vertebral cell and an intermediate neuron.  $y(t)$  represents the output of the cluster, and input  $p(t)$  is simulated by gaussian white noise with a mean value of 90 and a variance of 30.

Standard values of those parameters can be found in [7].

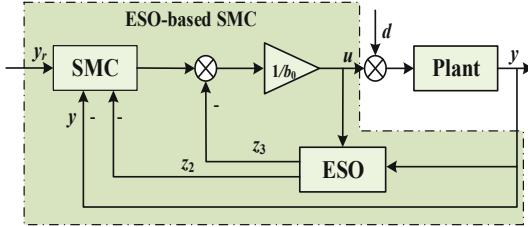
## 3 Control Strategy

NMM in Fig. 1 can be considered as [13].

$$\ddot{y} = f(y, \dot{y}, w) + b_0 u \quad (3)$$

where  $u$  is the regulation signal,  $y$  is the system output,  $b_0$  is a non-zero constant,  $w$  represents external disturbances, and  $f(y, \dot{y}, w)$  represents the total disturbance.

Based on the characteristics of the NMM and the control requirements, an ESO based SMC is designed. The ESO estimates the total disturbance, and the SMC suppresses the disturbance that is not fully compensated. Figure 2 presents the closed-loop scheme.



**Fig. 2.** The closed-loop system structure of the ESO-based SMC

### 3.1 Extended State Observer Design

An ESO is the core part in ADRC. It estimates internal and external disturbances in the control system, then compensate them to enhance the disturbance rejection of the system [14]. It is described as

$$\begin{cases} \dot{z}_1 = z_2 + \beta_1(y - z_1) \\ \dot{z}_2 = z_3 + \beta_2(y - z_1) + b_0 u \\ \dot{z}_3 = \beta_3(y - z_1) \end{cases} \quad (4)$$

where  $z_1, z_2, z_3$  are the estimation of the output value, the rate of change of the system output, and the total disturbance, respectively.  $b_0$  is control gain, and  $\beta_1, \beta_2, \beta_3$  are the gain of ESO. Parameterization of ESO is discussed in [15], one has  $\beta_1 = 3\omega_o, \beta_2 = 3\omega_o^2, \beta_3 = \omega_o^3$ . Here,  $\omega_o$  is the observer bandwidth.  $f(y, \dot{y}, w)$  in (3) can be estimated by  $z_3$ . Based on estimation and compensation of an ESO, system (3) becomes (5). Here,  $f$  represents  $f(y, \dot{y}, w)$  in (3), and  $\tilde{e}_3$  is the estimation error of  $z_3$ .

$$\ddot{y} = f + b_0 u = z_3 + b_0 u + \tilde{e}_3 \quad (5)$$

### 3.2 SMC Design

A PID sliding surface is chosen to reduce the trajectory tracking error [16].

$$s = \lambda_0 \dot{e} + \lambda_1 e + \lambda_2 \int_0^t e dt \quad (6)$$

where the coefficients  $\lambda_0, \lambda_1$ , and  $\lambda_2$  are positive constants,  $\lambda_0, \lambda_1, \lambda_2 \in R^+$ , and  $e = y - y_r$  is the tracking error.

The control law  $u$  is defined as

$$u = u_{eq} + u_{sw} \quad (7)$$

Let  $\dot{s} = \lambda_0 \ddot{e} + \lambda_1 \dot{e} + \lambda_2 e = 0$ . Do not consider unknown information, then  $u_{eq}$  is

$$u_{eq} = \frac{1}{b_0} \left( -z_3 - \frac{\lambda_1}{\lambda_0} z_2 - \frac{\lambda_2}{\lambda_0} y \right) \quad (8)$$

$u_{sw}$  is designed based on an exponent reaching law

$$u_{sw} = \frac{1}{b_0}(-k_1 s - k_2 \text{sign}(s)) \quad (9)$$

where  $k_1 > 0$  and  $k_2 > 0$ . Substituting (8) and (9) into (7), one has

$$u = \frac{1}{b_0}(-z_3 - \frac{\lambda_1}{\lambda_0}z_2 - \frac{\lambda_2}{\lambda_0}y - k_1 s - k_2 \text{sign}(s)) = \frac{1}{b_0}(u_0 - z_3) \quad (10)$$

To reduce input chattering, a quasi-sliding mode is adopted to eliminate chattering. So-called quasi-sliding mode refers to the mode that system trajectory is confined to a certain  $\Delta$  neighborhood of the ideal sliding mode. Here, the boundary layer function replaces the  $\text{sign}(s)$ , and  $\varepsilon$  is small.

$$BLF(s) = \frac{s}{|s| + \varepsilon} \quad (11)$$

### 3.3 Stability Analysis

Differentiate the sliding surface, one has

$$\dot{s} = \lambda_0(f + b_0 u) + \lambda_1 \dot{y} + \lambda_2 y \quad (12)$$

Substituting the control law (11) into (13), one has

$$\dot{s} = \lambda_0(f - z_3 - k_1 s - k_2 \text{sign}(s)) + \lambda_1(\dot{y} - z_2) \quad (13)$$

Then, following theorem is obtained.

**Theorem 1.** *For epilepsy closed-loop neural regulations, the sliding mode surface (6) is taken. Based on the ESO (4) and the control law (10), for any initial condition, the error trajectory will arrive at the sliding surface, and the closed-loop system is stable.*

*Proof.* Estimation errors  $\tilde{e}_2 = \dot{y} - z_2$  and  $\tilde{e}_3 = f - z_3$ . According to Ref. [17], estimation errors  $\tilde{e}_2$  and  $\tilde{e}_3$  are bounded by selecting appropriate observer gains. In other words,  $|\tilde{e}_2| \leq \tilde{e}_2^* = \sup_{t>0} |\tilde{e}_2|$  and  $|\tilde{e}_3| \leq \tilde{e}_3^* = \sup_{t>0} |\tilde{e}_3|$ . Then, if the sliding mode control is stable, the closed-loop system will be stable. Lyapunov function candidate is

$$V = \frac{1}{2}s^2 \quad (14)$$

Then

$$\begin{aligned} s\dot{s} &= s\lambda_0\tilde{e}_3 + s\lambda_1\tilde{e}_2 - \lambda_0 k_1 s^2 - \lambda_0 k_2 |s| \\ &\leq s\lambda_0\tilde{e}_3 + s\lambda_1\tilde{e}_2 - \lambda_0 k_2 |s| \\ &\leq |s| \lambda_0 (\tilde{e}_3^* + \frac{\lambda_1}{\lambda_0} \tilde{e}_2^* - k_2) \end{aligned} \quad (15)$$

Because  $\lambda_0 > 0$ , then  $k_2 > \tilde{e}_3^* + \frac{\lambda_1}{\lambda_0} \tilde{e}_2^*$ ,  $\dot{V} \leq 0$  can be obtained.

In summary, the ESO-based SMC is stable.

## 4 Simulations

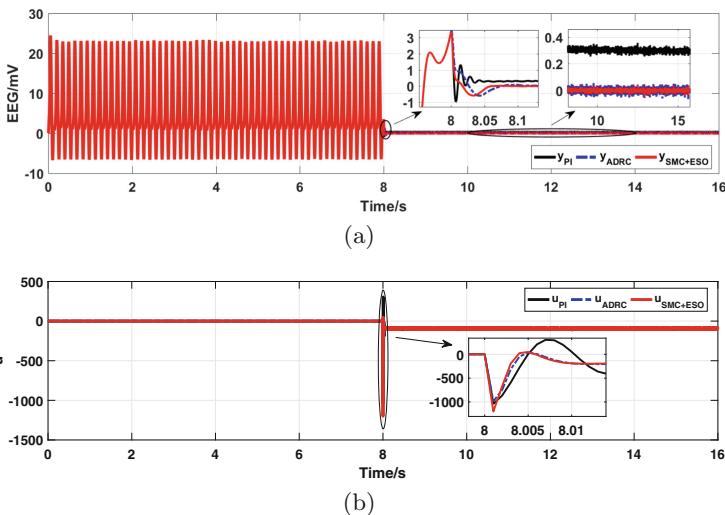
In this section, an NMM is taken to simulate EEG signals. Neuromodulations of ESO-based SMC, ADRC, and PI are compared. A group of parameters is taken to generate the epileptiform waves ( $H_e = 7$  and  $H_i = 22$ ). Simulations last for 16s, and the control signal is added from the 8th second. Cases without external disturbances and with sinusoidal disturbances are taken into consideration. Simulation parameters are listed in Table 1.

**Table 1.** Tunable parameters of PI, ADRC, and ESO-based SMC

	$b_0$	$\omega_c/k_p$	$\omega_o/k_i$	$\lambda_0$	$\lambda_1$	$\lambda_2$	$k_1$	$k_2$	$\varepsilon$
PI	—	310	2	—	—	—	—	—	—
ADRC	800	320	420	—	—	—	—	—	—
ESO+SMC	800	320	420	0.8	260	6500	430	6080	0.3

### 4.1 Case Without External Disturbance

No external disturbance, the EEG responses and regulation signals of the three control methods are shown in Fig. 3.



**Fig. 3.** Responses and regulation signals of the PI, ADRC, and ESO-based SMC

Figure 3(a) shows that three methods can control seizures for a short period of time. However, EEG signals modulated by ESO-based SMC response best. It

converges fastest and its fluctuations are smallest. Performance indexes of the three approaches are shown in Table 2.

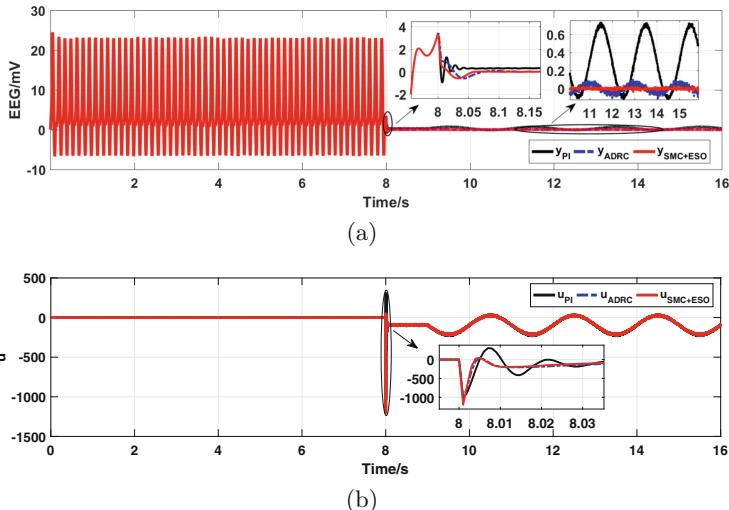
**Table 2.** Indexes of the PI, ADRC, and ESO-based SMC

	RMSE	MAE	ITAE	E
PI	0.2163	0.1502	28.7132	4.7220
ADRC	0.0582	0.0097	1.6993	4.4849
ESO+SMC	0.0463	0.0056	0.9647	4.5134
ESO+SMC vs PI	78.59%	96.27%	96.64%	4.42%
ESO+SMC vs ADRC	20.45%	42.27%	43.23%	-0.64%

In Table 2, RMSE represents the root-mean-square error, MAE represents the mean absolute error, ITAE is the integral of the time-multiplied absolute error, and E represents the energy consumption of a controller. The specific calculation formula is shown in (16).

$$RMSE = \sqrt{\frac{1}{k} \sum_{i=1}^k e_i^2}, MAE = \frac{1}{k} \sum_{i=1}^k |e_i|, ITAE = T \sum_{i=1}^k t_i |e_i|, E = \frac{T}{k} \sum_{i=1}^k u_i^2 \quad (16)$$

here  $e_i$  is the trajectory tracking error and  $T = 0.001$  is the sample time.



**Fig. 4.** Responses and regulation signals the PI, ADRC, and ESO-based SMC (sinusoidal disturbance exists)

From Table 2, it is easy to see that, compared with PI and ADRC, all indexes have been improved by similar energy consumption.

**Table 3.** Indexes of the PI, ADRC and ESO-based SMC (sinusoidal disturbance exists)

	RMSE	MAE	ITAE	E
PI	0.3031	0.1781	34.3018	9.1115
ADRC	0.0641	0.0201	3.7840	8.5267
ESO+SMC	0.0464	0.0060	1.0397	8.5246
ESO+SMC vs PI	84.69%	96.63%	96.97%	6.44%
ESO+SMC vs ADRC	27.61%	70.15%	72.52%	0.02%

## 4.2 Case with External Disturbance

To test the robustness,  $d = 120 \sin(\pi t)$  is taken. It is introduced from the 9th second. When control disturbance is added, the time-domain responses and control signals of the three methods are shown in Fig. 4.

Although ADRC can estimate and eliminate it to a certain extent, ESO-based SMC has stronger robustness and better performance than the ADRC. Data listed in Table 3 show that ESO-based SMC can achieve smaller RMSE, MAE, and ITAE with less energy. It confirms that the ESO-based SMC is more effective.

Above experiments discuss the time-domain responses in absence and presence of external disturbance. Simulation results show that RMAE, MAE, and ITAE of the ESO-based SMC are the smallest. In other words, the ESO-based SMC obtain the best results with a similar or lower energy consumption. It means that the ESO-based SMC is more robust to nonlinear, uncertain, and external disturbance.

## 5 Conclusion

To deal with high-amplitude epileptic activities caused by abnormal discharge, ESO-based SMC is designed in this paper. By comparison with PI and ADRC, one can see that the ESO-based SMC behaves best. At the same time, SMC can be realized only when state variables are available. However, the ESO-based SMC is less dependent on them and it provides a reference for the closed-loop neuromodulation of the epilepsy.

**Acknowledgments.** This work is supported by National Natural Science Foundation of China (61873005), the key program of Beijing Municipal Education Commission (KZ201810011012), and Support Project of High-level Teachers in Beijing Municipal Universities in the period of 13th Five-year Plan (CIT&TCD201704044).

## References

1. Zheng, Y., Jiang, Z., et al.: Acute seizure control efficacy of multi-site closed-loop stimulation in a temporal lobe seizure model. *IEEE Trans. Neural Syst. Rehab. Eng.* **20**(3), 419–428 (2019)
2. Lin, Z., Meng, L., Zou, J., et al.: Non-invasive ultrasonic neuromodulation of neuronal excitability for treatment of epilepsy. *Theranostics* **10**(12), 5514–5526 (2020)
3. Brodie, M.J., Besag, F., Ettinger, A.B., et al.: Epilepsy, antiepileptic drugs, and aggression: an evidence-based review. *Pharmacol. Rev.* **68**(3), 563–602 (2016)
4. Yuan, Y., Changqing, J., Yue, C., et al.: Development and prospect of neuromodulation technology. *Life Sci. Instr.* **16**, 20–28 (2018)
5. Yan, D.: Research progress of neuroregulation technology in epilepsy treatment. *Int. J. Neurosurg.* **44**(3), 302–307 (2017)
6. Wang, J.S., et al.: Closed-loop control of epileptiform activities in a neural population model using a proportional-derivative controller. *Chin. Phys. B* **24**(3), 1–8 (2015)
7. Wang, J.S., et al.: Suppressing epileptic activity in a neural mass model using a closed-loop proportional-integral controller. *Sci. Rep.* **6**(6), 1–12 (2016)
8. Liu, X., Liu, H.J., Tang, Y.G., et al.: Fuzzy PID control of epileptiform spikes in a neural mass model. *Nonlinear Dyn.* **71**(1), 13–23 (2013)
9. Cao, Y.Z., Ren, K.L., Su, F., et al.: Suppression of seizures based on the multi-coupled neural mass model. *Chaos* **25**(10), 1–13 (2015)
10. Shan, B.N., et al.: UKF-based closed loop iterative learning control of epileptiform wave in a neural mass model. *Cogn. Neurodyn.* **9**(1), 31–40 (2015)
11. Shan, B.N., Wang, J., Deng, B., et al.: Particle swarm optimization algorithm based parameters estimation and control of epileptiform spikes in a neural mass model. *Chaos* **26**(7), 1–9 (2016)
12. Jansen, B.H., Rit, V.G.: Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biol. Cybern.* **73**(4), 357–366 (1995)
13. Wei, W., Wei, X., Zuo, M., et al.: Seizure control in a neural mass model by an active disturbance rejection approach. *Int. J. Adv. Robot. Syst.* **16**(6), 1–15 (2019)
14. Han, J.: From PID to active disturbance rejection control. *IEEE Trans. Ind. Electron.* **56**(3), 900–906 (2009)
15. Gao, Z.: Scaling and bandwidth-parameterization based controller tuning. In: Proceedings of the American Control Conference, vol. 6, pp. 4989–4996. IEEE, Denver (2003)
16. Al-Jodah, A., et al.: A fuzzy disturbance observer based control approach for a novel 1-DOF micropositioning mechanism. *Mechatronics* **65**, 102317 (2020)
17. Chen, Z.Q., Sun, M.W., Yang, R.G.: On the stability of linear active disturbance rejection control. *Acta Automatica Sinica* **39**(5), 574–580 (2014)



# Research on General System Level Training Simulation Technology of Mid-And High-End Military UAV

Qing Zhang<sup>1,2(✉)</sup>, Jiahui Tong<sup>1,2</sup>, and Haifeng Li<sup>2</sup>

<sup>1</sup> Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing, China

yisanyue@163.com

<sup>2</sup> Beijing Electromechanical Engineering Institute, Beijing, China

**Abstract.** A common training simulation system is proposed for the general-purpose and efficient control training needs of the mid-and high-end military UAV system. The general modeling technology and process-configurable interface protocol modeling method are adopted to realize the rapid construction of flight simulation subsystem of various aircraft models. A scheme based on general bus monitoring and fast extraction equipment is designed to realize the fast embedded real-time control training of UAV with different physical devices. A self-correctable control and evaluation method of UAV training based on multi-sample statistics is proposed to enhance the targeted strengthening setting of training subjects for trainees, so as to improve training efficiency and training effect.

