



THE
UNIVERSITY OF
STRATHCLYDE
IN GLASGOW

IEEE
CACS

2015 21st INTERNATIONAL CONFERENCE ON AUTOMATION AND COMPUTING (ICAC)

Automation, Computing and Manufacturing for New Economic Growth

University of Strathclyde, Glasgow, UK
September 11-12, 2015

IEEE catalog number: USB-CFP1560R-USB

ISBN: 978-0-9926801-0-7



ICAC

2015 21st International Conference on Automation and Computing

University of Strathclyde, Glasgow, UK
September 11-12, 2015

IEEE Catalog Number: USB-CFP1560R-USB
ISBN: 978-0-9926801-0-7

2015 21st International Conference on Automation and Computing (ICAC)

Copyright © 2015 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Copyright and Reprint Permission

Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law, for private use of patrons, those articles in this volume that carry a code at the bottom of the first page, provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

Other copying, reprint, or reproduction requests should be addressed to IEEE Copyright Manager, IEEE Service Center, 445 Hoes Lane, P.O.Box 1331, Piscataway, NJ 08855-1331.

IEEE Catalog Number USB-CFP1560R-USB
ISBN 978-0-9926801-0-7

Additional copies of this publication are available from

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
+1 845 758 0400
+1 845 758 2633 (FAX)
email: curran@proceedings.com

Preface

Welcome to the 21st International Conference on Automation and Computing (ICAC'2015) held at University of Strathclyde.

The conference has a long successful history since its first conference in London (1995). The conference initially aimed to provide a forum for Chinese scientists, engineers, scholars and students in the UK to update technical knowledge and exchange ideas, and to provide a platform where joint research programs between those in both the UK and China can be formulated for mutual benefit. Since 2007, the conference had expanded towards a true international conference. In last few years, all papers presented in this conference series have entered the IEEE Xplore digital library so they can be searched by a much wide research community. This year the conference brings together researchers throughout the world, including Brazil, China, Czech Republic, Hong Kong, India, Ireland, Italy, Libya, Pakistan, Romania, Russia, Serbia, Singapore, Slovakia and others, to the UK to disseminate their scientific findings in all aspects of automation, computing, and manufacturing. We hope you have enjoyable time and find the conference stimulating and fruitful.

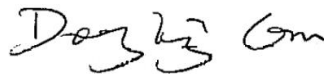
In general, the scope of the conference covers all the aspects of automation, computing, and manufacturing from fundamental research to engineering applications and advanced technologies. We are pleased to see the high quality of this year's papers, which contain both high level theoretical papers and practical application papers.

On behalf of the organising committee, we would like to take this opportunity to thank all those who have contributed to this conference, as well as the members of the organising committee and the international program committee for their fantastic work. Our sincere thanks also go to the conference sponsors for their suggestions and supports, particularly to International Journal of Automation and Computing for the sponsorship of best student paper awards. Finally, but not least, we want to thank many volunteers for their diligent works for the conference.

September 2015



Prof. Xichun Luo
Conference Chair



Prof. Dongbing Gu
Program Chair

Organizing Institutions

The Chinese Automation and Computing Society in the UK

University of Strathclyde, Glasgow, UK

Co-Sponsors

International Journal of Automation & Computing (IJAC)

Hefei University



Conference Chair

Xichun Luo University of Strathclyde

Program Chair

Dongbing Gu University of Essex

Conference Co-Chair

Hong Yue University of Strathclyde

Financial Chair

Lili Yang Loughborough University

Local Organisation Committee members

Xichun Luo Hong Yue Luis Rubio Wenlong Chang

Program and Organizing Committees

Baibing Li	Loughborough University	Bing Wang	Hull University
Dawei Gu	Leicester University	Dayou Li	University of Bedfordshire
Dingli Yu	Liverpool John Moores University	Dongling Xu	University of Manchester
Dongbing Gu	University of Essex	Erfu Yang	University of Strathclyde
Feng Dong	University of Bedfordshire	Fengshou Gu	Huddersfield University
Geyong Min	University of Exeter	Guiyun Tian	University of Newcastle
Guoping Liu	University of Glamorgan	Hong Wang	University of Manchester
Hong Yue	University of Strathclyde	Hongnian Yu	University of Bournemouth
Hongji Yang	Bath Spa University	Hui Yu	University of Portsmouth
Hujun Yin	University of Manchester	Huosheng Hu	University of Essex
Jiyin Liu	Loughborough University	Jianbo Yang	University of Manchester
Jianjun Gui	University of Essex	Jie Zhang	University of Newcastle
Jihong Wang	University of Warwick	Jinsheng Kang	Brunel university
Kai Cheng	Brunel Univrsity	Keliang Zhou	University of Glasgow
Leo Chen	Glasgow Caledonian University	Liangxiu Han	Manchester Metropolitan University
Lili Yang	Loughborough University	Luis Rubio	University of Strathclyde
Maozhen Li	Brunel Univrsity	Mianhong Wu	University of Derby
Qiang Xu	University of Huddersfield	Qinghua Wu	University of Liverpool
Qingde Li	Hull University	Quanmin Zhu	University of the West of England
Sen Wang	University of Essex	Sheng Chen	University of Southampton
Shengfeng Qin	University of Northumbria	Shengrong Bu	University of Glasgow
Shuanghua Yang	Loughborough University	Sijing Zhang	University of Bedfordshire
Tai Yang	University of Sussex	Wei Huang	University of Bedfordshire
Wenhan Zeng	University of Huddersfield	Wenhua Chen	Loughborough University
Wenlong Chang	University of Strathclyde	Wesam Jasim	University of Essex
Xiandong Ma	Lancaster University	Xiao Liu	University of College London
Xichun Luo	University of Strathclyde	Xun Chen	Liverpool John Moores University
Yang Dai	Coventry University	Yanmeng Xu	Burnel Univeristy
Yaochu Jin	University of Surrey	Yi Cao	Cranfield University
Yi Huang	University of Liverpool	Yong Yue	University of Bedfordshire
Yuchun Xu	Cranfield University	Yun Li	University of Glasgow
Zhen Tong	University of Strathclyde	Zhibin Yu	University of Glasgow
Zhijie Xu	University of Huddersfield	Zhiyong Zhang	DUVAS Technologies
Zidong Wang	Brunel Univrsity	Ziqiang Lang	University of Sheffield
Ziqiang Zhu	University of Sheffield		

Conference Awards

Best Student Paper Award in Automation

Best Student Paper Award in Computing

Best Student Paper Award in Manufacturer

TABLE OF CONTENTS

Keynote Speech

- 1 *Prof. William Edward Leithead, University of Strathclyde*
Optimal Operation of Large Wind Turbine Arrays through Farm-level Control
 - 2 *Prof. Kai Cheng, Brunel University*
Smart tooling, smart machines and smart manufacturing: working towards the Industry 4.0 and beyond
-

Paralell Sessions (A – 1, A – 2, A – 3, A – 4, A – 5)

A – 1 *Control Applications 1*

Chair: Octavian Stefan & Wenhua Chen

- 3 *Ashraf Khalil, Jihong Wang*
Stabilization of Load Frequency Control System Under Networked Environment
 - 9 *Octavian Stefan, Toma-Leonida Dragomir, Alexandru Codrean*
Observer-based Delay Compensation for Networked Control Systems – analysis and synthesis
 - 15 *Aiping Wang, Hong Wang, Ning Sheng, Xin Yin*
Performance Analysis for Operational Optimal Control for Complex Industrial Processes – the Square Impact Principle
 - 21 *Qichun Zhang, Zhuo Wang, Hong Wang*
Parametric Covariance Assignment using Reduced-order Closed-form Covariance Model
 - 27 *Adegboyega Ehinmowo, Yi Cao*
Stabilizing slug flow at large valve opening using active feedback control
-

A – 2 *System Diagnosis & Condition monitoring 1*

Chair: Xiandong Ma & Yanmeng Xu

- 33 *Long Zhang, Ziqiang Lang, Wen-Xian Yang*
Transmissibility Damage Indicator for Wind Turbine Blade Condition Monitoring
 - 37 *Peng Qian, Xiandong Ma, Yifei Wang*
Condition monitoring of wind turbines based on extreme learning machine
 - 43 *Abdulkarim Shaeboum, Samieh Abusaad, Niaoqing Hu, Fengshou Gu, Andrew D. Ball*
Detection and Diagnosis of Motor Stator Faults using Electric Signals from Variable Speed Drives
 - 49 *Mark Lane, Djon Ashari, Fengshou Gu, Andrew D. Ball*
Investigation of Motor Current Signature Analysis to Detect Motor Resistance Imbalances
 - 53 *Wuqiao Luo, Yun Li, Zhong Tian, Bo Gao, Ling Tong, Houjun Wang, Baoqing Zeng*
Survey of Greener Ignition and Combustion Systems for Internal Combustion Engines
-

A – 3 *Information systems 1*

Chair: Hui Yu & Lili Yang

- 59 *Sicong Ma, Siyan Li, Hongji Yang*
Creative Computing for Personalised Meta-Search Engine Based on Semantic Web
 - 65 *Xuan Wang, Hongji Yang*
Applying Semantic Web Technology to Poem Analysis
 - 71 *Lin Zou, Hongji Yang*
Creative Computing for Decision Making: Combining Game Theory and Lateral Thinking
 - 77 *Pree Thiengburanathum, Shuang Cang, Hongnian Yu*
A Decision Tree based Recommendation System for Tourists
 - 84 *Mohammad S. Hasan, Hongnian Yu*
Innovative Developments in HCI and Future Trends
-

A – 4 *Information systems 2*

Chair: Qingde Li & Shuanghua Yang

- 90 *Elias Eze, Sijing Zhang, Enjie Liu*
Message Dissemination Reliability in Vehicular Networks