**Keywords:** UAV · Training simulation · Bus monitoring and extraction · Self-correction · Control and evaluation

## 1 Introduction

Mid- and high-end military UAV is a kind of high-value, complex and sophisticated weapon rye, a very small hidden danger may lead to the destruction of the UAV, resulting in great economic loss and adverse effects. In order to improve the combat effectiveness of the UAV system, the UAV is required to have a complex and variable battlefield environment of high adaptability and survival capacity. As a result, there is a need for extensive flight training of UAV commanders and operators to improve their alleged proficiency and ability to respond to emergencies to maximize the safety of the drone. The life of the UAV system is limited, frequent live training is not allowed, and there are certain risks. Using the method of simulation and simulation instead of real-flying training is not only not limited by the space environment and weather conditions, but also can greatly reduce the training cost and improve the efficiency of training [1].

## 2 Training Overall Architecture Technology for Mid- to High-End Military UAV

### 2.1 Analysis of the Needs of Unmanned Aerial Systems Training

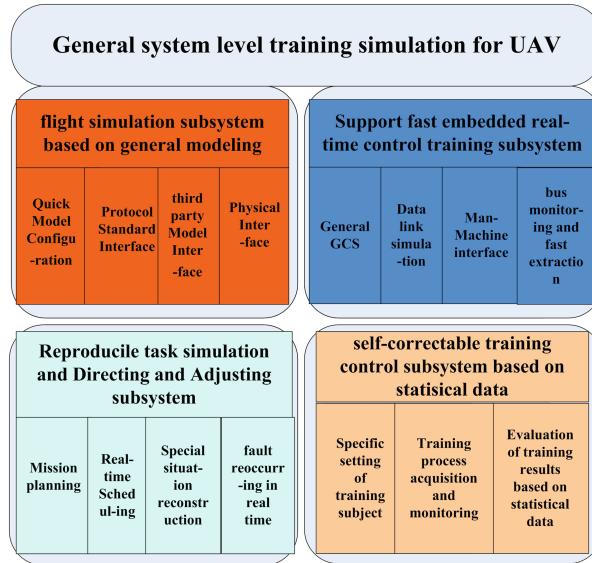
Military UAV are diverse, the types are very wide, and the use of various military types are also very different, such as target flight, detection and strike, electronic interference and so on, there are autonomous flight, remote control flight and other control methods, thus putting forward a variety of needs for the training of the UAV system [2]. Usually the maneuvering training needs of drones are generally in the following areas [3]:

- 1) Simulating the conventional flight operation of the UAV: the operation training of the normal flight process of the drone's ground self-test, take-off, flight, recovery/landing;
- 2) Simulated UAV combat mission operation: UAV load detection, interference, strike control training;
- 3) Special war situation training: typical failure re-emergence, dangerous conditions, bad weather and other complex battlefield environment flight control training;
- 4) Command and control training of drones;
- 5) Training in the handling and handling of unmanned aerial surveillance intelligence;
- 6) UAV communication management control training, etc.

In order to carry out in-depth excavation and find its consistency, the overall structure of the system-level general-purpose UAV training system is proposed, which supports the construction of the full operation process, multi-task mode and comprehensive control training simulation system, and improves the generalization, coverage and efficiency of the UAV control training simulation system [4,5].

### 2.2 The Overall Architecture of the UAV General Training Simulation System

Under the traction of system-level training requirements, the UAV system-level general training simulation system should have the simulation functions of flight process, mission flow, special war situation, command control, intelligence processing and so on, and realize the functions of flight control training, mission control training, special combat training, command and control training, etc., and the overall architecture design of the general training simulation system for demand should be compatible with the functional requirements, model requirements, interface requirements and hardware requirements of different UAV [6,7]. The system structure includes flight simulation subsystem based on general modeling, real-time control training subsystem supporting fast embedding, reproducible task simulation and directing and adjusting subsystem and self-correctable training control subsystem based on statistical data, as shown in Fig. 1.



**Fig. 1.** The composition of the UAV system-level general training simulation system

### 3 Key Technologies of General System Level Training Simulation for Mid-And High-end Military UAV

#### 3.1 Flight Simulation Modeling Method Based on Generalized Modeling

Different drones are different in terms of aerodynamic layout, guidance control, fire power system, mission load, avionics equipment and other system composition and workflow, mission requirements and so on. Simulation modeling is heavy and portability is poor, which leads to poor versatility and reuse of flight simulation subsystem. Therefore, the UAV flight simulation subsystem adopts the common modeling technology and process configurable interface protocol modeling method to realize the rapid construction of the flight simulation subsystem of various models, and to meet the needs of the UAV general training simulation system.

- 1) The construction of the UAV training simulation system adopts the object-oriented idea, establishes a common mathematical model library and universal algorithm, based on the graphical, dragable rapid modeling method, can quickly build a simulation model system with complete data flow relationship and parameter mapping, and model the model for the characteristics of the full-machine extension system function unit of the drone.
- 2) According to the typical high-end UAV conventional system composition mechanism, combining the types of interfaces, performance, characteristics, etc., to sort and organize. This paper analyzes the input and output relationships

- of each system, adopts the modeling mechanism based on the unified standard of protocol, targets the multi-class interface of the unmanned aerial system, and establishes the standard of the interface of the unmanned aerial system.
- 3) According to the full mission process of the UAV and the simulation of the characteristics of the equipment interface, it provides a unified and standardized interface for external third-party models, embedded real-installed equipment, etc., and realizes the interconnection between different forms of UAV system models and physical objects.

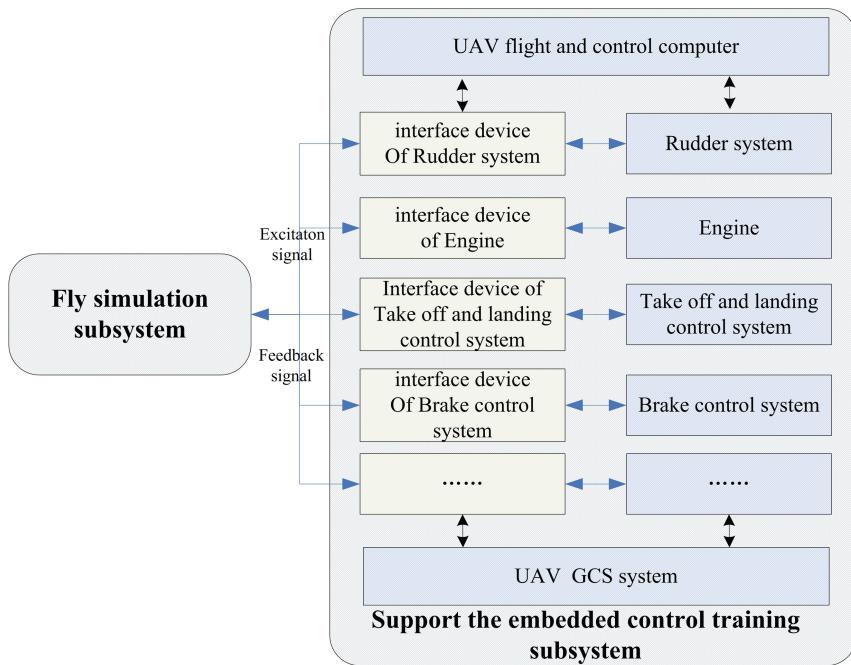
### 3.2 A Real-Time Control Training Scheme Supporting Fast Embedding of Real Equipment

**Regular Support for Real-Time Control Training Subsystem Composition:** In the development process, the mid- and high-end UAV pay special attention to the performance of some key equipment physical objects (such as engine, rudder system, brake control system, take-off and landing control system, etc.) in actual flight, which needs to be used to generate excitation signals by the flight simulation subsystem, to simulate the external environment and interface characteristics of its actual flight on the ground, and to match the operation training and inspection of its complete workflow and work performanceinst [8]. Usually the flight simulation subsystem uses the physical test interface equipment to achieve the control training embedded in the real equipment, as shown in Fig. 2.

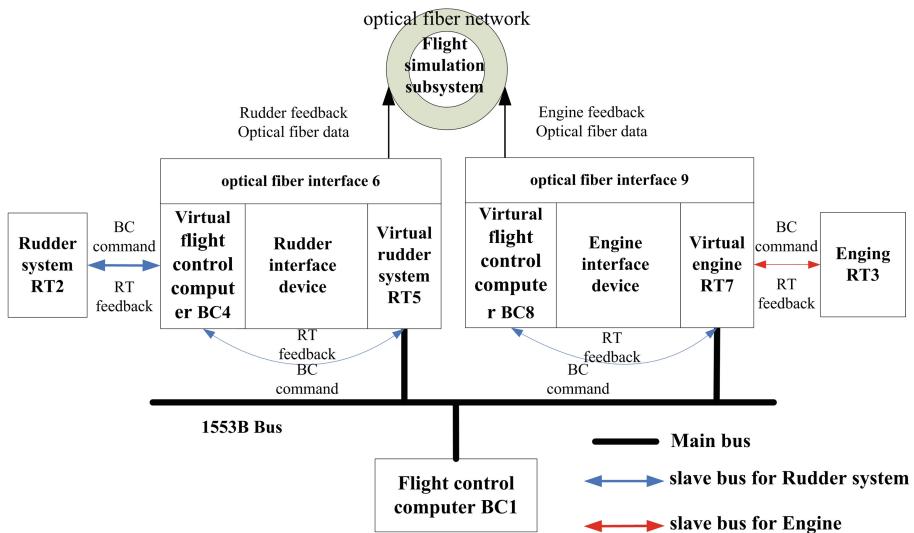
The test interface equipment is usually composed of interface computer, interface board, fiber optic board and interface software. It can be seen that different practical equipment training needs different interface equipment, multiple real equipment training also need slots of interface equipment, and when the various real equipment and flight control machine between the use of different communication interface (such as 1553B, RS-422, CAN, etc.), it is more necessary to use a number of different types of interface equipment. With the increase of physical equipment on board it will lead to the control training system is unusually large, and even lead to the inability to achieve real-time control training.

**Analysis of Real-Time Limitation:** Taking a certain UAV with 1553 B bus communication as an example, in addition to the flight control machine as the BC node, there are 9 RT nodes, such as rudder system, engine, take-off and landing control system, navigation system, etc., which require the real-time scheduling period 5 ms of the training simulation system. The closed-loop training simulation system is constructed with only one actual equipment of the flight control machine, which is equivalent to the interface equipment in Fig. 2. The computation time of flight simulation subsystem model is about 1 ms, the communication time between flight controller and pure virtual equipment is about 3 ms, the total time consumption of real-time cycle scheduling is 4 ms, which temporarily satisfies 5 ms time scheduling period constraints.

However, with the increase of the nodes of the rudder system and engine, according to the flow of the interface data in Fig. 3, a virtual node with two-way



**Fig. 2.** Regular support for real-time control training subsystem composition



**Fig. 3.** Real-time control training data flow diagram based on interface equipment

interface is added for each additional real device. When all the 9 real devices are tested, the total time consuming of the real-time periodic scheduling is:

$$1 \text{ ms} + (3/9) * 18 \text{ ms} = 7 \text{ ms} > 5 \text{ ms}. \quad (1)$$

So this scheme can not meet the requirement of training real-time simulation with a large number of real-time equipments. Therefore, it is necessary to take an effective method to collect the feedback of different real-time devices embedded in the control training, and to solve the problem of the universality of the closed-loop real-time solution of the flight simulation model.

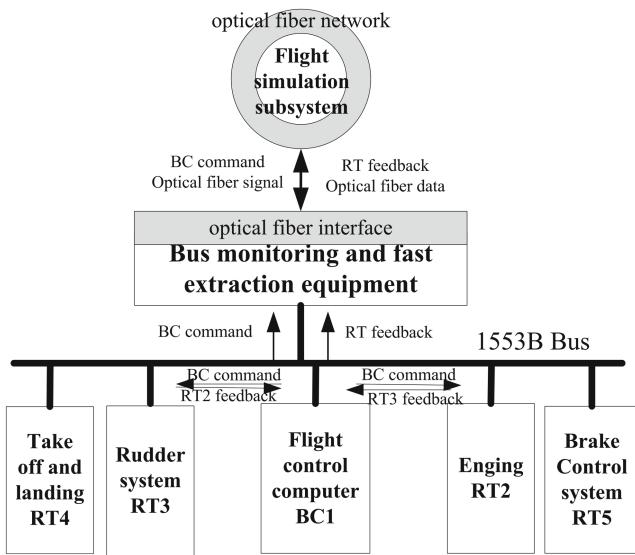
**Training of Embedded Control Based on Bus Monitoring and Fast Extraction:** In this paper, a real-time control training scheme based on universal bus monitoring and fast extraction equipment is designed to realize the fast embedding of different real-time devices. The real-time information interaction between the real-time equipment and the fast extraction equipment is carried out directly with the flight control machine, and the execution information is extracted in real time by the general bus monitoring and fast extraction equipment as a listener, and transmitted synchronously to the flight simulation subsystem to complete the real-time flight closed-loop solution. The general bus monitoring and fast extraction equipment is equipped with 1553B, RS-422, CAN and other conventional bus interfaces and fiber optic communication interfaces to quickly realize the response and support to the embedded control training of all kinds of real equipment of UAV with different interface forms.

Also in the form of 1553B bus interface rudder system, engine, take-off and landing control system, brake control system when the actual training of its interface data flow is shown in Fig. 4. Compared with the scheme in Fig. 3, the use of this scheme greatly simplifies the composition and data flow process of the control training subsystem, which not only reduces the complexity of the control training system and improves its generality.

### 3.3 Reproducible Task Simulation and Directing and Adjusting Mechanism

Mission simulation and guidance subsystem mainly plays the role of operational simulation and training task guidance control in the simulation of UAV control training. The subsystem designs a typical special situation and fault instant recovery guidance mechanism [9].

- 1) Dynamic task creation and scheduling based on full flight flow. It ensures that the typical task is issued and the real-time task scheduling is realized. Typical tasks are mainly issued by drone trainers (hereinafter referred to as instructors) according to the tactical situation of the battlefield, the training of the trainees (hereinafter referred to as the cadets) issued the task, using a pre-set loading method to carry out.



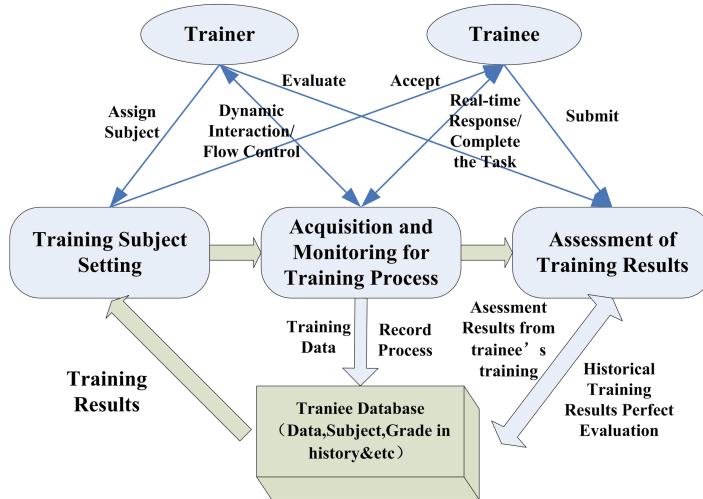
**Fig. 4.** Data flow diagram of real-installed control training based on universal bus monitoring and fast extraction equipment

- 2) The leading and adjusting mechanism of the typical combat situation and the instant recurrence of the fault. It mainly realizes the reproducible and reproducible training of the reconstruction of combat missions and the immediate recurrence of faults under typical combat situations. In the link to increase the Trainer's explanation link, emphasizing the repeated reproduction of training details, and then strengthen the training awareness and training effect of the students.

### 3.4 Self-correctable Training Control Method Based on Statistical Data

The subsystem of UAV training management is an important system to realize the evaluation of the whole UAV training simulation process management and training results [10]. In order to realize the effect of the targeted intensive training for different participants, a self-correctable training mechanism based on statistical data is designed, the training subjects can be set according to the historical achievements of the trainees, and the evaluation factors of the training evaluation system are constantly corrected based on the training data of multiple statistics, so as to perfect the automatic iteration of the training evaluation system.

**Targeted Settings for Training Subjects:** According to the statistical results of the personnel training loss elements, training subjects, training tasks,



**Fig. 5.** Data flow chart for self-correcting drone training control and evaluation based on statistical data

training scenes to strengthen the set-up, strengthen the targeted, to achieve the teaching and personal training efficiency of the efficient improvement.

**Training Process Acquisition and Monitoring:** In order to be able to judge the trainees's comprehensive performance more comprehensively, the actual manipulation process data during the training process of the trainees should be collected in real time, and the typical physiological characteristics of the human body in the training process should be collected and monitored, and the raw data of the subsequent training results are formed.

**Evaluation of Training Results Based on Statistics:** The first training results of the trainee by the trainer using the scoring system a prior value, mainly according to the training of the participants performance of the control accuracy, task completion and other comprehensive performance of the score. The training results are recorded in the training history library of the participants and become the important sample data of the reference factor weight adjustment and self-correction and optimization of the prior score system. It also becomes an important scoring reference for the trainee's next training performance. The implementation process for self-correctable control and evaluation of drone training based on statistical data is shown in Fig. 5.

## 4 Summary

This paper presents a solution for a general training simulation system that is compatible with the functional requirements, model requirements, interface

requirements and hardware requirements of different mid-and high-end military UAV systems, with the following technical advantages:

- 1) The rapid construction of flight simulation subsystem for a variety of models is realized by using general modeling technology and process-configurable interface protocol modeling method.
- 2) A scheme based on general bus monitoring and fast extraction equipment is designed to realize the fast embedded real-time control training of different physical devices of the UAV, which greatly reduces the complexity of the control training system and improves its versatility.
- 3) A typical combat situation and fault instant re-emergence guidance mechanism is proposed, which can not only realize the real simulation of the full flight process of the combat desired mission, but also effectively strengthen the emergency response capability of training the trainees of UAV in various emergency or sudden failure situations in complex battlefield environment.
- 4) A self-correctable training control and evaluation method based on multi-sample statistics is proposed, the scoring factor weight value of the correction experience scoring system is automatically adjusted by using the multi-sample statistics accumulated by repeated training. It not only greatly improves the intelligence of the UAV training simulation and evaluation system, and realizes the optimization and perfection of its autistic ring, so as to improve the training efficiency and enhance the training effect.