- 96 *Kapil Kanwal, Ghazanfar Safdar, Shyqyri Haxha*
Joint Resource Blocks Switching Off and Bandwidth Expansion for Energy Saving in LTE Networks
- 102 *Zivorad Mihajlovic, Ana Joza, Vladimir Milosavljevic, Vladimir Rajs, Milos Zivanov*
Energy Harvesting Wireless Sensor Node for Monitoring of Surface Water
- 108 *Nebrase Elmrabit, Shuang Hua Yang, Lili Yang*
Insider Threats in Information Security Categories and Approaches
- 114 *Naila Naheed, Munam Ali Shah, Sijing Zhang*
Energy Efficiency in Smartphones: A Survey on Modern Tools and Techniques

A – 5 *Manufacture Systems 1*

Chair: Zhijie Xu & Xun Chen

- 120 *Peiran Jiang, Liquan Wang, Xichun Luo*
Dynamic Analysis of An Underwater Leveling-Gripping System of An Jacket Platform Under Offshore Environmental Loads
- 124 *Guangbo Hao, Ronan Hand, Xianwen Kong, Wenlong Chang*
Design of compliant parallel grippers using the position space concept for manipulating sub-millimeter objects
- 130 *Xavier Herpe, Ross Walker, Xianwen Kong, Matthew Dunnigan*
Analysis and Characterisation of a Kinematically Decoupled Compliant XY Stage
- 136 *Khaldoon F. Brethee, Jingwei Gao, Andrew D. Ball*
Analysis of Frictional Effects on the Dynamic Response of Gear Systems and the Implications for Diagnostics
- 145 *Nasha Wei, Fengshou Gu, Tie Wang, Guoxing Li, Yuandong Xu, Longjie Yang, Andrew D. Ball*
Characterisation of Acoustic Emissions for the Frictional Effect in Engines using Wavelets based Multi-resolution Analysis

Parallel Sessions (*P1 – 1, P1 – 2, P1 – 3, P1 – 4, P1 – 5*)

P1 – 1 Control Applications 2

Chair: Parikshit Singh & Hong Yue

- 151 *Vildan V. Abdullin, Dmitry A. Shnayder*
Implementation of an Advanced Heating Control System at the University Academic Building
- 155 *Parikshit Singh, Surekha Bhanot, Harekrishna Mohanta*
Self-Tuned Fuzzy Logic Control of a pH Neutralization Process
- 161 *Lanxiang Zhu, Feng Yu, Dingwen Yu, Dingli Yu*
Decentralised PI Controller Design and Tuning Approaches
- 167 *Mengling Wang, Hong Yue, Jie Bao, William E. Leithead*
LIDAR-based Wind Speed Modelling and Control System Design
- 173 *Jan Chalupa, Robert Grepl, Václav Sova*
Design of Configurable DC Motor Power-Hardware-In-the-Loop Emulator for Electronic-Control-Unit Testing
- 179 *Paolo Righettini, Roberto Strada, Ehsan KhademOlama, Sirin Valilou*
Symbolic Kinematic and Dynamic Modelling toolbox for Multi-DOF Robotic Manipulators

P1 – 2 System Diagnosis & Condition Monitoring 2

Chair: Yi Cao & Dingli Yu

- 185 *Alessandro Mariani, Kary Thanapalan, Peter Stevenson, Thomas Stockley, Jonathan Williams*
Techniques for Monitoring and Predicting the OCV for VRLA Battery systems
- 191 *Ruirong Zhang, Yanmeng Xu, David Harrison, John Fyson, Darren Southree, Anan Tanwilaisiri*
A Study of the Performance of the Combination of Energy Storage Fibres
- 194 *Qing Tao, Wenlei Sun, Jianxing Zhou, Jinsheng Kang*
Study on the Inherent Characteristics of Planetary Gear Transmissions
- 199 *Adel Jaber, Pavlos Lazaridis, Yong Zhang, David Upton, Hamed Ahmed, Umar Khan, Bakhtiar Saeed, Peter Mather, Robert Atkinson, Martin Judd, Ian Glover, Maria Fatima Queiroz Vieira*
Comparison of Contact Measurement and Free-Space Radiation Measurement of Partial Discharge Signals

- 203 *Xiang Tian, Gaballa M. Abdallaa, Ibrahim Rehab, Fengshou Gu, Tie Wang, Andrew D. Ball*
Diagnosis of Combination Faults in a Planetary Gearbox using a Modulation Signal Bispectrum based Sideband Estimator
- 209 *Ibrahim Rehab, Xiang Tian, Fengshou Gu, Andrew D. Ball*
A Study of Diagnostic Signatures of a Deep Groove Ball Bearing Based on Nonlinear Dynamic Model

P1 – 3 Computer Vision

Chair: Shengfeng Qin & Yang Dai

- 216 *Jianjun Gui, Dongbing Gu, Huosheng Hu*
Pose Estimation Using Visual Entropy
- 222 *Cheng Zhao, Huosheng Hu, Dongbing Gu*
Building a grid-point cloud-semantic map based on graph for the navigation of intelligent wheelchair
- 229 *Rodrigo D. C. Silva, George A. P. Thé, Fátima N. S. de Medeiros*
Geometrical and statistical feature extraction of images for rotation invariant classification systems based on industrial devices
- 235 *Nicolas S. Pereira, Cinthya R. Carvalho, George A. P. Thé*
Point cloud partitioning approach for ICP improvement
- 240 *Shengfeng Qin, Huaiwen Tian*
On the Algorithm For Reconstruction of Polyhedral Objects From a Single Line Drawing
- 246 *Ye Zhang, Huanzhi Lou, Hui Yu*
Morphology Elements Research on Chinese Small-Sized Liquor Bottle Design

P1 – 4 Special Session: Industry 4.0

Chair: Yun Li & Hongnian Yu

- 252 *Paulino Rocher*
Invited Talk: Towards Industry 4.0
- 253 *Joo Hock Ang*
Forum Discussion: Design Knowledge Capture, Optimisation and Automation to Advance Industry 4.0
- 254 *Milan Gregor, Jozef Herčko, Patrik Grznár*
The Factory of the Future Production System Research
- 260 *Alfredo Alan Flores Saldivar, Yun Li, Wei-neng Chen, Zhi-hui Zhan, Jun Zhang, Leo Yi Chen*
Industry 4.0 with Cyber-Physical Integration: A Design and Manufacture Perspective
- 266 *Richard Martin*
Invited Talk: Supercomputing for Industry 4.0

P1 – 5 Manufacture Systems 2

Chair: Jinsheng Kang & Tai Yang

- 267 *Javier Zamorano Igual, Qiang Xu*
Determination of the material constants of creep damage constitutive equations using Matlab optimisation procedure
- 273 *Xin Yang, Zhongyu Lu, Qiang Xu*
The Interpretation of Experimental Observation Data for the Development of Mechanisms based Creep Damage Constitutive Equations for High Chromium Steel
- 279 *Xiangyu Teng, Dehong Huo, Wai Leong Eugene Wong, Manoj Gupta*
Experiment based investigation into micro machinability of Mg based metal matrix composites (MMCs) with nano-sized reinforcements
- 285 *Jianfeng Huang, Zhonglai Wang, Yuanxin Luo, Yun Li, Erfu Yang, Yi Chen*
Computational Investigation of Superalloy Persistent Slip Bands Formation
- 291 *Amir Mir, Xichun Luo, Amir Siddiq*
Numerical Simulation of Triaxial Tests to Determine the Drucker-Prager Parameters of Silicon
- 295 *Wenlei Sun, Li Cao, Qing Tao, Yuanhua Tan*
The Design and Simulation of Beam Pumping Unit
-

Parallel Sessions ($P2 - 1, P2 - 2, P2 - 3, P2 - 4$)

P2 – 1 Control Applications 3

Chair: Ehsan Khadem Olama & Luis Rubio

- 299 Yaser Alothman, Wesam Jasim, Dongbing Gu
Quad-rotor Lifting-Transporting Cable-Suspended Payloads Control
- 305 Paolo Righettini, Roberto Strada, Shirin Valilou, Ehsan Khadem Olama
Output Feedback Sliding Mode Controller with H-2 Performance for Robot Manipulator
- 311 Pengcheng Liu, Hongnian Yu, Shuang Cang
On Periodically Pendulum-Driven Systems for Underactuated Locomotion: a Viscoelastic Jointed Model
- 317 Václav Sova, Jan Chalupa, Robert Grepl
Fault Tolerant BLDC Motor Control for Hall Sensors Failure
- 323 Robert Grepl, Michal Matejasko, Bastl M., Zouhar F.
Design of a Fault Tolerant Redundant Control for Electro Mechanical Drive System

P2 – 2 System Diagnosis & Condition Monitoring 3

Chair: Robert Grepl & Qiang Xu

- 329 Sulaiman A. Lawal, Jie Zhang
Actuator Fault Monitoring and Fault Tolerant Control in Distillation Columns
- 335 Ning Sheng, Hong Wang
Fault Detection and Diagnosis for Operational Control Systems
- 341 Raphael Samuel, Yi Cao
Improved Kernel Canonical Variate Analysis for Process Monitoring
- 347 Lanxiang Zhu, Feng Yu, Dingwen Yu, A.M.S. Ertiame, Dingli Yu
Solution to Failure Detection of Closed-loop Systems and Application to IC Engines
- 353 Dianwei Wang, Jing Wang, Ying Liu, Zhijie Xu
An Adaptive Time-frequency Filtering Algorithm for Multi-component LFM Signals based on Generalized S-transform

P2 – 3 Information Systems 3

Chair: Dianwei Wang & Erfu Yang

- 359 Tai Yang, Zhong Li, Yu Shu
Applying Feedback to Stock Trading: Exploring A New Field of Research
- 365 Chinedu Eze, Tai Yang, Chris Chatwin, Dong Yue, Hong Yu
Research into Big Data for Smart Grids
- 371 Yang Pang, Kwaku Opong, Luiz Moutinho, Yun Li
Cash Flow Prediction Using a Grey-Box Model
- 377 Muhammad Bilal Shahid, Munam Ali Shah, Sijing Zhang, Safi Mustafa, Mushahid Hussain
Organization Based Intelligent Process Scheduling Algorithm (OIPSA)
- 383 Kumaran Rajarathinam, J. Barry Gomm, Dingli Yu, Ahmed Saad Abdelhadi
An Improved Search Space Resizing Method for Model Identification by Standard Genetic Algorithm