## References

1. Jin, C., Jin, L., et al.: Foundation and Application of Hardware in the Loop Simulation Technology. China Aerospace Publishing House, Beijing (2020)
2. Ruixuan, W., Xueren, L.: UAV System and Operational Use. National Defense Industrial Press, Beijing (2009)
3. Yuanping, Z., Lili, Z.: U.S. unmanned aircraft military system training. Trainer **9**(12), 29–37 (2011)
4. Dequan, D.: Requirement analysis of UAV training system. Electron. Prod. **1**(12), 45–81 (2015)
5. Kendall, J.: Deputy director of intelligence, operations and nuclear integration for flying training air education training command USAF flying training perspective. Air Education and Training Command United States Air Force (2009)
6. Yu, J., Kou, K., Chen, Y., Zhang, K.: Design simulation system for training and teaching of UAV. Research and exploration in laboratory, vol. 33, no. 7, pp. 221–224 (2014)
7. Xie, D.: Design and implementation of UAV simulation training system. China High Tech Enterprises, vol. 31, no. 1, pp. 30–32 (2015)
8. Yu, H., Peng, Z.: Training Simulation System of UAV. Ordnance Industry Automation, vol. 35, no. 7, pp. 18–21 (2016)
9. Qing, L., Zhihao, C., Yingxun, W., Lifang, C.: Method of UAV emergency disposition training based on fault tree analysis and snapshot. J. Beijing Univ. Aeronaut. Astronaut. **39**(11), 1548–1552 (2013)
10. Jiao, Y., Chen, Y., Li, R.: Discussion on simulation training teaching of UAV specialty. Exp. Technol. Manage. **34**(1), 109–111 (2017)



# Event-Based Robust State and Fault Estimation for Stochastic Linear System with Missing Observations and Uncertainty

Zhidong Xu, Bo Ding<sup>(✉)</sup>, and Tianping Zhang

School of Information Engineering, Yangzhou University, Jiangsu 225127, China  
zdimrs@163.com

**Abstract.** This paper investigates event-based robust state and fault estimation problem for stochastic linear system with missing observations and norm-bounded uncertainty. Missing observations is depicted by Bernoulli distributed process. The measurements transmit to estimator unless the current innovation disobey the Send-on-Delta (SoD) conditions. Filtering algorithm is constructed by virtue of augmented state method combined with two-stage robust event-based estimators. Upper bounds of the estimation error covariance are obtained respectively according to solving discrete Riccati difference equations. Subsequently, by appropriately formulating the filter gain matrices such that the upper bounds are minimal. Simulation results demonstrate the usefulness of the algorithm.

**Keywords:** Event-based · State and fault estimation · Stochastic linear system · Missing observations · Uncertainty

## 1 Introduction

The demand for increased productivity has led to a lot of challenging operating conditions for many practical system. These conditions increase the possibility of system failure. Therefore, a relatively accurate estimations of state and fault play a major part in the industrial field of practical significance. Nevertheless, system uncertainties can't be ignored for system linearizing, parameter changing and model simplification. Meanwhile, the true measurements may be corrupted by the restrictions on the transmission. Hence, under these circumstances, Kalman filter is no longer suitable. Therefore, we need to develop effective robust filter for system with missing observations and uncertainty.

Fault estimation is a powerful method that can accomplish the result of fault detection, fault identification and fault isolation. Recently, fruitful results have been achieved for fault estimation such as adaptive fault estimation scheme [1], the sliding mode fault estimation strategy [2], unknown input observers method

© The Editor(s) (if applicable) and The Author(s), under exclusive license

to Springer Nature Singapore Pte Ltd. 2021

Y. Jia et al. (Eds.): CISC 2020, LNEE 705, pp. 819–831, 2021.

[https://doi.org/10.1007/978-981-15-8450-3\\_85](https://doi.org/10.1007/978-981-15-8450-3_85)

[3] and descriptor observers method [4, 5]. In [6], intermediate estimator strategy was proposed to realize state and fault estimation.

The system parameters that describe input and output relationship are not often entirely known, which increase the difficulty of filter design. Many results have been done in the study of multifarious system with norm-bounded uncertainty [7–9]. In [10], a robust finite-horizon filtering has been proposed for stochastic system with missing measurement and norm-bounded uncertainty in state matrix. In [11], uncertainty has been extended into the state, output and white noise matrices. In [12], improved robust finite-horizon Kalman filtering has been introduced by reorganizing measurements. In [10–13], method of minimizing the upper bounds of the proposed filter has been utilized by designing gain matrices.

Recently, more and more scholars focus on the event triggering mechanism. Event triggering instead of time triggering can reduce energy waste and save communication resources. In [14], a SoD method has been proposed, which is a natural signal-dependent temporal sampling rule. Since then, different event-triggered schemes have been put forward in [15, 16]. An event-triggered controller for time delay system has been proposed in [17], which introduced the method of co-designing feedback gain and trigger parameters and analyzed the stability with an  $H_\infty$  norm bound. More recently, stochastic event triggers have been studied in [18, 19]. However, as far as the author knows, event-based state and fault estimation problem for stochastic linear system with missing observations and uncertainty has not been investigated.

In this paper, we aim to derive state and fault estimation based on event triggered scheme for stochastic linear system with missing observations and uncertainty. By assuming the fault as incipient fault, we augment fault as a part of state to realize fault estimation. According to the solutions to discrete Riccati equations, we obtain the upper bounds of the estimation error covariance and take the partial derivative of filter parameters for the trace of the upper bounds, and then filter parameters are derived.

Structure of this paper is as following. Sect. 2 presents stochastic linear system with missing observations and uncertainty and the event triggered condition. In Sect. 3, we formulate the event-based robust filter under the augmented state, and upper bounds of estimation error covariance are obtained separately. Besides, the filter parameters are designed to ensure the minimization of the upper bounds. Section 4 tests the validity of the filter. At last, conclusions are drawn in Sect. 5.

## 2 Problem Formulation

Consider the following stochastic linear system with missing observations and uncertainty

$$\begin{cases} x(s+1) = (A(s) + \Delta A(s))x(s) + D(s)f(s) + w(s) \\ y(s) = \xi(s)C(s)x(s) + v(s) \end{cases} \quad (1)$$

where  $s$  is the discrete time index,  $x(s) \in \mathbb{R}^n$ ,  $y(s) \in \mathbb{R}^m$  stand for system state, system output respectively.  $f(s) \in \mathbb{R}^p$  represents the additive fault, which might be treated as actuator fault.  $\Delta A(s)$  is the norm-bounded uncertainty.

$$\Delta A(s) = M(s)O(s)N(s) \quad (2)$$

where  $O^T(s)O(s) \leq I$ .  $\xi(s)$  describes the Bernoulli random process. 0 and 1 are the values of  $\xi(s)$ .

$$\begin{aligned}\mathbb{P}(\xi(s) = 1) &= \bar{\xi} \\ \mathbb{P}(\xi(s) = 0) &= 1 - \bar{\xi}\end{aligned} \quad (3)$$

The system noise  $w(s) \in \mathbb{R}^n$  and measurement noise  $v(s) \in \mathbb{R}^m$  are zero-mean, white random noise.  $Q(s) \geq 0$ ,  $R(s) \geq 0$  are covariance matrices of  $w(s)$  and  $v(s)$ .  $A(s), C(s), M(s), N(s), O(s)$  are identified matrices with suitable dimensions. Additionally, the fault  $f(s)$  is assumed as [20]

$$f(s+1) = f(s) + w^f(s) \quad (4)$$

where  $w^f(s)$  is virtual noise with covariance  $Q^f(s)$ . The initial fault  $f(0)$  is zero and the covariance  $P^f(0)$ .

*Remark 1.* The noise  $w^f(s)$  is virtual, we deem that the fault changes slowly or even does not change. As a result, the fault is incipient.

For ameliorating efficiency of transmission, we use event triggering to replace time triggering. When the measurement output meets certain conditions, the measurement output will be transmitted to the estimator. Here we use the method of SoD strategy [14].

$$f(d(s), \delta) = (d^T(s)d(s)) - \delta > 0 \quad (5)$$

where  $d(s) = \epsilon_m(s) - \epsilon(s)$ ,  $\epsilon(s)$  is the innovation sequence.  $\epsilon_m(s)$  is the transmitted innovation at the last event time and  $\delta$  is a known scalar. The current innovation sequence will be transmitted to estimator when (5) is satisfied, then the triggering instants  $0 \leq s_0 \leq s_1 \leq \dots \leq s_l \leq \dots N$  can be chosen as

$$s_{l+1} = \min \{s \in (0, N) | s > s_l, f(d(s), \delta) > 0\} \quad (6)$$

*Remark 2.* All measurements will be transmitted to estimator if  $\delta = 0$  and it becomes a time triggered one.  $(s_0, s_1, \dots, s_l, \dots)$  is the subset of  $(0, 1, \dots, N)$ .

### 3 Event-Based Robust Filter Proposed and Parameters Design

#### 3.1 Event-Based Robust Filter Proposed

Augmented state method is a simple and effective strategy commonly used in fault estimation. The stochastic linear system in augmented form is defined as follows

$$\begin{cases} X(s+1) = (\mathbf{A}(s) + \Delta \mathbf{A}(s))X(s) + W(s) \\ y(s) = \xi(s)\mathbf{C}(s)X(s) + v(s) \end{cases} \quad (7)$$

where  $X(s) = [x(s) \ f(s)]^T$ ,  $W(s) = [w(s) \ w^f(s)]^T$ ,  $\mathbf{A}(s) = \begin{bmatrix} A(s) & D(s) \\ 0 & I \end{bmatrix}$ ,  $\Delta\mathbf{A}(s) = \begin{bmatrix} \Delta A(s) & 0 \\ 0 & 0 \end{bmatrix}$ ,  $\mathbf{C}(s) = [C(s) \ 0]$ ,  $\mathbf{Q}(s)$  is the covirance of  $W(s)$ .

We consider the following event-based robust filter for the system (7) is represented as

$$\hat{X}(s|s) = \hat{X}(s|s-1) + L(s)\epsilon_m(s) \quad (8)$$

$$\tilde{X}(s+1|s) = \hat{\mathbf{A}}(s)\hat{X}(s|s-1) + K(s)\epsilon_m(s) \quad (9)$$

where  $s \in [s_l, s_{l+1})$ , (8) and (9) represent predictor and filter.  $L(s)$ ,  $\hat{\mathbf{A}}(s)$ ,  $K(s)$  are parameters to be selected in Subsect. 3.2. Due to it is impossible to calculate precise error covariance for (7) owing to the existing of the uncertain  $F(s)$ . We make estimation error as  $\tilde{X}(s|s) = X(s) - \hat{X}(s|s)$ ,  $\tilde{X}(s|s-1) = X(s) - \hat{X}(s|s-1)$ . Our goal is to find upper bounds of error covariance and are minimized.  $\bar{I}(s)$ ,  $\bar{A}(s)$  satisfy the following inequalities

$$\mathbb{E}[\tilde{X}(s|s)\tilde{X}^T(s|s)] \leq \bar{I}(s) \quad (10)$$

$$\mathbb{E}[\tilde{X}(s|s-1)\tilde{X}^T(s|s-1)] \leq \bar{A}(s) \quad (11)$$

### 3.2 Event-Based Robust Filter Parameters Design

According to the above analysis, we strive to find parameter  $L(s)$ ,  $\hat{\mathbf{A}}(s)$ ,  $K(s)$  which satisfy (10), (11).  $\tilde{\xi}(s)$  is denoted by  $\tilde{\xi}(s) = \xi(s) - \bar{\xi}$ . The innovation sequence has the format of  $\epsilon(s) = y(s) - \bar{\xi}\mathbf{C}(s)\hat{X}(s|s-1)$ . Then  $\epsilon(s)$  and  $\epsilon_m(s)$  can be represented by

$$\epsilon(s) = \tilde{\xi}(s)\mathbf{C}(s)X(s) + \bar{\xi}\mathbf{C}(s)\tilde{X}(s|s-1) + v(s) \quad (12)$$

$$\epsilon_m(s) = \epsilon(s) + d(s) \quad (13)$$

Substituting  $\epsilon_m(s)$  into event-based robust filter yields

$$\begin{aligned} \tilde{X}(s|s) &= (I - \bar{\xi}L(s)\mathbf{C}(s))\tilde{X}(s|s-1) - \tilde{\xi}(s)L(s)\mathbf{C}(s)\tilde{X}(s|s-1) \\ &\quad - \tilde{\xi}(s)L(s)\mathbf{C}(s)\hat{X}(s|s-1) - L(s)v(s) - L(s)d(s) \end{aligned} \quad (14)$$

$$\begin{aligned} \tilde{X}(s+1|s) &= (\mathbf{A}(s) + \Delta\mathbf{A}(s) - \bar{\xi}K(s)\mathbf{C}(s))\tilde{X}(s|s-1) \\ &\quad + (\mathbf{A}(s) + \Delta\mathbf{A}(s) - \hat{\mathbf{A}}(s))\hat{X}(s|s-1) \\ &\quad - \tilde{\xi}(s)K(s)\mathbf{C}(s)\tilde{X}(s|s-1) - \tilde{\xi}(s)K(s)\mathbf{C}(s)\hat{X}(s|s-1) \\ &\quad + W(s) - K(s)v(s) - K(s)d(s) \end{aligned} \quad (15)$$

The new augmented vector are as following

$$\Theta(s) = [\tilde{X}(s|s) \ \hat{X}(s|s)]^T, \Omega(s) = [\tilde{X}(s|s-1) \ \hat{X}(s|s-1)]^T,$$

And we have

$$\begin{aligned}\Theta(s) &= \bar{C}(s)\Omega(s) + \bar{M}_1(s)O(s)\bar{N}(s)\Omega(s) \\ \tilde{\xi}(s)\bar{C}_e(s)\Omega(s) &+ \bar{L}(s)v(s) + \bar{L}(s)d(s)\end{aligned}\quad (16)$$

$$\begin{aligned}\Omega(s+1) &= \bar{A}(s)\Omega(s) + \bar{M}(s)O(s)\bar{N}(s)\Omega(s) \\ + \tilde{\xi}(s)\bar{K}_c(s)\Omega(s) &+ [I, 0]^T W(s) + \bar{K}(s)v(s) + \bar{K}(s)d(s)\end{aligned}\quad (17)$$

where

$$\begin{aligned}\bar{C}(s) &= \begin{bmatrix} I - \bar{\xi}L(s)\mathbf{C}(s) & 0 \\ \bar{\xi}L(s)\mathbf{C}(s) & I \end{bmatrix}, \bar{C}_e(s) = \begin{bmatrix} -L(s)\mathbf{C}(s) & -L(s)\mathbf{C}(s) \\ L(s)\mathbf{C}(s) & L(s)\mathbf{C}(s) \end{bmatrix}, \\ \bar{A}(s) &= \begin{bmatrix} \mathbf{A}(s) - \bar{\xi}K(s)\mathbf{C}(s) & \mathbf{A}(s) - \hat{\mathbf{A}}(s) \\ \bar{\xi}K(s)\mathbf{C}(s) & \hat{\mathbf{A}}(s) \end{bmatrix}, \bar{K}_c(s) = \begin{bmatrix} -K(s)\mathbf{C}(s) & -K(s)\mathbf{C}(s) \\ K(s)\mathbf{C}(s) & K(s)\mathbf{C}(s) \end{bmatrix}, \\ \bar{M}(s) &= \begin{bmatrix} M(s) \\ 0 \end{bmatrix}, \bar{M}_1(s) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \bar{L}(s) = \begin{bmatrix} -L(s) \\ L(s) \end{bmatrix}, \bar{N}(s) = \begin{bmatrix} N(s) & N(s) \end{bmatrix}, \\ \bar{K}(s) &= \begin{bmatrix} -K(s) \\ K(s) \end{bmatrix}, \mathbf{Q}(s) = \begin{bmatrix} Q(s) & 0 \\ 0 & Q^f(s) \end{bmatrix}\end{aligned}$$

We find  $\bar{\Phi}(s)$ ,  $\bar{\Psi}(s)$  according to  $\tilde{\Phi}(s)$ ,  $\tilde{\Psi}(s)$  which satisfy the following inequality

$$[I, 0]\tilde{\Gamma}(s) \begin{bmatrix} I \\ 0 \end{bmatrix} \leq \bar{\Gamma}(s) \quad (18)$$

$$[I, 0]\tilde{\Lambda}(s) \begin{bmatrix} I \\ 0 \end{bmatrix} \leq \bar{\Lambda}(s) \quad (19)$$

where  $\tilde{\Gamma}(s) = \mathbb{E}[\Theta(s)\Theta^T(s)]$ ,  $\tilde{\Lambda}(s) = \mathbb{E}[\Omega(s)\Omega^T(s)]$ .