P2 – 4 Information Systems 4

Chair: Vildan Abdullin & Leo Chen

- 389 Wenbin Zhong, Wenlong Chang, Luis Rubio, Xichun Luo
Reconfigurable software architecture for a hybrid micro machine tool
- 393 Zhenying Xu, Baozhong Wu, Meng Zhang, Feng Zou, Yuehui Yan
Measurement Station Planning of Single Laser Tracker based on PSO
- 399 Naji Al-Messabi, Cindy Goh, Yun Li
Grey-box Identification for Photovoltaic Power Systems via Particle-Swarm Algorithm
- 406 Joo Hock Ang, Cindy Goh, Yun Li
Key Challenges and Opportunities in Hull Form Design Optimization for Marine and Offshore Applications

413 Author Index

Optimal Operation of Large Wind Turbine Arrays through Farm-level Control

Professor William Edward Leithead

University of Strathclyde, UK

Abstract

With the development of large offshore wind farms and attainment of high wind power penetration, it is no longer satisfactory for wind farms to be passive providers of generated power. Instead, offshore wind farms must become virtual generation plant that behave similarly to conventional generation. The power must be adjusted as required by the operators. To do so, requires flexible operation of the individual turbines and a wind farm controller to match power output to demand.

In addition to adjusting the power output, the wind farm controller could enable the wind farm to provide ancillary services such as curtailment, frequency support, voltage support, etc. Furthermore, there is extensive information regarding the local environment and conditions, including SCADA information, environmental information (wind direction, time of year, sea state etc., maintenance and repair logs information, wind farm layout information, condition monitoring and turbine health information) as well as individual wind turbine control information from nearby turbines. The potential to exploit this information through the wind farm controller to enable operators to make the most of their assets is substantial. In its most sophisticated form, the wind farm controller could control the operation of the individual wind turbines to achieve the most effective short and long term operation of all the assets in the wind farm. Accordingly, the general objectives for the wind farm controller is to maximise wind farm generated power, provide ancillary services, including curtailment, frequency support and voltage support, and minimise O&M costs.

An approach to wind farm control, that is hierarchical, decentralise and scalable, is presented.

Biographical sketch of the speaker

Professor William Edward Leithead was appointed in 1986 to the post of Lecturer in the Department of Electronic and Electrical Engineering at the University of Strathclyde. He was promoted to Senior Lecturer in 1990, Reader in 1995 and Professor in 1999. Prof. Leithead became Director of the Industrial Control Centre in 2006. He leads the wind energy research group within the department. Between 2002 and 2006, he was on secondment to assist with the establishment of the Hamilton Institute in Ireland. Prof. Leithead is a Member of the International Federation of Automatic Control Power Systems Committee and is on the Editorial Board of the international journal, Wind Energy. He has been the recipient of more than 40 research grants and is the author of more than 200 academic publications.

Since 1988, when Prof. Leithead established the wind energy group, it has grown to approximately 80 researchers and 10 academic staff and is now one of the leading wind energy research groups in Europe. His research interests in wind energy include the dynamic analysis of wind turbines, their dynamic modelling and simulation, control system design and optimisation of wind turbine design. Prof. Leithead has strong links to all aspects of the Wind Energy industry, manufacturers, component suppliers, developers, utilities and consultancies. He has been involved in the design process of a number of commercial machines and in many collaborative projects, including wind turbine controller design projects and 3 Energy Technology Institute projects, as a partner.

From October 2009 Prof. Leithead has been the Director of the EPSRC Centre for Doctoral Training (CDT) in Wind Energy Systems which is hosted by the wind energy group at Strathclyde University. He is, also, a member of the Executive Committee of the EPSRC Industrial Doctoral Centre in Offshore Renewable Energy and Chair of the EPSRC Supergen V Wind Energy Technologies Consortium. He is the Board Member for the UK of the European Academy of Wind Energy, Wind Energy Coordinator of the Energy Technology Partnership and sits on many national and international Wind Energy research advisory boards including the European Energy Research Alliance Joint Programme Wind Steering Committee, Scientific Advisory Board of the Norwegian Centre for Offshore Wind Technology (Trondheim), Scientific Advisory Board of the Norwegian Centre for Offshore Wind Energy (Bergen), Strategy Advisory Group Energy Technology Institute and Scottish Government Offshore Wind Industry Group.

Smart tooling, smart machines and smart manufacturing: Working towards the Industry 4.0 and beyond

Professor Kai Cheng

Brunel University, UK

Abstract

Smart manufacturing has tremendous potential and is becoming the key enabling technology for the future advanced manufacturing particularly in the Industry 4.0 context. Smart manufacturing processes are operated with some characteristics, e.g. being intelligent, connected, managed and secured. It will enable a new level of manufacturing capability and adaptability, including high process reliability, high precision, machining process optimization, plug-and-produce operations, and bespoke high value applications, etc.

This presentation will present some innovative design concepts and, in particular, the development of a number of smart tooling devices and smart machines, and their intrinsic relation and impact to smart manufacturing at an industrial scale. Practical implementation and application perspectives for these smart tooling and smart machines are explored and discussed, taking account of the requirements for smart manufacturing against a number of industrial applications, such as contamination-free machining, high speed smart drilling, machining of tool-wear-prone varifocal lenses and medical applications. Additional research on smart tooling implementation and application perspectives will also be presented, including: (a) plug-and-produce design principle, (b) novel cutting force modelling and the associated implementation algorithms, (c) piezoelectric film and surface acoustic wave transducers to measure cutting forces, (d) critical cutting temperature reduction and control in real-time machining, (e) Multi-physics based design and analysis of smart tooling and smart machines, and (f) application exemplars on smart machining.

The presentation will conclude with further discussion on the potentials and applications of smart tooling and smart machines development for Industry 4.0 and future manufacturing.

Biographical sketch of the speaker

Professor Kai Cheng holds the chair professorship in Manufacturing Systems at Brunel University. His current research interests focus on precision and micro manufacturing, design of high precision machines, smart tooling and smart machines, and sustainable manufacturing and systems. Professor Cheng has published over 180 papers in learned international journals and referred conferences, authored/edited 6 books and contributed 6 book chapters.

Professor Cheng is a fellow of the IMechE and IET. He is the head of the Micro/Nano Manufacturing Theme at Brunel University, which consists of 12 academics and over 50 research assistants/fellows and PhD students. Professor Cheng and the team are currently working on a number of research projects funded by the EPSRC, EU 7th Framework Programs, Innovate UK, Royal Academy of Engineering, KTP Programs and the industry. Professor Cheng is the European editor of the International Journal of Advanced manufacturing Technology and a member of the editorial board of International Journal of Machine Tools and Manufacture. Professor Cheng is also honored with the visiting professorship at Harbin Institute of Technology.

Stabilization of Load Frequency Control System Under Networked Environment

Ashraf Khalil

Electrical and Electronic Engineering Department
University of Benghazi
Benghazi, Libya
ashraf.khalil@uob.edu.ly

Jihong Wang

School of Engineering
University of Warwick
Coventry, UK
jihong.wang@warwick.ac.uk

Abstract—The deregulation of the electricity market made the open communication infrastructure an exigent need for future power system. In this scenario dedicated communication links are replaced by shared networks. These shared networks are characterized by random time delay and data loss. The random time delay and data loss may lead to system instability if they are not considered during the controller design stage. Load frequency control systems used to rely on dedicated communication links. To meet future power system challenges these dedicated networks will be replaced by open communication links which make the system stochastic. In this paper the stochastic stabilization of load frequency control system under networked environment is investigated. The shared network is represented by three states which are governed by Markov chains. A controller synthesis method based on the stochastic stability criteria is presented in the paper. A one-area load frequency control system is chosen as case study. The effectiveness of the proposed method for the controller synthesis is tested through simulation. The derived PI controller proves to be optimum where it is a compromise between compensating the random time delay effects and degrading the dynamic performance.

Keywords—component; networked control system; load frequency control; LFC; Markov chains; robust stabilization;

I. INTRODUCTION

The current challenges in power system and the recent advances in communication networks are the key drivers for adopting new open communication infrastructure in power systems [1]. The load frequency control (LFC) is one of the classical centralized power system control problems. The main goals of the LFC are: 1) To maintain uniform frequency, 2) Share the load between the generators, and 3) Control the tie-line interchange schedule [2]. Automatic Generation Control (AGC) and LFC system have been implemented in centralized scheme since the beginning of the interconnected power system [3]. The LFC is achieved by the AGC where the frequency deviation is used to sense the change in the load demand. In the AGC, a dedicated communication link is used to send the AGC signals. In the case of fault in the dedicated communication link, other communication links are used, usually voice communications through telephone lines [4]. Because of the increased number of ancillary services, the need for a duplex and distributed communication links becomes more pronounced [4].

Future power system needs to be decentralized, integrated, flexible, and open [5]. Communication

networks have grown rapidly while power system control centers remains far behind [5]. One of the promising solutions is the migrating from traditional SCADA (Supervisory Control and Data Acquisition) to TCP/IP (Transmission control protocol/ Internet Protocol) and Ethernet. The TCP/IP is becoming the de facto world standard for data transmission [1]. Open communication networks are reconfigurable and the hardware and the software are well developed. The migrating from traditional dedicated networks to open and distributed networks such as the Internet has been proposed by several system operators but they are still not adopted because of their non-deterministic characteristics. Although these new technologies are more flexible, reconfigurable, and have high bandwidth, there are some shortcomings. The main issues in the open communication infrastructure are the time delay, the data loss and the vulnerability to malicious attack [1]. The introduction of the open communication which is a shared network raises concerns about the stability of the LFC system.

To guarantee that the frequency is within the permissible range the Area Control Error (ACE) and the Generator Control Error (GCE) signals are distributed between the different areas through these shared networks. The ACE and the GCE are used to increase or decrease the generated power. Sending these signals through the communication link will introduce time delay and some of these data will be lost. One of the requirements of the future communication network is to be fault tolerant. Then the most important part for controller designer is to guarantee the stability of the LFC system with the time delay and data loss. There are some research works regarding the stability of the LFC system in the presence of the time delay and they are summarized in the following.

The methods reported in the literature focuses on estimating the maximum time delay margin which is the time delay that the LFC system can withstand before it becomes unstable. In the published research works the methods for linear time delay are applied to estimate the maximum time delay margin for LFC systems. There are mainly two types of methods that deal with linear time delay systems; the indirect methods based on Lyapunov stability theory and the direct methods which are based on tracking the eigenvalues of the characteristics equation. In [4] the authors used the simulation to study the impact of the time delay on the LFC system stability and discuss the communication requirement for the LFC system. They

considered both constant and random time delays. Although they investigated the impacts of the packet loss on the stability of the LFC system, they did not propose any method to estimate the control system requirements in terms of the maximum time delay margin. Their simulations show that the increased number of packet dropouts can result in system instability. The authors in [6] used a simple stability criterion but rather conservative that was reported in the 1980s and introduced in [7-8]. In [6] the problem of the LFC is formulated as a general time delay problem which is then solved in Linear Matrix Inequality (LMI) form, but it is only applicable to constant time delays. In [9] the time delay margin for LFC is calculated using stability criteria for linear time delay systems with variable time delay for one-area and multi-area LFC system. The effects of the PI controller, K_p and K_i , on the time delay margin are investigated for one-area and multi-area LFC system. In [10] the Genetic algorithms are used to tune the controllers and the authors consider the generation rate constraint (GRC), the dead band, and the time delay. The sliding mode control is used in [11] to design robust LFC system against uncertainty and time delay where the H_∞ optimal control is used to derive the sliding surface parameters. The general predictive control is investigated in [12] to compensate the effects of the bounded random time delay.

Most of the published research works in the literature concentrate on constant and varied delays. The random delay has not been considered in many papers and some of them treat LFC system under open communication as normal time delay system where the sampling, time delay and data dropouts are not considered. The random time delay and data loss should be considered during the LFC system design stage. The method presented in the paper is based on designing the controller while taking the stochastic nature of the network into consideration [13]. Under dedicated conventional closed communication links the assumptions of constant or varied time delay is justified, however, the time delay and data loss in open communication network are random and in many cases can be modeled using Markov chains. The paper focuses on the stochastic stabilization of LFC system under networked environment. The paper starts with the modeling of the LFC system with random time delay and packet loss. The LFC system with the random time delay and data dropouts is modeled as Standard Markovian Discrete-Time Jump Linear System. The stochastic stability of this type of systems is briefly discussed. Then the stability criterion is formulated as Bilinear Matrix Inequalities (BMIs). The BMIs are solved using V-K iteration method. In the V-K iteration method the problem is divided into three LMIs which are solved using Matlab. Simulation results show the merit of the proposed controller design method.

II. DYNAMIC MODEL OF ONE-AREA LFC SYSTEM WITH RANDOM TIME DELAY

The one-area LFC system is shown in Fig. 1. The main assumption is that all the generators are equipped with non-reheat turbines. The state-space linear model of one-area LFC system is expressed as [9]:

$$\begin{aligned}\dot{\bar{\mathbf{x}}}(t) &= \mathbf{A}_c \bar{\mathbf{x}}(t) + \mathbf{B}_c \mathbf{u}(t) + \mathbf{F}_c \Delta P_d \\ \bar{\mathbf{y}}(t) &= \mathbf{C}_c \bar{\mathbf{x}}(t)\end{aligned}\quad (1)$$

where;

$$\begin{aligned}\bar{\mathbf{x}}(t) &= [\Delta f \quad \Delta P_m \quad \Delta P_v]^T & \bar{\mathbf{y}}(t) &= ACE \\ \mathbf{A}_c &= \begin{bmatrix} -\frac{D}{M} & \frac{1}{M} & 0 \\ 0 & -\frac{1}{T_{ch}} & \frac{1}{T_{ch}} \\ -\frac{1}{RT_g} & 0 & -\frac{1}{T_g} \end{bmatrix} & \mathbf{B}_c &= \begin{bmatrix} 0 \\ 0 \\ \frac{1}{T_g} \end{bmatrix} & \mathbf{F}_c &= \begin{bmatrix} -\frac{1}{M} \\ 0 \\ 0 \end{bmatrix} \\ \mathbf{C}_c &= [\beta \quad 0 \quad 0]\end{aligned}$$

The parameters are given as: ΔP_d is the load deviation, ΔP_m is the generator mechanical output deviation, ΔP_v is the valve position deviation, Δf is the frequency deviation. M is the moment of inertia, D is the generator damping coefficient, T_g is the time constant of the governor, T_{ch} is the time constant of the turbine, R is the speed drop, and β is the frequency bias factor. For one-area LFC system, the area control error ACE is given as:

$$ACE = \beta \Delta f \quad (2)$$

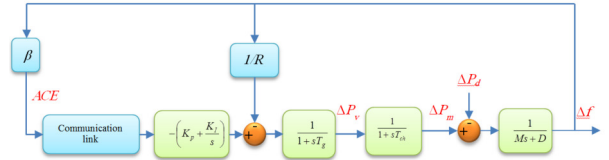


Figure 1. Dynamic model of one-area LFC scheme

The AGC has two components; the first is updated every 5 minutes for economical dispatch and the second is updated in the order of 1-5 seconds [6]. The later signal delay is the one considered in the paper. Stabilizing the system with conventional PI controller given as:

$$u(t) = -K_p ACE - K_i \int ACE \quad (3)$$

where K_p is the proportional gain, K_i is the integral gain and $\int ACE$ is the integration of the area control error. With the PI controller, the closed-loop system is expressed as follows:

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t - \tau(t)) + \mathbf{F}\Delta P_d \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t)\end{aligned}\quad (4)$$

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix} -\frac{D}{M} & \frac{1}{M} & 0 & 0 \\ 0 & -\frac{1}{T_{ch}} & \frac{1}{T_{ch}} & 0 \\ -\frac{1}{RT_g} & 0 & -\frac{1}{T_g} & 0 \\ \beta & 0 & 0 & 0 \end{bmatrix} & \mathbf{B} &= \begin{bmatrix} 0 \\ 0 \\ \frac{1}{T_g} \\ 0 \end{bmatrix} & \mathbf{F} &= \begin{bmatrix} -\frac{1}{M} \\ 0 \\ 0 \\ 0 \end{bmatrix} \\ \mathbf{C} &= \begin{bmatrix} \beta & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} & \mathbf{x}(t) &= [\Delta f \quad \Delta P_m \quad \Delta P_v \quad \int ACE]^T\end{aligned}$$

Digitizing system (3) and the controller (4) with sampling time, T_s , (3) and (4) becomes:

$$\mathbf{x}(k+1) = \mathbf{A}_d \mathbf{x}(k) + \mathbf{B}_d \mathbf{u}(k) + \mathbf{F}_d \Delta P_d(k) \quad (5)$$

$$\mathbf{u}(k) = \mathbf{K}(r_s(k))\mathbf{x}(k - r_s(k)) \quad (6)$$

where $\tau(k) = r_s(k) \cdot h$, h is the sampling period and $r_s(k)$ is a bounded random integer sequence governed by Markov Chain with $0 \leq r_s(k) \leq d_s < \infty$, and d_s represents the finite delay bound and the number of modes. \mathbf{A}_d , \mathbf{B}_d and \mathbf{F}_d are matrices with appropriate size and depend on the sampling rate. Before we proceed the following assumptions are made: The random time delay in the network is bounded, the number of data dropouts is finite, all the data are sent as single packet, and the time stamping is used where the old data are discarded. Introducing the augmenting state variable given as:

$$\bar{\mathbf{x}}(k) = [\mathbf{x}(k)^T \quad \mathbf{x}(k-1)^T \quad \cdots \quad \mathbf{x}(k-d_s)^T]^T$$

where $\bar{\mathbf{x}}(k) \in R^{(d_s+1)n}$, applying the controller (6) into (5) the closed-loop system becomes:

$$\bar{\mathbf{x}}(k+1) = (\bar{\mathbf{A}} + \bar{\mathbf{B}}\mathbf{K}(r_s(k))\bar{\mathbf{C}}(r_s(k)))\bar{\mathbf{x}}(k) \quad (7)$$

where;

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_d & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix} \quad \bar{\mathbf{B}} = \begin{bmatrix} \mathbf{B}_d \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

$$\bar{\mathbf{C}}(r_s(k)) = [\mathbf{0} \quad \cdots \quad \mathbf{0} \quad \mathbf{C}_d \quad \mathbf{0} \quad \cdots \quad \mathbf{0}]$$

$\bar{\mathbf{C}}(r_s(k))$ has all elements being zero except for the $r_s(k)^{\text{th}}$ block equals \mathbf{C}_d . Notice that the time delay and data loss are incorporated into $\bar{\mathbf{C}}(r_s(k))$. The closed-loop system (7) can be rewritten as;

$$\bar{\mathbf{x}}(k+1) = \mathbf{A}_{cl}(r_s(k))\bar{\mathbf{x}}(k) \quad (8)$$

The system represented by (8) is the standard Discrete-Time Markovian Jump Linear System (DTMJLS). The Markovian jump system is mostly used to study the stability and stabilization of system with abrupt changes [13]. The open communication network is modeled as a finite state Markov process with the following properties:

$$P\{r_s(k+1) = j \mid r_s(k) = i\} = p_{ij} \quad 0 \leq i, j \leq d_s$$

$$0 \leq p_{ij} \leq 1 \quad \sum_{j=0}^d p_{ij} = 1 \quad (9)$$

It should be noted that (9) incorporates the packets dropouts. p_{ij} is the transition probability from mode i to mode j . The general transition probability matrix is given by:

$$P = \begin{bmatrix} p_{00} & p_{01} & 0 & 0 & \cdots & 0 \\ p_{10} & p_{11} & p_{12} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 \\ p_{d0} & p_{d1} & p_{d2} & p_{d3} & \cdots & p_{d,d_s} \end{bmatrix} \quad (10)$$

The constraint (9) means the summation of the probabilities in every row is one. The assumption made is that the old data are discarded, this can be interpreted as;

$$p\{r_s(k+1) > r_s(k) + 1\} = 0 \quad (11)$$

From (11) the time delay can increase only at one step but it can decrease many steps as can be seen from (10). The diagonal elements in (10) represent the probability of successive equal time delays. The upper diagonal elements represent the possibility of receiving longer delays or increasing the network load. The zero elements represent the discard of the old data.