**Lemma 1** [23].  $c > 0$  is a scalar,  $a, b \in \mathbb{R}^n$  are arbitrary vectors, then

$$ab^T + ba^T \leq caa^T + c^{-1}bb^T \quad (20)$$

**Lemma 2** [8].  $A, M, N, O$  are matrices with suitable dimensions.  $OO^T \leq I$ . Assuming  $w$  is a symmetric positive-definite matrix and  $\iota$  is a random positive constant satisfy  $\iota^{-1}I - NXN^T > 0$ , then

$$(A + MON)w(A + MON)^T \leq A(w^{-1} - \iota N^T N)^{-1}A^T + \iota^{-1}MM^T \quad (21)$$

**Lemma 3** [22].  $s \in [0, N]$ , assume that  $X = X^T \geq 0$ ,  $Y = Y^T \geq 0$ ,  $\zeta_s(\cdot)$ :  $\mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$ . If

$$\zeta_s(X) \leq \zeta_s(Y), \forall X \leq Y \quad (22)$$

hence the solutions  $W_{s+1}$  and  $M_{s+1}$  of difference equations as follows

$$W_{s+1} = \Pi(W_s), W_{s+1} \leq \Pi(W_s), M_0 = W_0 \quad (23)$$

satisfy

$$M_{s+1} \leq W_{s+1} \quad (24)$$

**Corollary 1.** Suppose that  $\mu_1(s)$  is a positive constant, which satisfy  $\mu_1^{-1}(s)I - \bar{N}(s)\tilde{\Lambda}(s)\bar{N}^T(s) > 0$ , then

$$\begin{aligned}\tilde{\Gamma}(s) &\leq (1 + \kappa)\bar{C}(s)(\tilde{\Lambda}^{-1}(s) - \mu_1(s)\bar{N}^T(s)\bar{N}(s))^{-1}\bar{C}^T(s) \\ &\quad + \bar{\xi}(1 - \bar{\xi})\bar{C}_e(s)\bar{\Lambda}(s)\bar{C}_e^T(s) + (1 + \kappa^{-1})\bar{L}(s)\delta I\bar{L}^T(s) \\ &\quad (1 - 2\beta(s))\bar{L}(s)R(s)\bar{L}^T(s)\end{aligned}\quad (25)$$

$$\begin{aligned}\tilde{\Lambda}(s+1) &\leq (1 + \nu)\bar{A}(s)(\tilde{\Lambda}^{-1}(s) - \mu_1(s)\bar{N}^T(s)\bar{N}(s))^{-1}\bar{A}^T(s) \\ &\quad + (1 + \nu)\mu_1^{-1}(s)\bar{M}(s)\bar{M}^T(s) + \bar{\xi}(1 - \bar{\xi})\bar{K}_c(s)\bar{\Lambda}(s)\bar{K}_c^T(s) \\ &\quad + (1 + \nu^{-1})\bar{K}(s)\delta I\bar{K}^T(s) + (1 - 2\beta(s))\bar{K}(s)R(s)\bar{K}^T(s) + [I, 0]\mathbf{Q}(s)[I, 0]^T\end{aligned}\quad (26)$$

*Proof.* Calculating the covariance of  $\Theta$  and  $\Omega$  separately

$$\begin{aligned}\tilde{\Gamma}(s) &= [\bar{C}(s) + \bar{M}_1(s)O(s)\bar{N}(s)]\tilde{\Lambda}(s)[\bar{C}(s) + \bar{M}_1(s)O(s)\bar{N}(s)]^T \\ &\quad + \bar{\xi}(1 - \bar{\xi})\bar{C}_e(s)\tilde{\Psi}(s)\bar{C}_e^T(s) + \bar{L}(s)R(s)\bar{L}^T(s) + \bar{L}(s)\mathbb{E}[d(s)d^T(s)]\bar{L}^T(s) \\ &\quad + \bar{L}(s)\mathbb{E}[d(s)v^T(s)]\bar{L}^T(s) + \bar{L}(s)\mathbb{E}[v(s)d^T(s)]\bar{L}^T(s) \\ &\quad + [\bar{C}(s) + \bar{M}_1(s)O(s)\bar{N}(s)]E[\Omega(s)d^T(s)]\bar{L}^T(s) \\ &\quad + \bar{L}(s)\mathbb{E}[d(s)\Omega^T(s)][\bar{C}(s) + \bar{M}_1(s)O(s)\bar{N}(s)]^T\end{aligned}\quad (27)$$

Applying Lemma 1 there exists a constant  $\kappa > 0$ , then

$$\begin{aligned}&[\bar{C}(s) + \bar{M}_1(s)O(s)\bar{N}(s)]\mathbb{E}[\Omega(s)d^T(s)]\bar{L}^T(s) \\ &\quad + \bar{L}(s)\mathbb{E}[d(s)\Omega^T(s)][\bar{C}(s) + \bar{M}_1(s)O(s)\bar{N}(s)]^T \\ &\leq \kappa[\bar{C}(s) + \bar{M}_1(s)O(s)\bar{N}(s)]\tilde{\Psi}(s)[\bar{C}(s) + \bar{M}_1(s)O(s)\bar{N}(s)]^T \\ &\quad + \kappa^{-1}\bar{L}(s)\mathbb{E}[d(s)d^T(s)]\bar{L}^T(s)\end{aligned}\quad (28)$$

Applying Lemma 2 there exist a constant  $\mu_1(s) > 0$ , then

$$\begin{aligned}[\bar{C}(s) + \bar{M}_1(s)O(s)\bar{N}(s)]\tilde{\Psi}(s)[\bar{C}(s) + \bar{M}_1(s)O(s)\bar{N}(s)]^T \\ \leq \bar{C}(s)(\tilde{\Lambda}^{-1}(s) - \mu_1(s)\bar{N}^T(s)\bar{N}(s))\bar{C}^T(s)\end{aligned}\quad (29)$$

According to (5)

$$\mathbb{E}[v(s)d^T(s)] = -\beta(s)R(s) \quad (30)$$

where if  $s = s_l$ ,  $\beta(s) = 0$ ; otherwise,  $\beta(s) = 1$ .

In addition,  $d(s)$  would be zero if (5) is satisfied

$$d(s)d^T(s) \leq \delta \quad (31)$$

By operation

$$d(s)d^T(s) \leq \|d(s)\|^2 I = d(s)d^T(s)I \leq \delta I \quad (32)$$

and as a result

$$\mathbb{E}[d(s)d^T(s)] \leq \delta I \quad (33)$$

Together (28), (29), (30), (33) yields (25).

Similarly, where  $\nu > 0$  is a scalar, through mathematical operations we get (26).

**Theorem 1.**  $\mu_1(s)$  is a positive scalar and  $\Lambda(s)$  is a symmetric positive-definite matrix. If  $\mu_1(s)I - \bar{N}(s)\Lambda(s)\bar{N}^T(s) > 0$ , then  $\tilde{\Gamma}(s) \leq \Gamma(s)$  and  $\tilde{\Lambda}(s) \leq \Lambda(s)$ , where  $\tilde{\Gamma}(s), \tilde{\Lambda}(s)$  defined in Corollary 1. And  $\Gamma(s)$  and  $\Lambda(s)$  can be obtained by following equations

$$\begin{aligned} \Gamma(s) = & (1 + \kappa)\bar{C}(s)(\Psi^{-1}(s) - \mu_1(s)\bar{N}^T(s)\bar{N}(s))^{-1}\bar{C}^T(s) \\ & + \bar{\xi}(1 - \bar{\xi})\bar{C}_e(s)\Lambda(s)\bar{C}_e^T(s) + (1 + \kappa^{-1})\bar{L}(s)\delta I\bar{L}^T(s)(1 - 2\beta(s))\bar{L}(s)R(s)\bar{L}^T(s) \end{aligned} \quad (34)$$

$$\begin{aligned} \Lambda(s+1) = & (1 + \nu)\bar{A}(s)(\Lambda^{-1}(s) - \mu_1(s)\bar{N}^T(s)\bar{N}(s))^{-1}\bar{A}^T(s) \\ & + (1 + \nu)\mu_1^{-1}(s)\bar{M}(s)\bar{M}^T(s) + \bar{\xi}(1 - \bar{\xi})\bar{K}_c(s)\Lambda(s)\bar{K}_c^T(s) \\ & + (1 + \nu^{-1})\bar{K}(s)\delta I\bar{K}^T(s) + (1 - 2\beta(s))\bar{K}(s)R(s)\bar{K}^T(s) + [I, 0]\mathbf{Q}(s)[I, 0]^T \end{aligned} \quad (35)$$

*Proof.* According to Lemma 3 and Corollary 1 we derive Theorem 1 here we omit its proof for page limitation.

*Remark 3.* When calculating  $\tilde{\Gamma}(s)$  and  $\tilde{\Lambda}(s)$ , according to missing observations process  $\xi(s)$ , we derive that  $\mathbb{E}[\xi(s)] = 0$ ,  $\text{cov}[\xi(s)] = \bar{\xi}(1 - \bar{\xi})$ . Moreover,  $\beta(s)$  establishes a connection between  $d(s)$  and  $v(s)$ . However, Theorem 1 barely proves there exist upper bounds.

**Theorem 2.** Let  $\mu_1(s)$  be a positive scalar, such that  $\mu_1^{-1}(s)I - \bar{N}(s)\Psi(s)\bar{N}^T(s) > 0$  and  $\mu_1^{-1}(s)I - N(s)P(s)N^T(s) > 0$  hold then

$$\Lambda(s) = \begin{bmatrix} \bar{A}(s) & 0 \\ 0 & P(s) - \bar{A}(s) \end{bmatrix} \quad (36)$$

$$\begin{aligned} P(s+1) = & \mathbf{A}(s)(P^{-1}(s) - \mu_1(s)N^T(s)N(s))^{-1}\mathbf{A}^T(s) \\ & + \mathbf{Q}(s) + \mu_1^{-1}(s)M(s)M^T(s) \end{aligned} \quad (37)$$

$$\begin{aligned} \bar{\Lambda}(s+1) = & (1 + \nu)(\mathbf{A}(s) - \bar{\xi}K(s)\mathbf{C}(s))(\bar{A}(s) + \bar{A}(s)N^T(s)\sum^{-1}(s)N(s)\bar{A}(s)) \\ & (\mathbf{A}(s) - \bar{\xi}K(s)\mathbf{C}(s))^T + \mathbf{Q}(s) + (1 + \nu)\mu_1^{-1}(s)M(s)M^T(s) \\ & + \bar{\xi}(1 - \bar{\xi})K(s)\mathbf{C}(s)P(s)\mathbf{C}^T(s)K^T(s) + (1 + \nu^{-1})K(s)\delta I\bar{K}^T(s) \\ & + (1 - 2\beta(s))K(s)R(s)\bar{K}^T(s) \end{aligned} \quad (38)$$

The parameters of event-based robust filter are as following:

$$Y(s) = \mu_1^{-1}(s)I - N(s)\bar{\Lambda}(s)N^T(s) \quad (39)$$

$$\tilde{Y}(s) = \mu_1^{-1}(s)I - N(s)P(s)N^T(s) \quad (40)$$

$$\hat{\mathbf{A}}(s) = \mathbf{A}(s) + (\mathbf{A}(s) - \bar{\xi}K(s)\mathbf{C}(s)) \quad (41)$$

$$\bar{\Lambda}(s)N^T(s)Y^{-1}(s)N(s)$$

$$L(s) = \bar{\xi}(1 + \kappa)\bar{A}(s)[I + N^T(s)\tilde{Y}^{-1}(s)N(s)\bar{A}(s)]\mathbf{C}^T(s)S^{-1}(s) \quad (42)$$

$$K(s) = \bar{\xi}(1 + \nu)\mathbf{A}(s)\bar{\Lambda}(s)[I + N^T(s)Y^{-1}(s)N(s)\bar{\Lambda}(s)]\mathbf{C}^T(s)T^{-1}(s) \quad (43)$$

$$\begin{aligned} S(s) = & \bar{\xi}^2(1+\kappa)\mathbf{C}(s)\bar{\Lambda}(s)[I+N^T(s)\tilde{Y}^{-1}(s)N(s)\bar{\Lambda}(s)]\mathbf{C}^T(s) \\ & +\bar{\xi}(1-\bar{\xi})\mathbf{C}(s)P(s)\mathbf{C}^T(s)+(1+\kappa^{-1})\delta I+(1-2\beta(s))R(s) \end{aligned} \quad (44)$$

$$\begin{aligned} T(s) = & \bar{\xi}^2(1+\nu)\mathbf{C}(s)\bar{\Lambda}(s)[I+N^T(s)Y^{-1}(s)N(s)\bar{\Lambda}(s)]\mathbf{C}^T(s) \\ & +\bar{\xi}(1-\bar{\xi})\mathbf{C}(s)P(s)\mathbf{C}^T(s)+(1+\nu^{-1})\delta I+(1-2\beta(s))R(s) \end{aligned} \quad (45)$$

*Proof.* Obviously, for  $s=0$ , (36) holds. Define  $\bar{\Gamma}(s+1)=[I, 0]\Gamma(s+1)\begin{bmatrix} I \\ 0 \end{bmatrix}$ , assume that (36) holds for  $k$ , then

$$\begin{aligned} \bar{\Gamma}(s+1) = & (1+\kappa)(I-\bar{\xi}L(s)\mathbf{C}(s))\bar{\Lambda}(s)(I-\bar{\xi}L(s)\mathbf{C}(s))^T \\ & +(1+\kappa)((I-\bar{\xi}L(s)\mathbf{C}(s))\bar{\Lambda}(s)N^T(s)\tilde{Y}^{-1}(s)N(s)\bar{\Lambda}(s) \\ & (I-\bar{\xi}L(s)\mathbf{C}(s))^T+\bar{\xi}(1-\bar{\xi})L(s)\mathbf{C}(s)P(s)\mathbf{C}^T(s)L^T(s) \\ & +(1+\kappa^{-1})L(s)\delta IL^T(s)+(1-2\beta(s))L(s)R(s)L^T(s) \end{aligned} \quad (46)$$

For the sake of determining  $L(s)$  we take its first variation to the trace of (46) and set it zero, yields

$$\begin{aligned} \frac{\partial \text{tr}(\bar{\Gamma}(s))}{\partial L(s)} = & 2(1+\kappa)(I-\bar{\xi}L(s)\mathbf{C}(s))\bar{\Lambda}(s)(-\bar{\xi}\mathbf{C}(s))^T \\ & +2(1+\kappa)(I-\bar{\xi}L(s)\mathbf{C}(s))\bar{\Psi}(s)N^T(s)\tilde{Y}^{-1}(s) \\ & N(s)\bar{\Lambda}(s)(-\bar{\xi}\mathbf{C}(s))^T+2\bar{\xi}(1-\bar{\xi})L(s)\mathbf{C}(s)P(s)\mathbf{C}^T(s) \\ & +2(1+\kappa^{-1})\delta L(s)+2(1-2\beta(s))L(s)R(s)=0 \end{aligned} \quad (47)$$

then  $L(s)$  is obtained. Similarly  $\bar{\Lambda}(s+1)=[I, 0]\Lambda(s+1)\begin{bmatrix} I \\ 0 \end{bmatrix}$  decide  $\hat{\mathbf{A}}(s)$  and  $K(s)$ .

$$\begin{aligned} \bar{\Lambda}(s+1) = & (1+\nu)((\mathbf{A}(s)-\hat{\mathbf{A}}(s))(P(s)-\bar{\Lambda}(s))(\mathbf{A}(s)-\hat{\mathbf{A}}(s))^T \\ & +(\mathbf{A}(s)-\bar{\xi}K(s)\mathbf{C}(s))\bar{\Lambda}(s)(\mathbf{A}(s)-\bar{\xi}K(s)\mathbf{C}(s))^T \\ & +[\mathbf{A}(s)P(s)-\bar{\xi}K(s)\mathbf{C}(s)\bar{\Lambda}(s)-\hat{\mathbf{A}}(s)(P(s)-\bar{\Lambda}(s))] \\ N^T(s)\tilde{Y}^{-1}(s)N(s)[\mathbf{A}(s)P(s)-\bar{\xi}K(s)\mathbf{C}(s)\bar{\Lambda}(s)-\hat{\mathbf{A}}(s)(P(s)-\bar{\Lambda}(s))]^T \\ & +(1+\nu)\mu_1^{-1}(s)M(s)M^T(s)+\bar{\xi}(1-\bar{\xi})K(s)\mathbf{C}(s)P(s)\mathbf{C}^T(s)K^T(s) \\ & +(1+\nu^{-1})K(s)\delta IK^T(s)+(1-2\beta(s))K(s)R(s)K^T(s)+\mathbf{Q}(s) \end{aligned} \quad (48)$$

In order to determine  $\hat{\mathbf{A}}$  and  $K(s)$ , we take their first variation to the trace of (48) and set them zero separately, yields

$$\begin{aligned} \frac{\partial \text{tr}(\bar{\Lambda}(s+1))}{\partial \hat{\mathbf{A}}(s)} = & (\hat{\mathbf{A}}(s)-\mathbf{A}(s))(P(s)-\Lambda(s)) \\ & +(\mathbf{A}(s)P(s)-\bar{\xi}K(s)\mathbf{C}(s)\bar{\Lambda}(s)-\hat{\mathbf{A}}(s)(P(s)-\Lambda(s))) \\ N^T(s)\tilde{Y}^{-1}(s)N(s)(\Lambda(s)-P(s)) = 0 \end{aligned} \quad (49)$$

$$\begin{aligned} \frac{\partial \text{tr}(\bar{\Lambda}(s+1))}{\partial K(s)} = & 2(1+\nu)(\mathbf{A}(s)-\bar{\xi}K(s)\mathbf{C}(s))\bar{\Lambda}(s) \\ & (I+N^T(s)Y^{-1}(s)N(s)\bar{\Lambda}(s))(-\bar{\xi}\mathbf{C}(s))^T \\ & +2\bar{\xi}(1-\bar{\xi})K(s)\mathbf{C}(s)P(s)\mathbf{C}^T(s) \\ & +(1+\nu^{-1})\delta K(s)+2(1-2\beta(s))K(s)R(s)=0 \end{aligned} \quad (50)$$

Utilizing

$$\begin{aligned} N^T(s)\tilde{Y}^{-1}(s)N(s) &= N^T(s)Y^{-1}(s)N(s) \\ (I - (\Lambda(s) - P(s))N^T(s)\tilde{Y}^{-1}(s)N(s))^{-1} \end{aligned} \quad (51)$$

After calculation,  $\hat{\mathbf{A}}(s)$  and  $K(s)$  can be derived. Then

$$\hat{\mathbf{A}}(s) - \mathbf{A}(s) = (\mathbf{A}(s) - \bar{\xi}K(s)\mathbf{C}(s))\bar{\Lambda}(s)N^T(s)Y^{-1}(s)N(s) \quad (52)$$

Combing (48) with (52) yields (38). Define  $\bar{P}(s) = \mathbb{E}[X(s)X^T(s)]$ , by using Lemma 2.

$$\begin{aligned} \bar{P}(s+1) &\leq \mathbf{A}(\bar{P}^{-1}(s) - \mu_1(s)N^T(s)N(s))^{-1}\mathbf{A}^T(s) \\ &\quad + \mathbf{Q} + \mu_1(s)M(s)M^T(s) \end{aligned} \quad (53)$$

We obtain (37) by utilizing Lemma 3. Combining (35), (37) with (38) we can prove  $\Lambda(s+1)$  fulfills the structure of (36). According to mathematical induction (36) holds for  $s \in [0, N]$ .