III. STOCHASTIC STABILITY OF LFC SYSTEM WITH TIME DELAY AND PACKET DROPOUTS

The LFC system (8) is a stochastic hybrid system. The stochastic stability, mean square stability and the exponential mean square stability are all equivalent and every condition implies the almost sure (asymptotic) stability [14].

Definition 1: [14]

The system (8) is mean square stable if for every initial condition state, $(\bar{\mathbf{x}}_0, r_0)$,

$$\lim_{k \rightarrow \infty} E\|\bar{\mathbf{x}}(k)\|^2 = 0 \quad (12)$$

Definition 2: [14]

The system (8) is mean square stable with decay rate β if for every initial condition state, $(\bar{\mathbf{x}}_0, r_0)$,

$$\lim_{k \rightarrow \infty} \beta^k E\|\bar{\mathbf{x}}(k)\|^2 = 0 \quad \beta > 1 \quad (13)$$

The necessary and sufficient conditions for mean square stability for jump system are given in the following theorem.

Theorem 1 [14]: The mean square stability of (8) is equivalent to the existence of symmetric positive definite matrices $\mathbf{Q}_0, \dots, \mathbf{Q}_d$ satisfying the following condition:

$$\sum_{j=0}^d p_{ji} \mathbf{A}_i^T \mathbf{Q}_j \mathbf{A}_i < \mathbf{Q}_i, \quad i = 0, \dots, d$$

Replacing \mathbf{Q}_i by $\alpha \mathbf{Q}_i$ (where the decay rate or Lyapunov Exponent, $\beta = 1/\alpha$ and $\lim_{k \rightarrow \infty} \beta^k M(k) = 0$) on the right hand side, the closed-loop system becomes:

$$\sum_{j=0}^d p_{ji} (\mathbf{A}_i + \mathbf{B}_i \mathbf{K}_i \mathbf{C}_i)^T \mathbf{Q}_j (\mathbf{A}_i + \mathbf{B}_i \mathbf{K}_i \mathbf{C}_i) < \alpha \mathbf{Q}_i \quad (14)$$

where $i=0, \dots, d$. The coupled equations (14) are BMIs which are nonconvex and finding a global optimal solution is very difficult. The most widely used techniques for the solution is by iteration methods such the D-K, G-K and V-K iteration algorithms [15-16]. If we fix \mathbf{K}_i ($i = 0, \dots, d$) then we have a Generalized Eigenvalue Problem (GEVP) and if we fix \mathbf{Q}_i ($i = 0, \dots, d$) then we have Eigenvalue Problem (EVP) [15-16]. Both of these problems can be solved very efficiently using Matlab LMI toolbox. Equation (14) can be written as:

$$\alpha \mathbf{Q}_j - \begin{bmatrix} \tilde{\mathbf{A}}_0^T & \tilde{\mathbf{A}}_1^T & \dots & \tilde{\mathbf{A}}_d^T \\ p_{j0} \mathbf{Q}_0 & & & 0 \\ & & \ddots & \\ 0 & & & p_{jd} \mathbf{Q}_d \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{A}}_0 \\ \tilde{\mathbf{A}}_1 \\ \vdots \\ \tilde{\mathbf{A}}_d \end{bmatrix} > 0 \quad (15)$$

where $j = 0, 1, \dots, d$, $\tilde{\mathbf{A}}_i = \mathbf{A}_i + \mathbf{B}_i \mathbf{K}_i \mathbf{C}_i$. Using Schur complement to (15) then we have:

$$\begin{bmatrix} \alpha \mathbf{Q}_j & \tilde{\mathbf{A}}_0^T & \dots & \tilde{\mathbf{A}}_d^T \\ \tilde{\mathbf{A}}_0 & p_{j0}^{-1} \mathbf{Q}_0^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{A}}_d & 0 & \dots & p_{jd}^{-1} \mathbf{Q}_d^{-1} \end{bmatrix} > 0 \quad (16)$$

In the V-K algorithm the BMI is divided into two LMI's and by solving these two LMI's a local optimal solution can be found. The problem solution process is divided into three basic problems which are: Feasibility Problem (FP), Eigenvalue Problem (EVP), and Generalized Eigenvalue Problem (GEVP). These problems can be solved using the Matlab LMI toolbox. In the V-K algorithm, the problem is iterated between the EVP and the GEVP. The proof of the algorithm convergence is given in [15]. The detailed algorithm is shown in the flowchart in Fig 2. The algorithm starts with the initialization, and then if the solution is feasible the EVP and GEVP are iterated until the desired transition matrix is reached. In this improved algorithm the decay rate is maximized in both the EVP and the GEVP iterations. The initial transition probability matrix is chosen to be:

$$\mathbf{P}_0 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} \approx \begin{bmatrix} 1-n\epsilon & \epsilon_1 & \dots & \epsilon_n \\ 1-n\epsilon & \epsilon_1 & \dots & \epsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ 1-n\epsilon & \epsilon_1 & \dots & \epsilon_n \end{bmatrix}$$

It should be noted that the initial controller is designed for the free delay system and hence the initial solution is feasible. To get an initial feasible solution we have to start from small time delays and perturb the transition probability matrix toward longer time delays. The perturbation ϵ should be very small positive number in the order of 0.005.

IV. CASE STUDY: ONE-AREA LFC SYSTEM

The parameters of the LFC system shown in Fig. 1 are given as: $T_{ch}=0.3$, $T_g=0.1$, $R=0.05$, $D=1.0$, $\beta=21.0$ and $M=10$. Under open communication network the remote terminal unit (RTU) sends the signals to the central controller through the shared network, and then the controller sends the commands back. The two delays defined as feed forward and feedback delays. In most of the studies these two delays are aggregated into a single delay and this assumption is made in this paper [4,6]. In power systems the data collection is in the order of 1-5 s [1] and in the US the ACE signal is transmitted every 4 seconds [4]. In the light of these facts the sampling time is chosen to be 1 second. The shared network is chosen to have three different states; namely: Low, Medium and High.

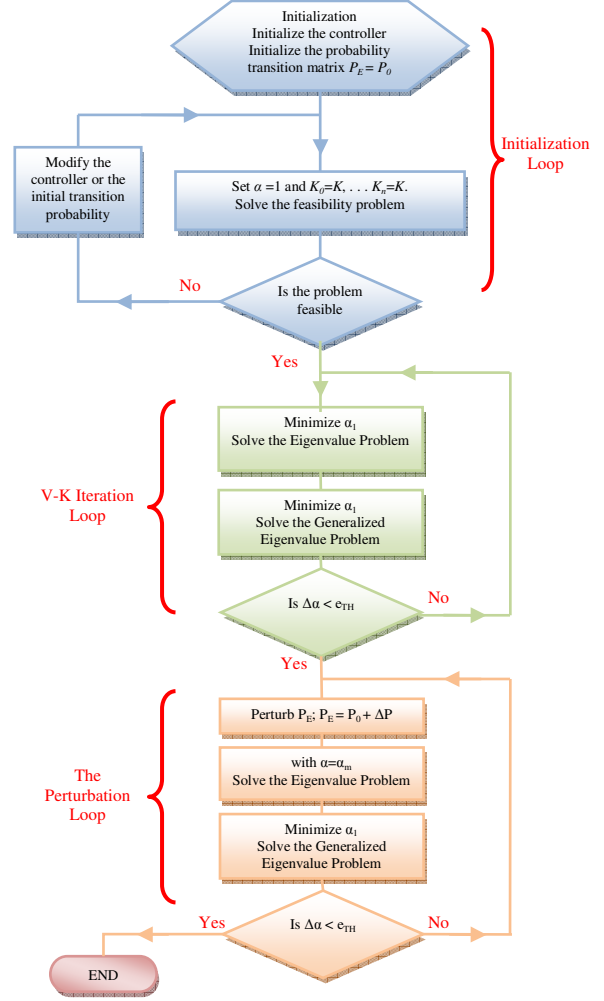


Figure 2. The V-K iteration algorithm

Firstly, the stochastic stability range of the PI controller gains as function of the sampling rate is investigated in the paper and is shown in Fig. 3. For positive values of the PI controller gains the stochastic stability region of the LFC system is semicircle. Fig. 3 is obtained by changing the values of the PI control gains (K_p and K_i) and solving the feasibility problem. The transition probability matrix for the network is given by:

$$\mathbf{P} = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.1 & 0.65 & 0.25 \\ 0.1 & 0.65 & 0.25 \end{bmatrix} \quad (17)$$

Increasing the sampling rate widens the stability region of the LFC system. This means with high sampling rates we can select K_p and K_i to be large which improves the LFC system performance. On the other hand increasing the sampling rate increases the load on the network which increases the time delay and data loss. With low sampling rates the stability region becomes small. Lower sampling rates degrade the performance of the system and a compromise between the performance of the LFC system and the sampling rate should be made. It should be noted that the size of the stability region has strong dependence on K_i .

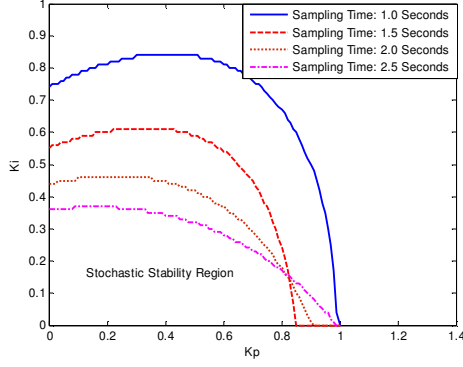


Figure 3. The stability region with different sampling rates

Choosing the initial PI controller as: $K_p = 0.8$ and $K_i = 0.8$. Using transition probability matrix with the following perturbation matrix:

$$P = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.4 & 0.5 & 0.1 \\ 0.4 & 0.5 & 0.1 \end{bmatrix} \quad \Delta P = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ -0.005 & 0.0025 & 0.0025 \\ -0.005 & 0.0025 & 0.0025 \end{bmatrix}$$

After 60 iterations the final controller gains are $K_p = 0.0057$ and $K_i = 0.178$ and the decay rate is: $\beta = 1.549$. Using theorem 1 in [17] for constant time delay the time delay margin is increased from 0.3519 s to 5.2329 s. The average time delay in the network is 1.16 s. The frequency deviation with and without random time delay compensation controller is shown in Fig. 4. Although the system without random time delay compensation is stable with this random time delay, it becomes unstable with different random seed because the initial controller fails to stabilize the system. This observation has been reported in [4]. The random time delay increases the oscillation in the frequency deviation and in the accumulative area control error as can be seen in Fig. 5.

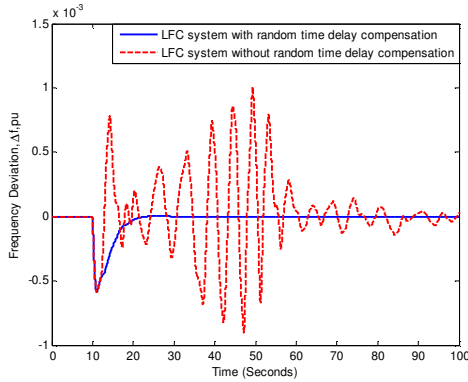


Figure 4. The frequency deviation, Δf

The random time delay is shown in Fig. 6. The frequency deviation with the initial and the final controller without time delay is shown in Fig. 7. The performance of the initial controller without time delay is better than the performance of the system with the final controller. For the system with the random time delay the final controller compensates the effects of the random time delay without degrading the system performance.

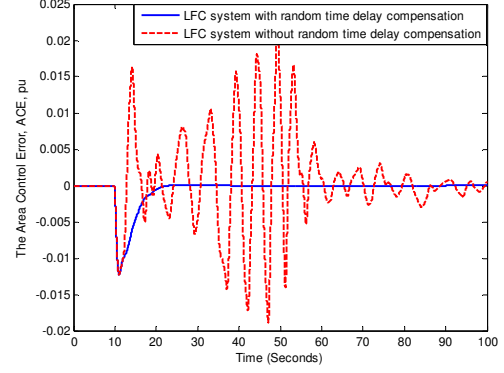


Figure 5. The Area Control Error, ACE

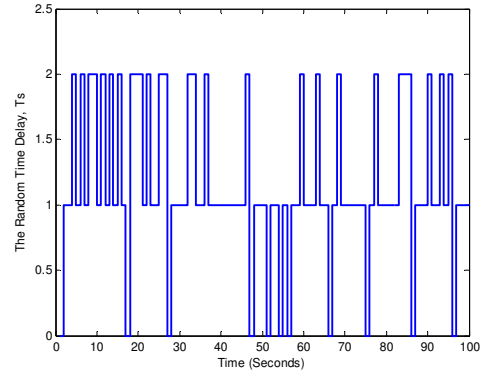


Figure 6. The random time delay

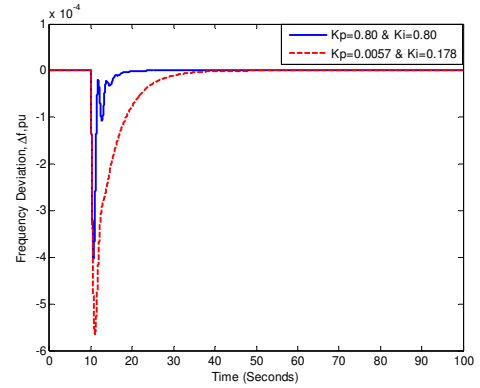


Figure 7. The frequency deviation, Δf

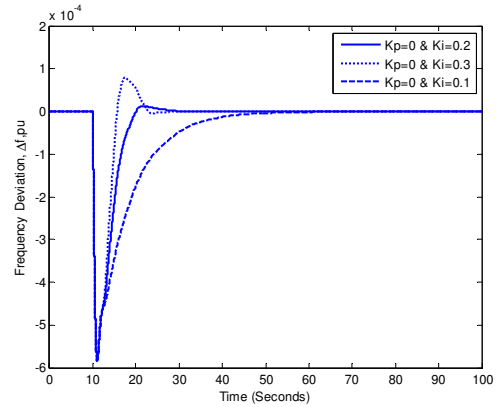


Figure 8. The frequency deviation, Δf

It should be noted that the PI controller derived through the V-K iteration lies in the optimum region ($K_p = 0.0057$ and $K_i = 0.178$), this fact proves that the V-K iteration achieves the optimum controller.

In the studies reported for LFC system with constant and varied time delay the results show that the maximum time delay decreases with increasing the integral gain while our results for the random time delay show that there is an optimum value for K_i . The frequency deviation of the LFC system with different values of K_p and K_i are shown in Fig. 8. The response shows that the optimum value of K_p and K_i is achieved with the V-K iteration. Increasing the integral gain beyond 0.2 increases the oscillations because of the random time delay impact. Decreasing the integral gain below 0.2 makes the response of the LFC system slower and hence increases the ACE.

The controller derived through the V-K iteration decreases the oscillation resulting from the random time delay dramatically. These large oscillations may damage the equipment; additionally they degrade the LFC system performance, cause overload on transmission lines and cause interference to the protection system [18]. They can even lead to power system collapse. Practically the PI controller gains are chosen to be large in order to achieve faster time response, however under random time delay these values must kept low. The work in this paper shows the feasibility of achieving the stability of LFC system under random time delay which mimics the control of LFC system under networked environment. The challenge for open communication networks is to guarantees the reliability, availability and immunity against malicious attack. Enhancing the availability, reliability and security of the shared networks will make the migrating from conventional dedicated communication links to open communication network possible. In this paper the conventional PI controller is used and in future studies different controllers can be used.

V. CONCLUSION

In this paper the stochastic stabilization of one-area LFC system is investigated. Under open communication network the time delay and data loss are usually random and in many cases can be modeled using Markov chains. The conventional LFC system with Markovian random time delay is modeled as Standard Markovian Linear Jump System. The stochastic stability of the LFC system is formulated as Bilinear Matrix Inequality which is non-convex. The BMIs are divided into three LMIs problems and the V-K iteration algorithm is used to derive the controller. The stabilizing controller shows good performance despite the existence of the random time delay. Furthermore the derived controller achieves the

optimum performance as proved through simulations. The controller design method is to be extended to multi-area LFC system. Also the H_∞ performance could be used in the V-K iteration algorithm.

REFERENCES

- [1] Mak, K.-H. & Holland, B. 2002. "Migrating electrical power network SCADA systems to TCP/IP and Ethernet networking". *Power Engineering Journal*, 16, (6) 305-311
- [2] Hadi Saadat, "Power system analysis", McGraw Hill Companies, 1999.
- [3] Fardanesh, B. 2002. "Future trends in power system control". *IEEE Computer Applications in Power*, 15, (3) 24-31
- [4] S. Bhowmik, K. Tomovic, and A. Bose, "Communication models for third party load frequency control," *IEEE Trans. Power Syst.*, 19, (1), pp. 543-548, Feb. 2004.
- [5] Wu, F. F., Moslehi, K., & Bose, A., 2005. "Power system control centers: Past, present, and future", 11 edn, IEEE., pp. 1890-1908.
- [6] Yu, X. & Tomovic, K. 2004. "Application of linear matrix inequalities for load frequency control with communication delays". *IEEE Transactions on Power Systems*, 19, (3) 1508-1515
- [7] Mori, T. 1985. "Criteria for asymptotic stability of linear time-delay systems". *IEEE Transactions on Automatic Control*, AC-30, (2) 158-161
- [8] Mori, T. & Kokame, H. 1989. "Stability of $\dot{x}(t)=Ax(t)+Bx(t-\tau)$ ". *IEEE Transactions on Automatic Control*, 34, (4) 460-462
- [9] L. Jiang, W. Yao, Q. H. Wu, J. Y. Wen, S. J. Cheng, "Delay-dependent stability for load frequency control with constant and time-varying delays", *IEEE Trans. Power Syst.*, vol. 27, no. 2, 2012, pp. 932-941.
- [10] H. Golpîra, H. Bevrani, H. Golpîra, "Application of GA optimization for automatic generation control design in an interconnected power system", *Energy Conversion and Management* 52 (2011) 2247-2255
- [11] K. Vrdoljak, I. Petrovic, and N. Peric, "Discrete-time sliding mode control of load frequency in power systems with input delay," in *Proc. 12th Int. Power Electronics and Motion Control Conf., EPE-PEMC, Potoroz, Slovenia, Aug. 2006*, pp. 567-572.
- [12] J. H. Zhang, J. H. Hao, G. L. Hou, "Automatic generation controller design in deregulated and networked environment using predictive control strategy", *Proceedings of the 17th World Congress The International Federation of Automatic Control* Seoul, Korea, July 6-11, 2008, 9410-9415.
- [13] A. F. Khalil and J. Wang, "Robust stabilization of networked control systems using the markovian jump system approach," in *The Proceeding of the United Kingdom International Control Conference (IEEE)*, September 2012, 316-321.
- [14] L. El Ghaoui and M. A. Rami, "Robust state-feedback stabilization of jump linear systems via LMIs," *International Journal of Robust and Nonlinear Control*, vol. 6, no. 9-10, pp. 1015-1022, Nov.1996.
- [15] David Banjerdpongchai, "parametric robust controller synthesis using linear matrix inequalities." University of Stanford, 1997.
- [16] X. Lin, A. Hassibi, and J. P. How, "Control with random communication delays via a discrete-time jump system approach," in *Proceedings of the 2000 American Control Conference. ACC*, vol.3 ed Danvers, MA, USA: American Autom. Control Council, 2000, pp. 2199-2204.
- [17] A. F. Khalil and W. Jihong, "A new stability and time-delay tolerance analysis approach for Networked Control Systems," in *The Proceeding of the 49th IEEE conference on Control and Decision* 2010, pp. 4753-4758.
- [18] H. Bevrani, G. Ledwich, J. J. Ford and Z. Y. Dong, "On power system frequency control in emergency conditions" *Journal of Electrical Engineering & Technology*, Vol.3, No.4, pp.499-508, 2008.