*Remark 4.* By designing parameters of estimator, we minimize the upper bound proposed before. (36) gives the structure of the upper bound  $\Psi(s)$ . (38) takes  $\bar{\eta}$ ,  $\delta$ ,  $\beta(s)$  into account.

## 4 Simulation Example

For the sake of verifying the validity of the algorithm proposed, we choose the vehicle lateral dynamic system as the simulation model [21], which is formulated as follows

$$\left\{ \begin{array}{l} \begin{bmatrix} \beta_s(s+1) \\ r_y(s+1) \end{bmatrix} = \begin{bmatrix} -\frac{c_{av}+c_{ah}}{mv_{ref}} & \frac{l_h c_{ah} - l_v c_{av}}{mv_{ref}^2} - 1 \\ \frac{l_h c_{ah} - l_v c_{av}}{I_Z} & -\frac{l_v^2 c_{av} + l_h^2 c_{ah}}{I_Z v_{ref}} \end{bmatrix} \begin{bmatrix} \beta_s(s) \\ r_y(s) \end{bmatrix} + \begin{bmatrix} \frac{c_{av}}{mv_{ref}} \\ \frac{l_v c_{av}}{I_Z} \end{bmatrix} \delta_l(s) \\ \quad + Df(s) + w(s) \\ \\ \begin{bmatrix} \alpha_y(s) \\ r_y(s) \end{bmatrix} = \begin{bmatrix} -\frac{c_{av}+c_{ah}}{m} & \frac{l_h c_{ah} - l_v c_{av}}{mv_{ref}} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_s(s) \\ r_y(s) \end{bmatrix} + \begin{bmatrix} \frac{c_{av}}{m} \\ 0 \end{bmatrix} \delta_l(s) + v(s) \end{array} \right. \quad (54)$$

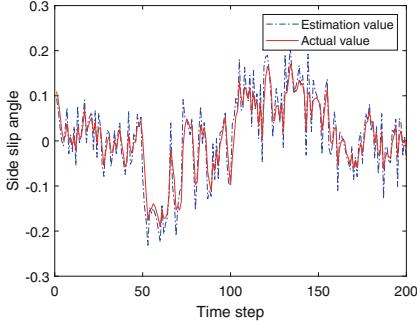
where  $\beta_s(s)$  is vehicle side slip angle,  $r_y(s)$  is the yaw rate,  $m$  is total mass,  $v_{ref}$  is vehicle longitude velocity,  $c_{av}$  is front tire cornering stiffness,  $c_{ah}$  is rear tire cornering stiffness,  $\delta_l(s)$  is vehicle steering angle,  $I_Z$  is moment of inertia about the z-axis,  $l_h$  is distance from the CG to the rear axle,  $l_v$  is distance from the CG to the front axle.  $w(s)$  and  $v(s)$  are uncorrelated zero-mean white noise.  $f(s)$  is the fault in steering angle actuator. Here the speed is 50 km/h and sampling interval is 0.1 s.

$$\begin{aligned} A(s) &= \begin{bmatrix} 0.6333 & -0.0672 \\ 2.0570 & 0.6082 \end{bmatrix}, C(s) = \begin{bmatrix} -152.7568 & 1.2493 \\ 0 & 1 \end{bmatrix}, M(s) = \begin{bmatrix} 0.2 & 0.1 \end{bmatrix}^T, \\ O(s) &= \sin(0.5k), N(s) = \begin{bmatrix} 0.1 & 0.1 \end{bmatrix} \end{aligned} \quad (55)$$

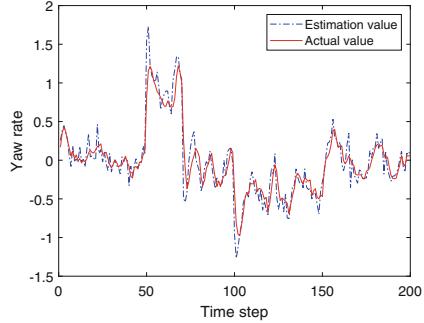
The control input  $\delta_l(s) = 0$  and the fault is described by

$$f(s) = \begin{cases} 0.18, & 50 \leq s \leq 50; \\ -0.1, & 100 \leq s \leq 150; \\ 0, & \text{else;} \end{cases} \quad (56)$$

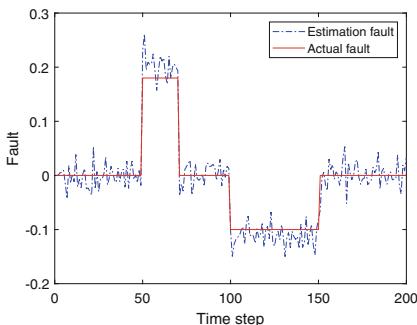
We assume that  $x(0) = [0.1 \ 0]^T$ ,  $P(0) = 0.01I_2$ ,  $Q(s) = 0.01I_2$ ,  $Q^f(s) = 0.01I_2$ ,  $R(s) = 0.01I_2$ ,  $\xi = 0.9$ ,  $\delta = 0.5$ ,  $\kappa = 2$ ,  $\nu = 3$ ,  $\mu_1(s) = 0.2$ .



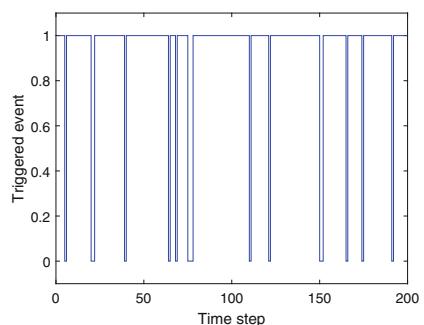
**Fig. 1.** Actual side slip angle and estimation



**Fig. 2.** Actual yaw rate and estimation

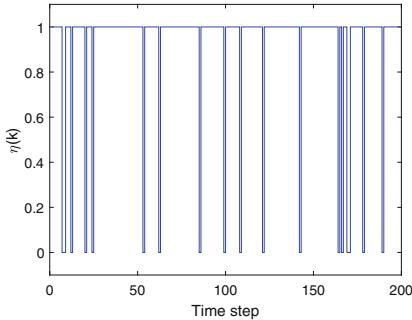


**Fig. 3.** Actuator fault and estimation

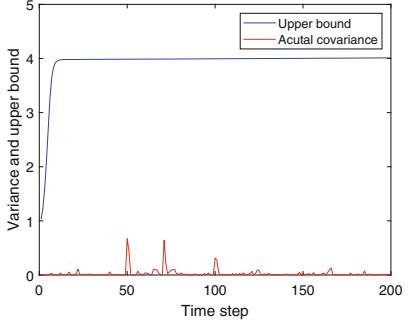


**Fig. 4.** Event triggered strategy

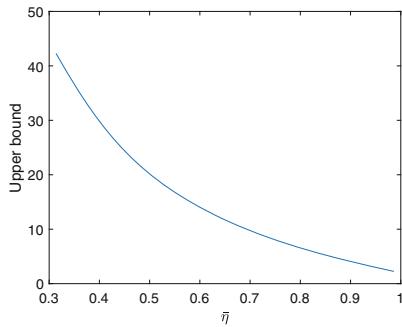
Figure 1, 2, 3 give the actual side slip angle, yaw rate and steering angle actuator fault and their estimations. It is shown that the filter is quite valid. Figure 4, 5 show the effect of the event-triggered mechanism and missing observations phenomenon. Furthermore, values 1 and 0 indicate whether the measurements are transmitted to the estimator or not. The trace of actual error covariance and its upper bound are drawn in Fig. 6. We control  $\bar{\eta}$  as the only variable, connection between upper bounds and probability of missing phenomenon gives in Fig. 7.



**Fig. 5.** Missing observations phenomenon



**Fig. 6.** Trace of error covariance and its upper bound



**Fig. 7.** The upper bound under  $0.3 \leq \bar{\eta} \leq 1$

## 5 Conclusion

In this paper, the event-based robust state and fault estimator was proposed for stochastic linear system with missing observations and uncertainty. Based on the augmented state method and two-stage robust event-based estimators, the filter algorithm was derived. Meanwhile, uncertainty, missing observation and event triggered mechanism were all considered in the filter algorithm. By utilizing discrete Riccati difference equation, the upper bounds of the estimation error covariance were obtained. Finally, the simulation of vehicle lateral dynamic system demonstrated the usefulness of the presented filter.

**Acknowledgements.** This work was supported by National Natural Science Foundation (NNSF) of China under Grant 61803330.

## References

1. Wang, H., Daley, S.: Actuator fault diagnosis: an adaptive observer-based technique. *IEEE Trans. Autom. Control* **41**(7), 1073–1078 (1996)

2. Jiang, B., Staroswiecki, M., Cocquempot, V.: Fault estimation in nonlinear uncertain system using robust/sliding-mode observers. *IEE Proc.-Control Theory Appl.* **151**(1), 29–37 (2004)
3. Martinez, G.R., Diop, S.V.: Diagnosis of nonlinear system using an unknown-input observer: an algebraic and differential approach. *IEE Proc.-Control Theory Appl.* **151**(1), 130–135 (2004)
4. Gao, Z.W., Ding, S.X.: Actuator fault robust estimation and fault-tolerant control for a class of nonlinear descriptor system. *Automatica* **43**(5), 912–920 (2007)
5. Liu, M., Cao, X.B., Shi, P.: Fuzzy-model-based fault-tolerant design for nonlinear stochastic system against simultaneous sensor and actuator faults. *IEEE Trans. Fuzzy Syst.* **21**(5), 789–799 (2012)
6. Zhu, J.W., Yang, G.H., Wang, H., Wang, F.L.: Fault estimation for a class of nonlinear system based on intermediate estimator. *IEEE Trans. Autom. Control* **61**(9), 2518–2524 (2015)
7. Zhe, D., Zheng, Y.: Finite-horizon robust Kalman filtering for uncertain discrete time-varying system with uncertain-covariance white noises. *IEEE Sig. Process. Lett.* **13**(8), 493–496 (2006)
8. Xie, L.H., Soh, Y.C., Souza, C.E.: Robust Kalman filtering for uncertain discrete-time system. *IEEE Trans. Autom. Control* **39**(6), 1310–1314 (1994)
9. Liu, Y., Chen, C.L.P., Wen, G.X., Tong, S.C.: Adaptive neural output feedback tracking control for a class of uncertain discrete-time nonlinear system. *IEEE Trans. Neural Netw.* **22**(7), 1162–1167 (2011)
10. Wang, Z.D., Yang, F.W., Ho, D.W.C., Liu, X.H.: Robust finite-horizon filtering for stochastic system with missing measurements. *IEEE Sig. Process. Lett.* **12**(6), 437–440 (2005)
11. Rezaei, H., Esfanjani, R.M., Sedaaghi, M.H.: Improved robust finite-horizon Kalman filtering for uncertain networked time-varying system. *Inf. Sci.* **293**, 263–274 (2015)
12. Mohamed, S., Nahavandi, S.: Robust finite-horizon Kalman filtering for uncertain discrete-time system. *IEEE Trans. Autom. Control* **57**(6), 1548–1552 (2012)
13. Wang, S.Y., Fang, H.J., Tian, X.G.: Event-based robust state estimator for linear time-varying system with uncertain observations and randomly occurring uncertainties. *J. Franklin Inst.-Eng. Appl. Math.* **354**(3), 1403–1420 (2017)
14. Miskowicz, M.: Send-on-delta concept: an event-based data reporting strategy. *Sensors* **6**(1), 49–63 (2006)
15. Tabuada, P.: Event-triggered real-time scheduling of stabilizing control tasks. *IEEE Trans. Autom. Control* **52**(9), 1680–1689 (2007)
16. Trimpe, S., Dandrea, R.: Event-based state estimation with variance-based triggering. *Conf. Decis. Control* **59**(12), 3266–3281 (2012)
17. Yue, D., Tian, E.G., Han, Q.L.: A delay system method for designing event-triggered controllers of networked control system. *IEEE Trans. Autom. Control* **58**(2), 475–481 (2012)
18. Han, D., Mo, Y.L., Wu, J.F., Weerakkody, S., Sinopoli, B., Shi, L.: Stochastic event-triggered sensor schedule for remote state estimation. *IEEE Trans. Autom. Control* **60**(10), 2661–2675 (2015)
19. Yang, C., Yang, W., Shi, H.B.: Communication-saving design by stochastic event triggers. *J. Franklin Inst.* **356**(17), 10532–10546 (2019)
20. Ding, B., Fang, H.J.: Fault prediction for nonlinear stochastic system with incipient faults based on particle filter and nonlinear regression. *ISA Trans.* **68**, 327–334 (2017)

21. Steven, X.D.: Model-Based Fault Diagnosis Techniques: Design Schemes, Algorithms, and Tools. Springer, Heidelberg (2008). (Chapter 3)
22. Theodor, Y., Shaked, U.: Robust discrete-time minimum-variance filtering. *IEEE Trans. Sig. Process.* **44**(2), 181–189 (1996)
23. Hu, J., Wang, Z.D., Gao, H.J., Stergioulas, L.K.: Extended Kalman filtering with stochastic nonlinearities and multiple missing measurements. *Automatica* **48**(9), 2007–2015 (2012)



# Sliding Mode Control for a Constant Force Suspension System

Yuxin Jia, Yingmin Jia<sup>(✉)</sup>, Kai Gong, Yao Lu, and Meng Duan

The Seventh Research Division and the Center for Information and Control,  
School of Automation Science and Electrical Engineering,  
Beihang University (BUAA), Beijing 100191, China  
[{yuxinjia,ymjia}@buaa.edu.cn](mailto:{yuxinjia,ymjia}@buaa.edu.cn)

**Abstract.** The constant force suspension system is the core device of suspended gravity compensation platform. It can provide a microgravity environment similar to space for test objects, and is of great significance to the development of ground testing and verification technology for space missions. In this paper, a sliding mode controller based on a new reaching law is designed for the dynamic model of a constant force suspension system. The simulation results show that, compared with the existing controller, the controller suppresses the overshoot and chattering while reducing the settling time, thereby improving the stability and practical value of the system.

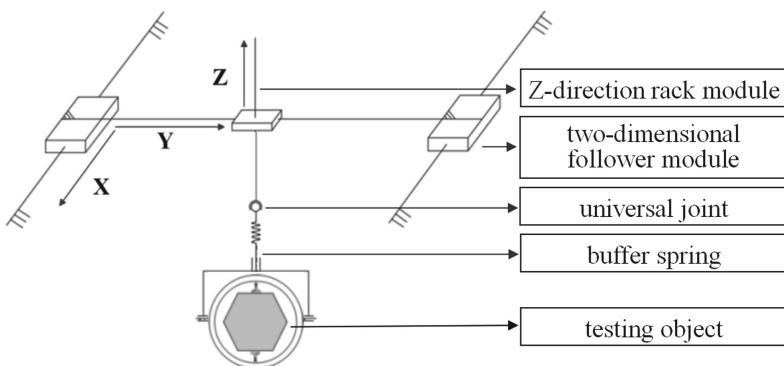
**Keywords:** Constant force suspension system · Dynamic model · Reaching law · Sliding mode controller

## 1 Introduction

Aerospace technology is the supporting technology for exploring and utilizing space resources, which embodies a country's comprehensive capabilities. According to the development strategy of China's aerospace technology, China's space station will be built around 2022, which puts forward higher requirements for the implementation ability of the space mission. Space missions are characterized by high risks and high costs. The execution steps is mainly divided into scheme design, ground verification, and mission implementation. Simulating the microgravity environment on the ground to test and verify the execution scheme is an important step to reduce the risk and improve the reliability of space mission [1–3]. In response to this strategic demand in the aerospace field, many scientific research institutions have conducted a lot of research.

The Ranger test system of the University of Maryland in the United States uses liquid buoyancy to counteract gravity, providing a microgravity environment for spacecraft ground testing[4]. In order to simulate the attitude movement of satellites in space, the U.S. Naval Postgraduate School developed a three-degree-of- freedom satellite simulator TASS using air floatation[5]. The European Space Agency (ESA) uses the aircraft modified on the basis of Airbus A300 to simulate

microgravity environment by parabolic flight, which has been used in astronaut training and space equipment testing for many times[6]. In addition, China's National Microgravity Laboratory adopts the method of free fall to simulate microgravity environment. In the microgravity tower with a height of more than 100m, it can provide a short-term microgravity environment of 3.5s for the test object[7]. Among the above methods, the liquid flotation method has a large damping, which is not suitable for the high-speed moving test objects. Generally, the air flotation method can only simulate the movement of five degrees of freedom, which limits the movement of the test object. The parabolic flight and free fall method can only simulate the microgravity environment for a short time, and the construction and maintenance costs are high[8]. Fortunately, there is a suspension method that can overcome these shortcomings, which counteracts the gravity of the test object by vertical tension. The mechanical structure of this method is simple and easy to establish in the laboratory. It has been applied to the field of spacecraft ground testing by many national scientific research institutions. In 2019, China's space agency (CNSA) adopted the method of suspension to counteract the partial gravity of the Mars Lander on the earth, so as to simulate the hover, descent and obstacle avoidance motion of the lander in the gravity environment of Mars. The constant force suspension system is the core device of the gravity compensation experiment platform established by the suspension method. Therefore, it is very meaningful to study the effective control strategy to ensure that the system output is constant.[9].



**Fig. 1.** Schematic diagram of constant force suspension system

In this paper, a new sliding mode controller is designed for the dynamic model of a constant force suspension system. Compared with the existing controller[9], while the chattering is suppressed, the settling time is shortened and the practical value of the system is improved.