Observer-based Delay Compensation for Networked Control Systems – analysis and synthesis

Octavian Stefan, Toma-Leonida Dragomir
Department of Automation and Applied Informatics
Politehnica University Timisoara
Timisoara, Romania
{octavian.stefan, toma.dragomir}@upt.ro

Alexandru Codrean
Department of Automation
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
alexandru.codrean@aut.utcluj.ro

Abstract — The current study focuses on the stability of a generic observer-based delay compensation structure for networked control systems with time-varying delays. After a brief presentation of the design principles, the stability conditions are derived in terms of maximum delay bounds, using a Lyapunov functional. Experimental results validate the entire approach on a specific case study.

Keywords — networked control systems; time delay; disturbance observers; stability analysis;

I. INTRODUCTION

Networked control systems (NCS) gained an increasing attention from the scientific community in the last years because of the rapid development of communication technology ([1], [2]). Low cost, high performance and reliability make the NCS a viable solution for distributed control systems.

Although NCS have a lot of advantages over the conventional control systems, there are also some shortcomings induced by the network component like time varying delays, information loss, limited communication capacity, that tend to complicate the design and analysis phases of the NCS ([3]). Multiple control strategies have been developed by the scientific community in order to overcome these shortcomings. The most important of them are based on: robust control ([4]), optimal stochastic control ([5]), event based control ([6]), model predictive control ([7]), gain scheduling ([8]) and adaptive Smith predictor ([9]).

One solution of interest for time-varying delay compensation in NCS considers the transmission delay as an unknown additive disturbance at the input of the process [10]. The disturbance is estimated by using a communication disturbance observer (CDOB) and then its effect on the control system is filtered from the controller's point of view. The main advantage of this approach is that it needs no a priori information about the time delay instantaneous values or variation speed.

Although several particular CDOB-based network control structures have been designed and analyzed in previous studies, the current study addresses the stability analysis and control synthesis of a generic observer-based delay compensation structure for the general case involving time-varying delays. In

[10], [11] and subsequent studies the stability of a CDOB-based NCS was proven only for constant time delay values.

The remainder of this paper is organized as follows. Section II introduces the observer-based network control structure. Section III presents the NCS's control design. Section IV analyses the NCS's stability. Section V presents an illustrative example. Section VI states some final conclusions.

II. OBSERVER-BASED NETWORKED CONTROL STRUCTURE

The considered observer-based NCS structure, proposed in [12], is presented in Fig. 1. The aim of the networked control structure is to reject the disturbance effect of the network and of the local disturbance, in order to ensure the imposed process control behavior. At the local side containing the process, the disturbance d is compensated by a feedback loop composed of a disturbance observer (DOB) and a disturbance compensator (DCO). The controller, placed at the remote side, is separated from the process by the network, considered as a discrete-time nonlinear system.

As a design hypothesis, both channels of the network are modeled as time-varying delay elements and the process as linear time invariant (LTI). As consequence, first, the round trip time (RTT) delay of the network can be obtained by combining both time delay elements. Second, the effect of the RTT is considered as a delay disturbance d_n , acting at input of the process, and defined as the difference between the transmitted control signal u and the received one u_n ([10])

$$d_n(t) = u_n(t) - u(t), \quad (1)$$

with

$$u_n(t) = u(t - \tau). \quad (2)$$

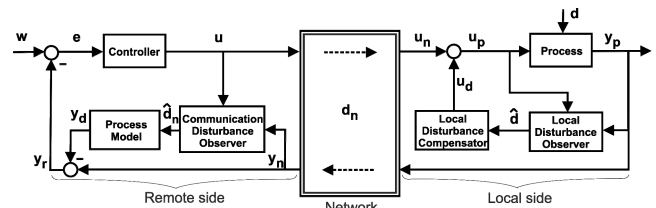


Fig. 1. Observer-based networked control structure

The current study presents the design principles for the NCS from Fig. 1, analysis the stability of the network control structure, and finally pursues an extensive validation, with respect to the network delay bounds which ensure stability.

III. CONTROL DESIGN

Consider a LTI process of the form

$$\begin{cases} \dot{\mathbf{x}}_p(t) = \mathbf{A}_p \mathbf{x}_p(t) + \mathbf{b}_p u_p(t) + \mathbf{b}_{pd} d(t) \\ y_p(t) = \mathbf{c}_p^T \mathbf{x}_p(t) \end{cases}, \quad (3)$$

with $\mathbf{x}_p \in \mathfrak{R}^p$, $u_p \in \mathfrak{R}$, $d \in \mathfrak{R}$, $y_p \in \mathfrak{R}$ and the output feedback controller

$$\begin{cases} \dot{\mathbf{x}}_c(t) = \mathbf{A}_c \mathbf{x}_c(t) + \mathbf{b}_c (w(t) - y_r(t)) \\ u(t) = \mathbf{c}_c^T \mathbf{x}_c(t) + d_c (w(t) - y_r(t)) \end{cases}, \quad (4)$$

with $\mathbf{x}_c \in \mathfrak{R}^c$ and $w \in \mathfrak{R}$.

The controller is designed as if it were directly connected to the process ($u = u_p$, $y_r = y_p$).

Next, consider the local feedback loop (DOB+DCO), with the disturbance d acting on the process. For slow variations in time of d , a first order exogenous system, $\dot{d}(t)=0$, can be used in the design of the DOB. The observer's parameters are then adopted such that the estimated disturbance converges to the real disturbance d . The DCO is designed to ensure complete local disturbance compensation in steady state regime. The local feedback loop can be framed in the state space form as

$$\begin{cases} \dot{\mathbf{x}}_{co}(t) = \mathbf{A}_{co} \mathbf{x}_{co}(t) + \mathbf{b}_{coc} u_p(t) + \mathbf{b}_{cop} y_p(t) \\ u_d(t) = \mathbf{c}_{co}^T \mathbf{x}_{co}(t) + d_{co} y_p(t) \end{cases}. \quad (5)$$

The CDOB provides an estimate for the network disturbance \hat{d}_n defined as in (1). Because \hat{d}_n is induced by digital network transmissions, it can be regarded as a staircase signal that can be obtained from a first order exogenous system $\dot{\hat{d}}_n(t)=0$, which is then included in the CDOB. The CDOB parameters are chosen such that:

i) the transfer function that relates the output \hat{d}_n to the input d_n should behave as a low pass filter with a sufficiently high cut-off frequency (i.e. $\hat{d}_n \rightarrow d_n$ for an imposed frequency domain);

ii) the transfer function that relates the output \hat{d}_n to the input d should attenuate low frequency signal components in order to reject the residual error of the local compensation loop.

The CDOB can be described by the following state-space model

$$\begin{cases} \dot{\mathbf{x}}_o(t) = \mathbf{A}_o \mathbf{x}_o(t) + \mathbf{b}_o u(t) + \mathbf{b}_{op} y_p(t) \\ \hat{d}_n(t) = \mathbf{c}_o^T \mathbf{x}_o(t) + d_o y_p(t) \end{cases}. \quad (6)$$

As a final step, the CDOB is coupled with the process model

$$\begin{cases} \dot{\mathbf{x}}_{mp}(t) = \mathbf{A}_p \mathbf{x}_{mp}(t) + \mathbf{b}_p \hat{d}_n(t) \\ y_d(t) = \mathbf{c}_p^T \mathbf{x}_{mp}(t) \end{cases} \quad (7)$$

in order to reject the delay disturbance from the controller's point of view.

IV. STABILITY ANALYSIS

In order to analyze the stability of the networked control structure from Fig. 1, first, three simplifying assumptions will be made:

A₁: The local disturbance d and the reference w are assumed to be null.

A₂: The network is idealized as a time-varying delay transfer element placed on the direct path, with the delay equal to the RTT. Consequently, the influence of data loss and the communication channels' limited capacity are neglected (the packet loss rate is assumed to be negligible with respect to the sample period, while the digital bandwidth of the channel is assumed to be sufficiently large).

A₃: The RTT delay is assumed to be bounded: $0 \leq \tau(t) \leq \tau_{max}$.

Because $w(t)=0$, the input to the controller C is $-y_r(t) = -(y_n(t) - y_d(t))$. As specified in assumption A₂, the network is replaced by a single time varying delay on the direct path (equal to the RTT), such that $u_n(t) = u(t - \tau(t))$ and $y_n(t) = y_p(t)$.