## 2 Dynamic Model

The constant force suspension system studied in this paper consists of a Z-direction rack module, a two-dimensional follower module, a universal joint and a buffer spring, and its structural diagram is shown in Fig. 1[10]. The two-dimensional follower module includes X-direction and Y-direction linear modules. When the test object is moved by its own driving force, the two-dimensional follower module can drive the Z-direction rack module to quickly track the trajectory of the test object, so as to ensure that the buffer spring remain vertical.

In addition, a tension sensor is installed between the Z-direction rack module and the universal joint. The system controls the rack to move up and down according to the sensor feedback, thereby ensuring that the output force of the buffer spring can completely or partially counteract the gravity of the test object. The universal joint is used to ensure the displacement freedom of the test object. The physical meaning of the symbols in the system is shown in Table 1.

**Table 1.** The physical meaning of symbols in the system

$g$	Acceleration of gravity
$l_0$	The initial length of buffer spring
$l$	The length variation of buffer spring
$k$	Stiffness coefficient of buffer spring
$r$	Gear radius of Z-direction rack module
$T_\alpha$	Output torque of the motor in Z-direction rack module
$M$	The mass of test object
$m_1, m_2$	The mass of the gear and rack in Z-direction rack module
$m_x, m_y$	The load of the motor in X-direction and Y-direction linear modules
$\omega_x, \omega_y$	Orthogonal decomposition value of the swing angle between buffer spring and Z-direction rack module
$F_x, F_y$	Driving force of the motor in X-direction and Y-direction linear modules
$F_{0x}, F_{0y}, F_{0z}$	Driving force of test object

Through the Lagrange equation, the dynamic model of the constant force suspension system can be described by

$$(H_2 - H_1 H_2^{-T} H_3) \ddot{x} + (J_1 - H_1 H_2^{-T} J_2) \dot{x} + (N_1 - H_1 H_2^{-T} N_2) = U - H_1 H_2^{-T} \delta \quad (1)$$

The variables defined in (1) is as follows:

$$H_1 = \begin{bmatrix} \varphi_x & 0 & 0 \\ 0 & \varphi_y & 0 \\ 0 & 0 & \varphi_1 r^2 \end{bmatrix}, H_2 = \begin{bmatrix} \omega_x & d & 0 \\ \omega_y & 0 & d \\ -r & 0 & 0 \end{bmatrix}, H_3 = \begin{bmatrix} \hat{\omega} & d\omega_x & d\omega_y \\ d\omega_x & d^2 & 0 \\ d\omega_y & 0 & d^2 \end{bmatrix}$$

$$J_1 = \begin{bmatrix} 2\dot{\omega}_x & 0 & 0 \\ 2\dot{\omega}_y & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, J_2 = \begin{bmatrix} 2\omega_{xy} & 0 & 0 \\ 2d\dot{\omega}_x & 0 & 0 \\ 2d\dot{\omega}_y & 0 & 0 \end{bmatrix}$$

$$N_1 = [0 \ 0 \ gr\varphi_2]^T, N_2 = [l\phi \ 0 \ 0]^T, U = \frac{1}{M} [F_x \ F_y \ T_\alpha]^T, \delta = \frac{1}{M} [F_{0z} \ F_{0x} \ F_{0y}]^T$$

$$x = [l \ \omega_x \ \omega_y]^T, \varphi_x = \frac{M + m_x}{M}, \varphi_y = \frac{M + m_y}{M}, \varphi_1 = \frac{0.5m_1 + m_2 + M}{M}$$

$$\varphi_2 = \frac{m_2 + M}{M}, d = l + l_0 + l_1, \phi = \frac{k}{M}, \widehat{\omega} = \omega_x^2 + \omega_y^2 + 1, \omega_{xy} = \omega_x\dot{\omega}_x + \omega_y\dot{\omega}_y$$

where  $x$  is the state vector of the system,  $U$  is the input vector,  $\delta$  is the driving force vector of the test object, which is regarded as system interference.  $l_1 = Mg/k$ ,  $d = l + l_0 + l_1$  is the actual length of the buffer spring. Since the research in this paper is that the output force of the system completely counteract the gravity of the test object, define  $l = 0$ . If  $x_d$  is defined as the expected value of the state vector, then the control objective of the constant force suspension system is to design the input vector  $U$ , in order that the state vector  $x$  approaches  $x_d$ .

### 3 Controller Design

In this section, a new reaching law is proposed, and then a new sliding mode controller is designed for the dynamic model of the constant force suspension system. Define the switching surface as

$$s = \alpha e + \beta \operatorname{sgn}(\dot{e})^\gamma = 0 \quad (2)$$

where  $s = [s_1 \ s_2 \ s_3]^T$ ,  $\operatorname{sgn}(x)^\gamma = [|x_1|^\gamma \ \operatorname{sign}(x_1) \cdots |x_n|^\gamma \ \operatorname{sign}(x_n)]^T$ .  $e = x - x_d$  is the state deviation of the system,  $\alpha = \operatorname{diag}(\alpha_1 \ \alpha_2 \ \alpha_3)$  and  $\beta = \operatorname{diag}(\beta_1 \ \beta_2 \ \beta_3)$  are positive definite diagonal matrices,  $\gamma$  is a constant and satisfies  $1 < \gamma < 2$ . The new reaching law designed in this paper is

$$\dot{s} = -\Gamma \beta \operatorname{diag}(|\dot{e}|^{\gamma-1})[q_1 s + q_2 \operatorname{sgn}(s)^\rho] \quad (3)$$

where,  $\Gamma = \operatorname{diag}(\gamma \ \gamma \ \gamma)$ ,  $q_1 = \operatorname{diag}(q_{11} \ q_{12} \ q_{13})$  and  $q_2 = \operatorname{diag}(q_{21} \ q_{22} \ q_{23})$  are positive definite diagonal matrices,  $\rho$  is a constant and satisfies  $0 < \rho < 1$ . Based on the dynamic model (1), switching surface (2) and reaching law (3), the controller designed in this paper is

$$\begin{aligned} U &= U_0 + U_1 \\ U_0 &= J\dot{x} + N + H\ddot{x}_d \\ U_1 &= -H(q_1 s + q_2 \operatorname{sgn}(s)^\rho) - H\Gamma^{-1}\beta^{-1}\alpha \operatorname{sgn}(\dot{e})^{2-\gamma} \end{aligned} \quad (4)$$

where  $H = H_2 - H_1 H_2^{-T} H_3$ ,  $J = J_1 - H_1 H_2^{-T} J_2$ ,  $N = N_1 - H_1 H_2^{-T} N_2$ .

#### Finite-Time Stability:

Select the positive definite function as

$$V = \frac{1}{2} s^T s \quad (5)$$

$$\dot{V} = s^T \dot{s} = s^T (\alpha \dot{e} + \Gamma \beta \text{diag}(|\dot{e}|^{\gamma-1}) \ddot{e}) \quad (6)$$

Substitute controller (4) into dynamic model (1) will lead to

$$\begin{aligned} H\ddot{x} &= H\ddot{x}_d - H(q_1 s + q_2 \text{sgn}(s)^\rho) - H\Gamma^{-1} \beta^{-1} \alpha \text{sgn}(\dot{e})^{2-\gamma} - H_1 H_2^{-T} \delta \\ \ddot{e} &= -\Gamma^{-1} \beta^{-1} \alpha \text{sgn}(\dot{e})^{2-\gamma} - (q_1 s + q_2 \text{sgn}(s)^\rho - \hat{\delta}) \end{aligned} \quad (7)$$

where  $\hat{\delta} = -H^{-1} H_1 H_2^{-T} \delta$ . Combining (6) and (7), we can obtain the following equation:

$$\dot{V} = -s^T \Gamma \beta \text{diag}(|\dot{e}|^{\gamma-1}) (q_1 s + q_2 \text{sgn}(s)^\rho - \hat{\delta}) \quad (8)$$

Through equivalent transformation, (8) can be written as

$$\dot{V} = -s^T \Gamma \beta \text{diag}(|\dot{e}|^{\gamma-1}) ((q_1 - \text{diag}(\hat{\delta}) \text{diag}^{-1}(s)) s + q_2 \text{sgn}(s)^\rho) \quad (9)$$

Assuming that  $q_1 - \text{diag}(\hat{\delta}) \text{diag}^{-1}(s)$  is a positive definite diagonal matrix, then (9) can be written as

$$\dot{V} = -s^T \hat{q}_1 s - s^T \hat{q}_2 \text{sgn}(s)^\rho \quad (10)$$

The variables defined in (10) is as follows:

$$\begin{aligned} \hat{q}_1 &= \Gamma \beta \text{diag}(|\dot{e}|^{\gamma-1}) (q_1 - \text{diag}(\hat{\delta}) \text{diag}^{-1}(s)) = \text{diag}(\hat{q}_{11} \ \hat{q}_{12} \ \hat{q}_{13}) \\ \hat{q}_2 &= \Gamma \beta \text{diag}(|\dot{e}|^{\gamma-1}) q_2 = \text{diag}(\hat{q}_{21} \ \hat{q}_{22} \ \hat{q}_{23}) \end{aligned}$$

Define,  $\hat{q}'_1 = \min(\hat{q}_{11} \ \hat{q}_{12} \ \hat{q}_{13})$ ,  $\hat{q}'_2 = \min(\hat{q}_{21} \ \hat{q}_{22} \ \hat{q}_{23})$ , we can obtain the following inequality:

$$\begin{aligned} s^T \hat{q}_1 s &\geq 2\hat{q}'_1 V \\ s^T \hat{q}_2 \text{sgn}(s)^\rho &\geq \hat{q}'_2 (s_1^{\rho+1} + s_2^{\rho+1} + s_3^{\rho+1}) \geq \hat{q}'_2 (2V)^{\frac{\rho+1}{2}} \end{aligned} \quad (11)$$

Substitute (11) into (10) will lead to

$$\dot{V} \leq -2\hat{q}'_1 V - \hat{q}'_2 (2V)^{\frac{\rho+1}{2}} \quad (12)$$

By analyzing inequality (12), it can be found that the stability time of the system satisfies the inequality (13). Therefore, the system will be stable in a limited time:

$$T \leq \frac{1}{\hat{q}'_1(1-\rho)} \ln \frac{\hat{q}'_1 V^{\frac{1-\rho}{2}}(x_0) + \hat{q}'_2 2^{\frac{\rho-1}{2}}}{\hat{q}'_2 2^{\frac{\rho-1}{2}}} \quad (13)$$

## 4 Simulation Experiment

In this section, the controller of the constant force suspension system based on the new reaching law is simulated and compared with the existing sliding mode controller [9]. The system model parameters are  $k = 700\text{N/m}$ ,  $l_0 = 0.48\text{m}$ ,  $r = 0.02\text{m}$ ,  $g = 10\text{m/s}^2$ ,  $M = 15.2\text{kg}$ ,  $m_1 = 0.54\text{kg}$ ,  $m_2 = 4.2\text{kg}$ ,  $m_x = 15.9\text{kg}$ ,

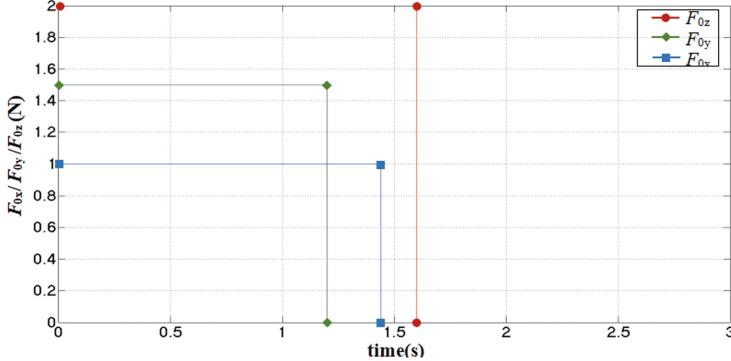


Fig. 2. The driving force of the test object

$m_y = 42\text{kg}$ . According to the practical application of the system, the expected value of the state vector is defined as  $x_d = [0 \ 0 \ 0]^T$ .

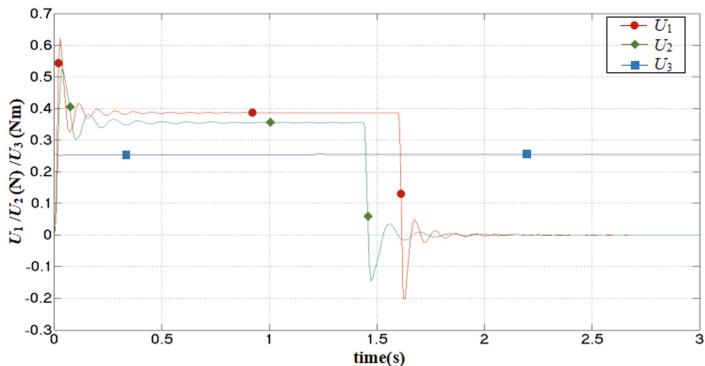
This paper uses  $\dot{e}/(|\dot{e}| + 0.05)$  instead of the sign function  $\text{sign}(\dot{e})$  to reduce the chattering of the system, in the simulation experiment. Assume that the driving force of the test object, i.e. the system interference is the pulse signal as shown in Figure 2. The parameters of the sliding mode controller (4) are selected as:

$$\begin{aligned} q_1 &= \text{diag}(90 \ 100 \ 40), q_2 = \text{diag}(30 \ 8 \ 10), q_3 = \text{diag}(20 \ 40 \ 40) \\ \beta &= \text{diag}(0.4 \ 1 \ 1), \gamma = 1.6, \rho = 1/3 \end{aligned}$$

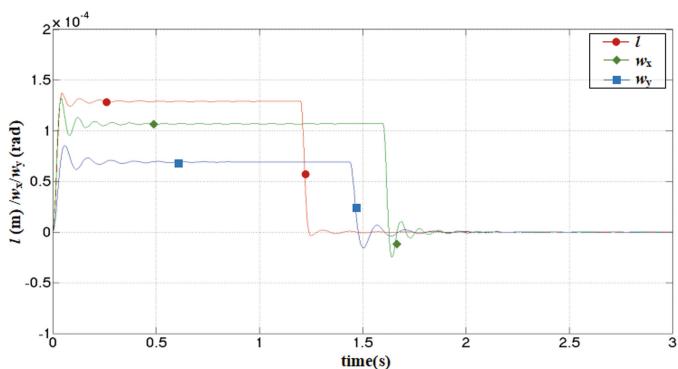
The output curve of the controller based on the new reaching law is shown in Figure 3. The output  $U = [U_1 \ U_2 \ U_3]^T = 1/M[F_x \ F_y \ T_\alpha]^T$ . Through the system state response curve in Figure 4, it can be found that the motion state of the system is closely related to the driving force of test object, which is consistent with the practical application. In this simulation experiment, when  $t \geq 2.2\text{s}$ , the state deviation of the system reaches the interval  $\|e\| \leq 1.83 \times 10^{-4}$ . The overshoot of the length variation  $l$  and the orthogonal decomposition value  $\omega_x$  and  $\omega_y$  of the swing angle are

$$\sigma_l \approx 1.37 \times 10^{-4}\text{m}, \sigma_{\omega_x} \approx 1.33 \times 10^{-4}\text{rad}, \sigma_{\omega_y} \approx 8.53 \times 10^{-5}\text{rad}$$

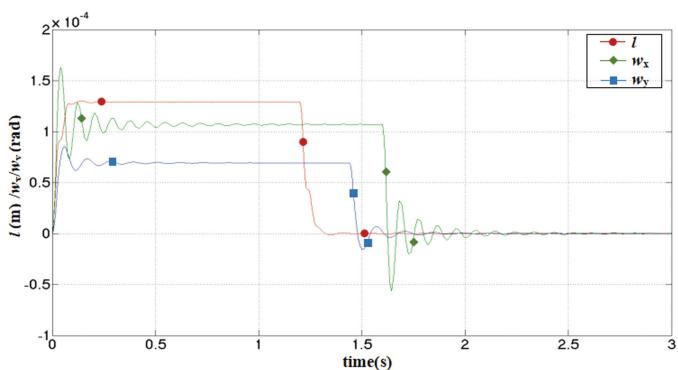
Applying the existing sliding mode controller[9] to this system, the simulation results are shown in Figure 5. It can be found through comparison that the controller based on the new reaching law proposed in this paper reduces the overshoot of  $\omega_x$  by 18.7%, when the overshoot of  $l$  and  $\omega_y$  doesn't change by more than 5.6%. In addition, the time for the system state deviation to reach the interval  $\|e\| \leq 1.83 \times 10^{-4}$  is shortened by nearly 0.25s.



**Fig. 3.** Output of the controller based on the new reaching law



**Fig. 4.** State response of the control system based on the new reaching law



**Fig. 5.** State response of the existing control system

## 5 Conclusions

In this paper, a new reaching law is proposed, and then a new sliding mode controller is designed for the constant force suspension system. Through theoretical analysis, the finite time stability of the system is proved. The simulation results show that, compared with the existing controller, the controller suppresses the overshoot and chattering while reducing the settling time. Therefore, the stability of the system is improved.

**Acknowledgement.** This work was supported by the NSFC (61327807, 61521091, 61520106 010, 61134005) and the National Basic Research Program of China (973 Program: 2012CB821200, 2012CB821201).

## References

1. Kim, B.M., Velenis, E., Kriengsiri, P., et al.: Designing a low-cost spacecraft simulator[J]. *IEEE Contr. Syst. Mag.* **23**(4), 26–37 (2003)
2. Nakayama, A., Hirata, T., Tsujita, K.: A study of a gravity compensation system for the spacecraft prototype test by using multi-robot system[J]. *Artif. Life Robot.* **25**(1), 81–88 (2020)
3. Zhao, Z., Kangjia, F., Li, M., Li, J., Xiao, Y.: Gravity compensation system of mesh antennas for in-orbit prediction of deployment dynamics[J]. *Acta Astronaut.* **167**, 1–13 (2020)
4. Jacobs, S.E., Akin, D.L., Braden, J.R.: System overview and operations of the mx-2 neutral buoyancy space suit analogue[C]. In: International Conference on Environmental Systems (2006)
5. Spencer, M., Chernesky, V., Baker, J., et al.: Bifocal relay mirror experiments on the NPS three axis spacecraft simulator[C]. In: AIAA Guidance, Navigation, and Control Conference and Exhibit (2002)
6. Pletser, V., Rouquette, S., Friedrich, U., et al.: European parabolic flight campaigns with Airbus ZERO-G: Looking back at the A300 and looking forward to the A310[J]. *Adv. Space Res.* **56**(5), 1003–1013 (2015)
7. Yuan, J., Zhu, Z., Ming, Z., Luo, Q.: An innovative method for simulating microgravity effects through combining electromagnetic force and buoyancy[J]. *Adv. Space Res.* **56**(2), 355–364 (2015)
8. Boge, T., Ma, O.: Using advanced industrial robotics for spacecraft rendezvous and docking simulation[C]. In: 2011 IEEE International Conference on Robotics and Automation. IEEE (2011)
9. Jia, J.: Research on Control and Implementation of an Active Gravity Compensation System. Beihang University (2018)
10. Jia, J., Jia, Y., Sun, S.: Preliminary design and development of an active suspension gravity compensation system for ground verification[J]. *Mech. Mach. Theory* **128**, 492–507 (2018)



# Real-Time Coverage Path Planning of a UAV with Threat and Value Zone Constraints

Yan Liu<sup>1(✉)</sup>, Hao Li<sup>3</sup>, and Zhi Liu<sup>2</sup>

<sup>1</sup> Hengxiang Control Technology Company Limited, Xi'an, China  
yan.L70163.com

<sup>2</sup> National Key Laboratory of Science and Technology on Aircraft Control, Xidian University, Xi'an, China

<sup>3</sup> School of Artificial Intelligence,  
Xi'an, FACRI, China

**Abstract.** Coverage path planning (CPP) is the task of finding a path that covers every point of an area of interest. Unmanned Aerial Vehicles (UAVs) are being widely used in various fields, such as rescuing, photography, agriculture, and surveillance. However, most of the research focused on finding the optimal path that can cover a specific area without threat and value zone constraints. This paper proposes an entry-and-exit point geometric pattern based CPP algorithm that can avoid no-fly zones and visit high-value zones effectively. The algorithm does not require iterative search and optimization, so it has a real-time performance. Finally, the performances of the proposed method on simulation and real environments are presented and analyzed.

**Keywords:** Coverage path planning · Threat zone avoidance · Value zone visit · Unmanned aerial vehicle · Geometric pattern

## 1 Introduction

Compared with manned air-crafts, Unmanned Aerial Vehicles (UAVs) are more suitable for dangerous and dirty missions [1]. Over the past decade, UAVs have been used in a wide range of applications, such as smart farming, civil security, photogrammetry, wildfire tracking, and surveillance, etc. [2]. Coverage Path Planning (CPP) is an NP-hard motion planning problem. The CPP problem has become more popular in our daily life, especially for UAV and cleaning robot applications. For UAV, CPP aims at determining an optimal flight path that can cover an Area Of Interest (AOI), while avoiding no-fly zones (NFZs) or threat zones.

The CPP problem can be divided into two main categories: complete coverage path planning and incomplete coverage path planning. For complete CPP, there is no need to consider the high-value zone (HVZ) visit problem. This is because the UAV can visit any point outside the no-fly zones. However, for incomplete

CPP problem, we need to take it into consideration [3]. The CPP problem of UAVs is more complicated than that of ground-based vehicles. This is because a UAV usually carries a small amount of electric energy and is difficult to be controlled. Thus, there should be as few turns as possible in the coverage path. Furthermore, the coverage path should be short enough to reduce energy consumption and flight time [4]. The area to be covered is usually a polygonal area, and non-polygonal area can be transformed to multi-polygonal areas by decomposition methods such as cellular decomposition [4] and convex decomposition [5,6]. Therefore, This paper will focus on CPP of a convex polygonal area.

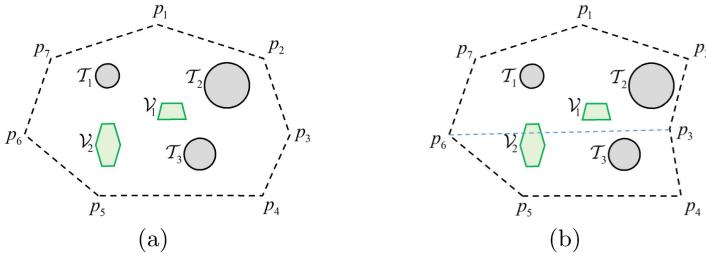
Existing CPP methods can be summarized into five main categories: geometry-based, search-based, optimization-based and structure inspection based methods. For simple AOI (such as convex polygonal region), geometric patterns are sufficient to explore such areas. The most common patterns are the back-and-forth (BF) [7] and the spiral [8]. However, these algorithms are inefficient when dealing with irregular-shaped areas in complex scenarios. What's more, the paths generated by these algorithms usually have a large turning angle when bypassing the no-fly area and can not visit the high-value area. The search-based methods usually discretize the AOI into grid cells firstly and then apply a cost to each moving step. Finally, search algorithms such as  $D^*$  [9] or  $A^*$  [10] is utilized to find the shortest coverage path. However, the grid-based methods usually demand high computational time leading to inefficient and expensive paths, making them not usable in real-world scenarios [11]. The optimization-based algorithms model the CPP problem as a mathematical optimization problem and then the optimal path is solved by the optimization algorithm. The most commonly used optimization algorithm is Genetic Algorithm (GA) [12–14]. The disadvantages of these algorithms are two folds. First, different initialization parameters lead to different paths. Second, the optimization time of solving the path is too long to meet the need for real-time planning. Recently, structure inspection based methods [15,16] utilize the structure information obtained from sensors (such as a camera) equipped on UAV to assist planning. However, these algorithms can only be used on UAVs with specific devices and need high-performance computing resources.

In order to overcome these shortcomings, in this paper, we propose a real-time CPP algorithm with the ability to avoid the NFZ and visit the HVZ based on entry-and-exit point (EE) geometric pattern. The rest of this paper is organized as follows. Section 2 describes the coverage path planning problem and introduces the BF line scan CPP method. Then, the designs of geometric patterns for avoiding NFZs and visit HVZs are described in detail in Sect. 3. Experimental results of simulation and real scenarios are presented in Sect. 4. Finally, the conclusions are given in Sect. 5.

## 2 Related Works

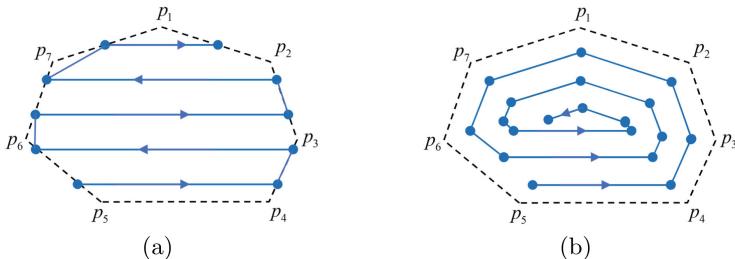
The polygonal area of interest (AOI) can be expressed as an ordered point set  $\mathcal{A} = \{p_i = (x_i, y_i)\}_{i=1}^{N_p}$ , where,  $p_i$  represents the  $i$ -th vertex of the polygon,

$(x_i, y_i)$  is the two-dimensional euclidean coordinate of  $p_i$ . As illustrated in Fig. 1, the AOI is a polygonal zone with seven vertexes and edges. There are three circular threat zones  $T_1, T_2, T_3$  and two polygonal high-value zones  $V_1, V_2$  in the AOI. Figure 1(a) demonstrates a convex polygonal AOI while Fig. 1(b) demonstrates a concave one. For concave AOI in Fig. 1(b), it can be decomposed into two convex polygonal AOIs by line  $p_3p_6$ . The most commonly used decomposition method is cellular decomposition [4].



**Fig. 1.** Different polygonal areas of interest with three circular no-fly zones and two high-value polygonal zones explored during CPP. (a) Convex polygonal area. (b) Concave polygonal area.

Once a complex polygonal AOI scenario is decomposed into simple convex polygonal areas, we can perform the CPP method to generate a coverage path for each sub-area. The back-and-forth (BF) and spiral geometric patterns based coverage methods are the most commonly used two algorithms to generate waypoints. According to [18], the BF approach without AOI decomposition presents good and trustworthy results in relation to the spiral approach while the spiral approach generates shorter paths in rounded-shape AOI with large inner angles. Figure 2 gives the results of the above two methods, in which, the blue solid line represents the coverage path. However, the original BF or spiral algorithm does not support NFZ avoidance and HVZ visiting. In this paper, we focus on extending them to support these two functions.



**Fig. 2.** Coverage waypoints generated by back-and-forth and spiral method respectively. (a) Back-and-forth. (b) Spiral.

### 3 Methodology

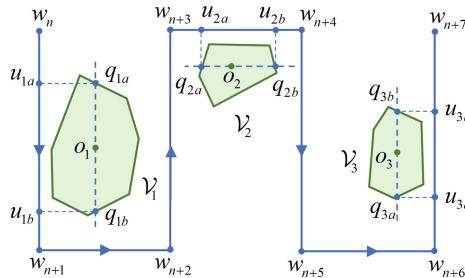
In this section, we describe our proposed scheme for real-time CPP of a UAV with NFZ and HVZ constraints.

The first step of our scheme is the initial CPP without NFZ and HVZ constraints. In order to obtain real-time performance, we choose the simple BF or spiral geometric pattern approach to generate the initial waypoints. The second step is high-value zone visit planning and the last step is no-fly zone avoidance planning.

#### 3.1 High-Value Zone Visit

High-Value Zone (HVZ) visit re-planning aims to refine the paths and make the paths cross all HVZs. Denote the waypoints generated by BF as an ordered set  $\mathcal{W} = \{w_n = (x_n, y_n)\}_{n=1}^{N_w}$ . Let  $\mathcal{V} = \{\mathcal{V}_m = \{(x_n, y_n)\}_{n=1}^{N_{\mathcal{V}_m}}\}_{m=1}^{M_V}$  represent all the polygonal HVZs in the AOI, where,  $M_V$  is the number of HAVs,  $N_{\mathcal{V}_m}$  is the number of vertexes of the  $m$ -th HVZ  $\mathcal{V}_m$  and  $(x_n, y_n)$  is the 2d-coordinates of the  $n$ -th vertex.

We hope that the re-planned path can pass through most parts of each HVZ. Therefore, we make the re-planned path pass through all the centroid of each HVZ. The centroid  $o_m$  of  $\mathcal{V}_m$  can be computed by averaging the 2d-coordinates of all the vertexes in  $\mathcal{V}_m$ .



**Fig. 3.** High-value zone visit.

As we know, two adjacent waypoints  $w_n, w_{n+1}$  form a linear track  $\overrightarrow{w_n w_{n+1}}$ . In order to obtain the shortest visit path, find the directed line segment  $\vec{l}_m = \overrightarrow{w_m w_{m+1}}$  that is closest to the centroid  $o_m$ . Let  $\vec{s}_m = \overrightarrow{q_{ma} q_{mb}}$  be the directed line segment that pass through centroid  $o_m$  and has the same direction as  $\vec{l}_m$ , where,  $q_{ma}$  and  $q_{mb}$  are the entry-point and exit-point when a UAV flies over value zone  $\mathcal{V}_m$  along the directed line segment  $\vec{s}_m$  respectively. Denote  $u_{ma}, u_{mb}$  as the vertical projection points of  $q_{ma}, q_{mb}$  onto line  $\vec{l}_m$  respectively. The re-planned path can be obtained by inserting points  $u_{ma}, q_{ma}, q_{mb}, u_{mb}$  between

waypoint  $w_{n_m}$  and  $w_{n_m+1}$  sequentially. Figure 3 gives a demonstration of the above re-planning process with three HVZs  $\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3$ .

It's a little bit complicated to obtain entry point  $q_{ma}, q_{mb}$  by solving equations. We get them in another way. Firstly, discretize the directed line  $\overrightarrow{s_m}$  into points. Secondly, determine whether these points are in  $\mathcal{V}_m$  and get a flag vector  $\mathbf{f}_m = \{0, \dots, 0, 1, \dots, 1, 0, \dots, 0\}$ . Thirdly, apply 1-order gradient ( $\mathbf{f}'_m(i) = \mathbf{f}_m(i+1) - \mathbf{f}_m(i)$ ) on  $\mathbf{f}_m$  to get  $\mathbf{f}'_m$ . Finally, find the position  $i_1, i_{-1}$  of 1, -1 in  $\mathbf{f}'_m$  and then we obtain  $q_{ma} = \mathbf{f}(i_1), q_{mb} = \mathbf{f}(i_{-1})$ .

### 3.2 No-Fly Zone Avoidance

The initially planned coverage paths may cross NFZs, we proposed an entry-and-exit (EE) point pattern based NFZ avoidance algorithm to overcome the problem. Suppose the NFZ is circular and can be represented by  $(x, y, r)$ , where,  $x, y$  are the center point coordinates of the circle, and  $r$  is radius. By finding the entry point and exit point when a UAV passes through the NFZ along the initial path, a short arc from the entry point to the exit point is constructed as a new NFZ avoidance path. Let  $\mathcal{T} = \{\mathcal{T}_m = (x_m, y_m, r_m)\}_{m=1}^{N_T}$  be all the NFZs, where  $N_T$  is the number of threat zones. Denote  $\mathcal{W} = \{w_n = (x_n, y_n)\}_{n=1}^{N_w+4N_V}$  as the waypoints after HVZ visit re-planning, where  $N_w$  is the number of initial waypoints,  $N_V$  is the number of HVZs.

Let  $\mathcal{T}_m^d = (x_m, y_m, r_m + d)$ , where  $d$  is the global minimum safe distance to NFZs. For each NFZ  $\mathcal{T}_m = (x_m, y_m, r_m)$ , find all the EE point pairs  $\mathcal{E}_m = \{q_{ma_k} q_{mb_k}\}_{k=1}^{K_m}$  when the UAV flies over  $\mathcal{T}_m^d$  along the initially planned path. Where,  $K_m$  is the number of EE point pairs of the  $m$ -th NFZ. Denote  $\theta_{ma_k}, \theta_{mb_k} \in [-\pi, \pi]$  as the angle from unit vector  $\vec{e} = (1, 0)$  to vector  $\overrightarrow{o_m q_{ma_k}}$  and  $\overrightarrow{o_m q_{mb_k}}$  respectively. The range of the central angle of the shorter arc  $\widehat{q_{ma_k} q_{mb_k}}$  can be computed by

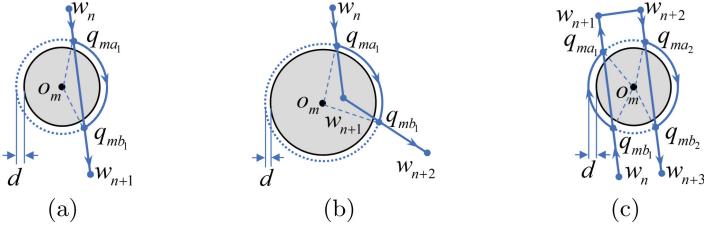
$$\begin{cases} [\theta_{ma_k}, \theta_{mb_k}], & \text{if } |\theta_{mb_k} - \theta_{ma_k}| < \pi \\ [\theta_{ma_k}, \text{sign}(\theta_{ma_k})\pi; \text{sign}(\theta_{mb_k})\pi, \theta_{mb_k}], & \text{else} \end{cases} \quad (1)$$

where,  $\text{sign}(\cdot)$  is the sign function, interval  $[\theta_{ma_k}, \text{sign}(\theta_{ma_k})\pi; \text{sign}(\theta_{mb_k})\pi, \theta_{mb_k}]$  is composed of interval  $[\theta_{ma_k}, \text{sign}(\theta_{ma_k})\pi]$  and interval  $[\text{sign}(\theta_{mb_k})\pi, \theta_{mb_k}]$ . Discretize the interval expressed in Eq. 1 into  $N_\theta$  angles and substitute these angles into Eq. 2 to get the re-planned waypoints for avoiding NFZ  $\mathcal{T}_m$ .

$$\begin{cases} x = x_m + (r_m + d)\cos(\theta) \\ y = y_m + (r_m + d)\sin(\theta) \end{cases} \quad (2)$$

Figure 4 illustrates different cases when the UAV flies over an NFZ. Figure 4(a) and Fig. 4(b) show the case with single EE point pair while Fig. 4(c) shows the case with multiple EE point pairs.

Solving the entry point  $q_{ma_k}$  and exit point  $q_{mb_k}$  by solving the equations composed of flight paths and boundaries of NFZs is difficult. We utilize the same method described in Sect. 3.1 to obtain the coordinates of EE point pairs.



**Fig. 4.** No-fly zone avoidance.

## 4 Experimental Results

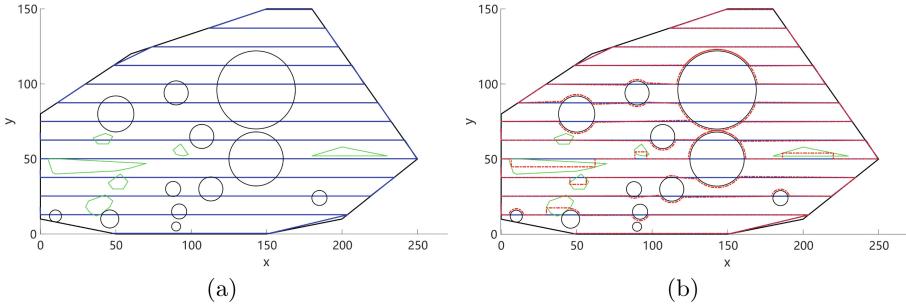
To evaluate our proposed entry-and- exit point based algorithm, we test the algorithm in simulation and real scenarios respectively. The experimental platform is Intel(R) Core(TM) m3-6Y30 CPU, 0.90 GHz, 5W, 4GB RAM, MATLAB R2015b.

In the simulation scenario, the area of interest is a convex polygonal region specified by vertex set  $\mathcal{A} = \{(0, 10), (50, 0), (100, 0), (150, 0), (200, 10), (250, 50), (180, 150), (150, 150), (90, 130), (60, 120), (0, 80)\}$  in counter-clockwise order. There are a total of twelve no-fly zones  $\mathcal{T} = \{(10, 12, 4), (46, 10, 6), (50, 80, 12), (88, 30, 5), (92, 15, 5), (90, 5, 3), (90, 94, 8), (107, 65, 8), (113, 30, 8), (143, 96, 26), (143, 50, 18), (185, 24, 5)\}$  and six high-value zones  $\mathcal{V} = \{\{(40, 13), (46, 18), (48, 22), (43, 26), (32, 22), (30, 18), (33, 14), (35, 13), (37, 12)\}, \{(10, 40), (50, 42), (60, 44), (70, 47), (13, 50), (5, 50), (8, 41)\}, \{(50, 30), (55, 30), (58, 35), (53, 40), (45, 34), (48, 30)\}, \{(40, 60), (45, 60), (48, 65), (43, 67), (35, 64), (38, 60)\}, \{(90, 53), (95, 52), (98, 53), (93, 60), (88, 56)\}, \{(180, 52), (230, 52), (200, 58)\}\}$  in simulated scenario as illustrated in Fig. 5. In Fig. 5, the boundaries of threat zones are indicated by black circles, the boundaries of value zones are indicated by green polygons.

We first use the back-forth scan method to generate an initial coverage path. The initially planned path is shown in Fig. 5(a) and is plotted as blue solid line. We can see that the path crosses all most every threat zone and can't visit each value zone. In order to avoid the threat zones and visit the value zones, we need to replan the initialized path. We first utilize our proposed high-value zone visit method to replan the path and make it visit all the high-value zones. Next, the proposed no-fly zone avoidance method is applied to avoid the UAV entering the no-fly zones. The results are shown in Fig. 5(b) and indicated by red paths.

We repeatedly run our proposed algorithm 1000 times, the average time consumed is summarized in Table 1, the total planning time is 51.23ms. We can draw the conclusion that our proposed algorithm can get a real-time performance on mobile low-pressure processing platform.

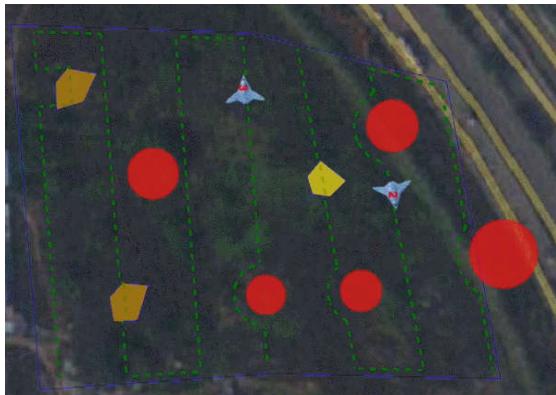
We also test our proposed algorithm in reality, and the results are illustrated in Fig. 6. In the real scenario, the area of interest has a hexagonal shape. There are five no-fly zones and three high-value zones in the area. The hexagonal area is non-convex, thus it is divided into two quadrilateral parts for two UAVs



**Fig. 5.** CPP results of our proposed algorithm in simulated environment scenario. The boundary of the polygonal area to be investigated is plotted as the black polygonal line. (a) Initially planned path generated by back-and-forth scan algorithm. (b) Replanned path generated by our proposed algorithm.

**Table 1.** The average running time of the proposed algorithm

Back-forth coverage	High-value zone visit	No-fly zone avoidance
23.70ms	21.62ms	5.91ms



**Fig. 6.** CPP results of our proposed algorithm in real environment scenario. The five no-fly zones are indicated by red circular regions and the three high-value zones are indicated by yellow polygonal regions. The blue polygon encloses the area to be investigated. The green dashed lines indicate the planned path by our proposed algorithm.

to cover. We generate two initial coverage paths use the back-and-forth scan method for each quadrilateral area, the two paths are then replanned by our proposed algorithm respectively. As shown in Fig. 6, the planned paths can avoid all the threat zones and visit all the high-value zones while covering the full reconnaissance area.

## 5 Conclusion

This paper studies an entry-and-exit point geometric pattern based coverage path planning algorithm for UAVs with no-fly and high-value zone constraints. Since it is non-iterative and non-search based, it has low computational complexity and real-time performance. Future works will focus on improving the entry-and-exit point geometric pattern-based CPP algorithm by introducing smoothing and shortest path search.

## References

1. Jiao, Y.S., Wang, X.M., Chen, H., Li, Y.: Research on the coverage path planning of UAVS for polygon areas. In: 2010 5th IEEE Conference on Industrial Electronics and Applications, pp. 1467C1472 (2010)
2. Cabreira, T.M., Brisolara, L.B., Ferreira Jr, P.R.: Survey on coverage path planning with unmanned aerial vehicles. *Drones* **3**, (2019). <https://doi.org/10.3390/drones3010004>
3. Stack, J.R., Smith, C.M.: Combining random and data-driven coverage planning for underwater mine detection. In: Oceans 2003. Celebrating the Past Teaming Toward the Future (IEEE Cat. No. 03CH37492), pp. 2463C2468 (2003)
4. Li, Y., Chen, H., Er, M.J., Wang, X.: Coverage path planning for UAVs based on enhanced exact cellular decomposition method. *Mechatronics* **21**(5), 876–885 (2011). <https://doi.org/10.1016/j.mechatronics.2010.10.009>
5. Coombes, M., Fletcher, T., Chen, W.-H., Liu, C.: Optimal polygon decomposition for UAV survey coverage path planning in wind. *Sensors* **18**(7), 2132 (2018). <https://doi.org/10.3390/s18072132>
6. Nielsen, L.D., Sung, I., Nielsen, P.: Convex decomposition for a coverage path planning for autonomous vehicles: Interior extension of edges. *Sensors* **19**(19), 4165 (2019). <https://doi.org/10.3390/s19194165>
7. Huang, W.H.: Optimal line-sweep-based decompositions for coverage algorithms. In: Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation, pp. 27C32 (2001)
8. Cabreira, T.M., Di Franco, C., Ferreira, P.R., Buttazzo, G.C.: Energy-aware spiral coverage path planning for UAV photogrammetric applications. *IEEE Robot. Autom. Lett.* **3**(4), 3662–3668 (2018)
9. Dakulović, M., Horvatić, S., Petrović, I.: Complete coverage D\* algorithm for path planning of a floor-cleaning mobile robot. *IFAC Proc. Vol.* **44**(1), 5950–5955 (2011). <https://doi.org/10.3182/20110828-6-IT-1002.03400>
10. Dogru, S., Marques, L.: A\*-Based solution to the coverage path planning problem. In: ROBOT 2017: Third Iberian Robotics Conference, pp. 240C248. Springer International Publishing, Cham (2018)
11. Cabreira, T.M., Ferreira, P.R., Di Franco, C., Buttazzo, G.C.: Grid-based coverage path planning with minimum energy over irregular-shaped areas with Uavs. In: 2019 International Conference on Unmanned Aircraft Systems (ICUAS), pp. 758C767 (2019)
12. Nagib, G., Gharieb, W.: Path planning for a mobile robot using genetic algorithms. *IEEE Proceedings of Robotics*, 185C189 (2004). <https://doi.org/10.1109/ICEEC.2004.1374415>

13. Mohamed, A.Y., Mohamed, T.L.: The path planning of cleaner robot for coverage region using Genetic Algorithms. *J. Innov. in Digit. Ecosyst.* **3**(1), 37–43 (2016). <https://doi.org/10.1016/j.jides.2016.05.004>
14. Schäfle, T.R., Mohamed, S., Uchiyama, N., Sawodny, O.: Coverage path planning for mobile robots using genetic algorithm with energy optimization. In: 2016 International Electronics Symposium (IES), pp. 99C104 (2016). <https://doi.org/10.1109/ELECSYM.2016.7860983>
15. Sina, S.M., Christoforos, K., Emil, F., Dariusz, K., George, N.: Cooperative coverage path planning for visual inspection. *Control Eng. Pract.* **74**, 118–131 (2018). <https://doi.org/10.1016/j.conengprac.2018.03.002>
16. Almadhoun R., Taha T., Dias J., Seneviratne L., Zweiiri Y.: Coverage path planning for complex structures inspection using unmanned aerial vehicle (UAV). In: Intelligent Robotics and Applications. ICIRA 2019. Lecture Notes in Computer Science, vol 11744. Springer, Cham (2019)
17. Song, Z., Zhang, H., Zhang, X., Zhang, F.: Unmanned aerial vehicle coverage path planning algorithm based on cellular automata. In: 2019 15th International Conference on Computational Intelligence and Security (CIS), pp. 123C126 (2019). <https://doi.org/10.1109/CIS.2019.00034>
18. Maza, I., Ollero, A.: Multiple UAV cooperative searching operation using polygon area decomposition and efficient coverage algorithms. In: Distributed Autonomous Robotic Systems 6, pp. 221C230. Springer (2007)

# Author Index

## B

- Bao, Xinbiao, 194  
Bi, Zhenfa, 428, 475  
Bian, Jilong, 409

## C

- Cai, Qiang, 225  
Cao, Tianyu, 242, 261, 270  
Chai, Yi, 127  
Chen, Chuanfang, 688  
Chen, Jiyang, 659  
Chen, Li, 377  
Chen, Mou, 18  
Chen, Na, 69  
Chen, Penghao, 145  
Chen, Qiang, 419  
Chen, Qiaoyu, 287, 518  
Chen, Yangzhou, 36  
Chen, Zengqiang, 185  
Cheng, Hangyang, 69, 155

## D

- Dai, Haofei, 659  
Dai, Xianhua, 549  
Deng, Haipeng, 351  
Deng, Weiwei, 541  
Deng, Yibiao, 225  
Ding, Bo, 819  
Dong, Dengwei, 251  
Dong, Jinfeng, 646  
Du, Dongsheng, 508  
Duan, Meng, 832  
Duo, Jingyun, 342, 560

## F

- Fan, Jingzi, 317  
Fan, Songtao, 569  
Fan, Yimin, 136  
Fang, Yu, 446, 455  
Fang, Yuan, 549  
Feng, Qingquan, 333

## G

- Gan, Yiming, 69, 155, 165, 175  
Gao, Dai, 579  
Gao, Hai, 358  
Gao, Han, 308  
Ge, Xiaoxing, 446  
Gong, Guanghong, 437  
Gong, Kai, 832  
Gu, Weikun, 251  
Gu, Xiang, 117  
Guo, Jiangzhen, 778  
Guo, Jiaxin, 127  
Guo, Xuemei, 549, 646  
Guo, Yao, 669  
Guo, Yongheng, 194  
Guo, Zirong, 739

## H

- Han, Cunwu, 624  
Han, Sumin, 730  
He, Xuehui, 155, 165  
He, Yi, 624  
Hou, Ao, 730  
Hu, Chaofang, 669  
Hu, Longhui, 708

Hu, Weikang, 389  
 Hu, Yumei, 792  
 Hua, Yu, 541  
 Huang, Fengguang, 721  
 Huang, Jun, 599, 608  
 Huang, Xinghua, 590

**J**  
 Ji, Feng, 48  
 Ji, Lei, 342, 560  
 Jia, Yingmin, 1, 27, 419, 569, 832  
 Jia, Yuxin, 832  
 Jiang, Xiong, 194  
 Jiang, Yunxia, 760  
 Jiang, ZhengLin, 760

**K**  
 Kong, Deyu, 688

**L**  
 Lei, Lai, 778  
 Leo, Rongfan, 615  
 Li, Bin, 297  
 Li, Haifeng, 810  
 Li, Haisheng, 225  
 Li, Hao, 840  
 Li, Jiayi, 217  
 Li, Jinfeng, 409  
 Li, Ping, 801  
 Li, Runze, 437  
 Li, Shaowei, 333  
 Li, Shouyi, 18  
 Li, Shurong, 634  
 Li, Xinqing, 78  
 Li, Yaxin, 1  
 Li, Yuanyuan, 590  
 Li, Zhiyong, 333  
 Liang, Xinglong, 679  
 Liang, Yiming, 351  
 Liao, Guobo, 127  
 Lin, Manfei, 541  
 Lin, Yue, 569  
 Liu, Dan, 278  
 Liu, Fan, 175  
 Liu, Fei, 769  
 Liu, Jialun, 99  
 Liu, Jinkun, 242  
 Liu, Lei, 624  
 Liu, Ningxi, 242, 261, 270  
 Liu, Shizhao, 333  
 Liu, Shuang, 615  
 Liu, Wei, 730  
 Liu, Xiangsheng, 760  
 Liu, Yan, 840

Liu, Yang, 136, 527  
 Liu, Zhaojiang, 659  
 Liu, Zhenzu, 495  
 Liu, Zhi, 840  
 Liu, Zhongxin, 185  
 Lou, Yunjiang, 251  
 Lu, Fengshun, 194  
 Lu, Yao, 832  
 Luan, Yizhong, 659  
 Luo, Yaqin, 590  
 Lv, Jing, 110  
 Lyu, Jianting, 579

**M**  
 Ma, Jirong, 351, 792  
 Ma, Qinghua, 351, 792  
 Ma, Sile, 659  
 Ma, Zhengguang, 308  
 Mei, Panpan, 495  
 Meng, Deyuan, 739, 749  
 Meng, Jun, 590  
 Mo, Lipo, 57, 367

**Q**  
 Qi, Guanqiu, 590  
 Qi, Jinpeng, 99  
 Qi, Long, 194  
 Qian, Houbin, 145  
 Qu, Xiaoyu, 326

**R**  
 Ren, Lin, 760  
 Ren, Yuanhong, 165

**S**  
 Shen, Song, 10  
 Shen, Xizhong, 615  
 Shi, Xuqing, 377  
 Shi, Zhongsuo, 110  
 Shu, Yudan, 377  
 Song, Lailong, 760  
 Su, Chunying, 579  
 Sun, Qinglin, 234, 464  
 Sun, Wenxu, 659  
 Sun, Yujiao, 769

**T**  
 Tan, Panlong, 464  
 Tang, Wei, 446  
 Tian, Lin, 721  
 Tjan, Patrick, 669  
 Tong, Dongbing, 287, 518  
 Tong, Jiahui, 810

**W**

- Wang, Chao, 342  
Wang, Chaoli, 10, 696, 708  
Wang, Dongxiao, 333  
Wang, Fuyong, 185  
Wang, Fuzhong, 730, 778  
Wang, Guoli, 549, 646  
Wang, Guoqing, 206  
Wang, Jiangang, 333  
Wang, Jiangyun, 326, 437  
Wang, Juan, 399, 485  
Wang, Kang, 527  
Wang, Lei, 508  
Wang, Lijun, 242, 261, 270  
Wang, Lin, 696  
Wang, Na, 136  
Wang, Peng, 634  
Wang, Shuaiwei, 351, 792  
Wang, Wei, 721  
Wang, Xin, 579  
Wang, Yang, 297  
Wang, Yao, 287  
Wang, Yuhui, 18  
Wang, Zhimin, 10  
Wang, Zongkai, 475  
Wei, Wei, 801  
Wei, Xinjiang, 78, 88  
Weng, Miao, 455  
Wu, Aiguo, 389  
Wu, Lianghong, 495  
Wu, Qingxian, 18  
Wu, Yuxin, 749

**X**

- Xia, Chengyi, 399, 485  
Xia, Xiaonan, 446, 455  
Xiao, Gang, 206  
Xu, Jiansheng, 688  
Xu, John, 48  
Xu, Jun, 679  
Xu, Ming, 688  
Xu, Yijie, 155  
Xu, Zeyang, 57  
Xu, Zhidong, 819  
Xu, Ziwen, 27

**Y**

- Yan, Jiaxuan, 242, 261, 270  
Yan, Jin, 225  
Yang, Guobao, 428  
Yang, Hongyong, 769  
Yang, Jiajun, 455  
Yang, Lin, 599

Yang, Shujun, 351, 792

- Yang, Xiansheng, 251  
Yang, Xueqing, 69, 155, 165, 175  
Yang, Yuanyuan, 760  
Ye, Linfei, 688  
Yi, Yang, 117, 145  
Yin, Hongpeng, 127  
Yin, Yanhui, 185  
You, Lihong, 88  
Yu, Li, 464  
Yu, Meiyang, 769  
Yu, Yongguang, 57, 367  
Yuan, Xiaolin, 367

**Z**

- Zhan, Jingyuan, 36  
Zhang, Bing, 399, 485  
Zhang, Haoran, 599, 608  
Zhang, Hongqiang, 495  
Zhang, Jun, 206  
Zhang, Kai, 297  
Zhang, Li, 778  
Zhang, Liyan, 399  
Zhang, Min, 608  
Zhang, Qing, 810  
Zhang, Qingyuan, 485  
Zhang, Shanshan, 358  
Zhang, Shaobo, 234  
Zhang, Tengfei, 1  
Zhang, Tianping, 117, 145, 541, 819  
Zhang, Weicun, 217  
Zhang, Xiaoli, 117  
Zhang, Yan, 409  
Zhang, Yunhao, 446  
Zhao, Huanyu, 508  
Zhao, Lin, 278, 317  
Zhao, Lingxue, 669  
Zhao, Long, 342, 560  
Zhao, Yong, 419  
Zhao, Yongguo, 308  
Zhao, Yuyin, 455  
Zheng, Jianqiang, 351, 792  
Zheng, Zewei, 527  
Zhou, Hao, 399, 485  
Zhou, Hongbiao, 508  
Zhou, Wuneng, 69, 155, 165, 175, 287, 518  
Zhou, Zhengxin, 760  
Zhu, Houjie, 99  
Zhu, Ming, 527  
Zhu, Zhiqin, 590  
Zou, Junchen, 99  
Zou, Shengrong, 377  
Zuo, Min, 801