The closed loop model of the NCS is given by

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{A}_d \mathbf{x}(t - \tau(t)), \quad (8)$$

where $\mathbf{x}(t) = [\mathbf{x}_p^T(t) \quad \mathbf{x}_{co}^T(t) \quad \mathbf{x}_o^T(t) \quad \mathbf{x}_{mp}^T(t) \quad \mathbf{x}_c^T(t)]^T$ and

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_p - \mathbf{b}_p d_{co} \mathbf{c}_p^T & -\mathbf{b}_p \mathbf{c}_{co}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{b}_{cop} \mathbf{c}_p^T - \mathbf{b}_{coc} d_{co} \mathbf{c}_p^T & \mathbf{A}_{co} - \mathbf{b}_{coc} \mathbf{c}_{co}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{b}_{oc} d_c \mathbf{c}_p^T + \mathbf{b}_{op} \mathbf{c}_p^T & \mathbf{0} & \mathbf{A}_o & \mathbf{b}_{oc} d_c \mathbf{c}_p^T & \mathbf{b}_{oc} \mathbf{c}_c^T \\ \mathbf{b}_p d_o \mathbf{c}_p^T & \mathbf{0} & \mathbf{b}_p \mathbf{c}_o^T & \mathbf{A}_p & \mathbf{0} \\ -\mathbf{b}_c \mathbf{c}_p^T & \mathbf{0} & \mathbf{0} & \mathbf{b}_c \mathbf{c}_p^T & \mathbf{A}_c \end{bmatrix}, \quad (9)$$

$$\mathbf{A}_d = \begin{bmatrix} \mathbf{b}_p d_c \mathbf{c}_p^T & \mathbf{0} & \mathbf{0} & \mathbf{b}_p d_c \mathbf{c}_p^T & \mathbf{b}_p \mathbf{c}_c^T \\ \mathbf{b}_{coc} d_c \mathbf{c}_p^T & \mathbf{0} & \mathbf{0} & \mathbf{b}_{coc} d_c \mathbf{c}_p^T & \mathbf{b}_{coc} \mathbf{c}_c^T \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

The system (8) is now in standard form for time delay systems, for which several stability methods were investigated in the literature in the last two decades. The methods can be classified according to three categories: delay independent, delay dependent and rate dependent, delay dependent and rate independent. Because in practice the delay is usually bounded in a certain range (A_3), and the rate of the delay variation is usually unknown, here the focus will be on the delay dependent and rate independent case.

The aim is to find the maximum range $[0, \tau_{\max}]$ in which the time delay can vary, for which the NCS is still stable. In other words, to find the maximum value of τ_{\max} for which the system is stable. To this end, the following theorem gives the sufficient conditions for stability for a given delay range $[0, \tau_{\max}]$. The theorem, along with the proof, is an adaption of the results from [13].

Theorem 1: The system (8) with time-varying delay $\tau(t)$ of upper bound τ_{\max} is asymptotically stable if there exist symmetric positive definite matrices \mathbf{P} , \mathbf{Q} , \mathbf{Z} , and matrices \mathbf{N}_1 , \mathbf{N}_2 , \mathbf{S}_1 , and \mathbf{S}_2 such that the following linear matrix inequality (LMI) holds

$$\begin{bmatrix} \mathbf{P}\mathbf{A} + \mathbf{A}^T\mathbf{P} + \mathbf{Q} + \mathbf{N}_1 + \mathbf{N}_1^T & \mathbf{P}\mathbf{A}_d + \mathbf{N}_2^T - \mathbf{N}_1 + \mathbf{S}_1 & -\mathbf{S}_1 & \tau_{\max}\mathbf{N}_1 & \tau_{\max}\mathbf{S}_1 & \tau_{\max}\mathbf{A}^T\mathbf{Z} \\ * & \mathbf{S}_2 + \mathbf{S}_2^T - \mathbf{N}_2 - \mathbf{N}_2^T & -\mathbf{S}_2 & \tau_{\max}\mathbf{N}_2 & \tau_{\max}\mathbf{S}_2 & \tau_{\max}\mathbf{A}_d^T\mathbf{Z} \\ * & * & -\mathbf{Q} & 0 & 0 & 0 \\ * & * & * & -\tau_{\max}\mathbf{Z} & 0 & 0 \\ * & * & * & * & -\tau_{\max}\mathbf{Z} & 0 \\ * & * & * & * & * & -\tau_{\max}\mathbf{Z} \end{bmatrix} < 0 \quad (10)$$

where * stands for symmetric term in a symmetric matrix.

Proof: Consider the Lyapunov functional candidate ([13])

$$\begin{aligned} V(\mathbf{x}_t) = & \mathbf{x}^T(t)\mathbf{P}\mathbf{x}(t) + \int_{t-\tau_{\max}}^t \mathbf{x}^T(s)\mathbf{Q}\mathbf{x}(s) ds \\ & + \int_{-\tau_{\max}}^0 \int_{t+\theta}^t \mathbf{x}^T(s)\mathbf{Z}\dot{\mathbf{x}}(s) ds d\theta \end{aligned} \quad (11)$$

where $\mathbf{P} = \mathbf{P}^T > 0$, $\mathbf{Q} = \mathbf{Q}^T > 0$, $\mathbf{Z} = \mathbf{Z}^T > 0$ and $\mathbf{x}_t \in C([t_a, t_b], \mathbb{R}^n)$ defined by $\mathbf{x}_t(\theta) = \mathbf{x}_t(t + \theta)$, with $-\tau_{\max} \leq \theta \leq 0$, $t \in [t_a + \tau_{\max}, t_b]$. $C([t_a, t_b], \mathbb{R}^n)$ is a Banach space of continuous functions mapping the interval $[t_a, t_b]$ into \mathbb{R}^n , with the norm $\|\mathbf{x}_t\| = \sup_{[-\tau_{\max} \leq \theta \leq 0]} \|\mathbf{x}_t(\theta)\|_2$.

Based on the Leibniz-Newton formula, the following equations hold for any matrices \mathbf{N}_1 , \mathbf{N}_2 , \mathbf{S}_1 , and \mathbf{S}_2 (weighting matrices)

$$\begin{aligned} & 2[\mathbf{x}^T(t)\mathbf{N}_1 + \mathbf{x}^T(t - \tau(t))\mathbf{N}_2] \\ & \left[\mathbf{x}(t) - \mathbf{x}(t - \tau(t)) - \int_{t-\tau(t)}^t \dot{\mathbf{x}}(s) ds \right] = 0 \end{aligned} \quad (12)$$

$$\begin{aligned} & 2[\mathbf{x}^T(t)\mathbf{S}_1 + \mathbf{x}^T(t - \tau(t))\mathbf{S}_2] \\ & \left[\mathbf{x}(t - \tau(t)) - \mathbf{x}(t - \tau_{\max}) - \int_{t-\tau_{\max}}^{t-\tau(t)} \dot{\mathbf{x}}(s) ds \right] = 0 \end{aligned} \quad (13)$$

Additionally, the following equation also holds

$$-\int_{t-\tau_{\max}}^t \dot{\mathbf{x}}^T(s)\mathbf{Z}\dot{\mathbf{x}}(s) ds = -\int_{t-\tau(t)}^t \dot{\mathbf{x}}^T(s)\mathbf{Z}\dot{\mathbf{x}}(s) ds - \int_{t-\tau_{\max}}^{t-\tau(t)} \dot{\mathbf{x}}^T(s)\mathbf{Z}\dot{\mathbf{x}}(s) ds \quad (14)$$

By making use of Leibniz's integral rule, the derivative of $V(\mathbf{x}_t)$ along the solutions of (8) can be written as

$$\begin{aligned} \dot{V}(\mathbf{x}_t) = & 2\mathbf{x}^T(t)\mathbf{P}\dot{\mathbf{x}}(t) + \mathbf{x}^T(t)\mathbf{Q}\mathbf{x}(t) - \mathbf{x}^T(t - \tau_{\max})\mathbf{Q}\mathbf{x}(t - \tau_{\max}) \\ & - \mathbf{Q}\mathbf{x}(t - \tau_{\max}) + \tau_{\max}\dot{\mathbf{x}}^T(t)\mathbf{Z}\dot{\mathbf{x}}(t) - \int_{t-\tau_{\max}}^t \dot{\mathbf{x}}^T(s)\mathbf{Z}\dot{\mathbf{x}}(s) ds \end{aligned} \quad (15)$$

The further use of equations (12) - (14), and after some calculations and regrouping yield

$$\begin{aligned} \dot{V}(\mathbf{x}_t) \leq & 2\mathbf{x}^T(t)\mathbf{P}\dot{\mathbf{x}}(t) + \mathbf{x}^T(t)\mathbf{Q}\mathbf{x}(t) - \mathbf{x}^T(t - \tau_{\max})\mathbf{Q}\mathbf{x}(t - \tau_{\max}) \\ & + \tau_{\max}\dot{\mathbf{x}}^T(t)\mathbf{Z}\dot{\mathbf{x}}(t) - \int_{t-\tau(t)}^t \dot{\mathbf{x}}^T(s)\mathbf{Z}\dot{\mathbf{x}}(s) ds - \int_{t-\tau_{\max}}^{t-\tau(t)} \dot{\mathbf{x}}^T(s)\mathbf{Z}\dot{\mathbf{x}}(s) ds \\ & + 2[\mathbf{x}^T(t)\mathbf{N}_1 + \mathbf{x}^T(t - \tau(t))\mathbf{N}_2] \left[\mathbf{x}(t) - \mathbf{x}(t - \tau(t)) - \int_{t-\tau(t)}^t \dot{\mathbf{x}}(s) ds \right] \\ & + 2[\mathbf{x}^T(t)\mathbf{S}_1 + \mathbf{x}^T(t - \tau(t))\mathbf{S}_2] \left[\mathbf{x}(t - \tau(t)) - \mathbf{x}(t - \tau_{\max}) - \int_{t-\tau_{\max}}^{t-\tau(t)} \dot{\mathbf{x}}(s) ds \right] \\ & \leq \zeta^T(t) [\Gamma + \bar{\mathbf{A}}^T \tau_{\max} \mathbf{Z} \bar{\mathbf{A}} + \tau_{\max} \mathbf{N} \mathbf{Z}^{-1} \mathbf{N}^T + \tau_{\max} \mathbf{S} \mathbf{Z}^{-1} \mathbf{S}^T] \zeta(t) \\ & - \underbrace{\int_{t-\tau(t)}^t [\zeta^T(t) \mathbf{N} + \dot{\mathbf{x}}^T(s) \mathbf{Z}] \mathbf{Z}^{-1} [\mathbf{N}^T \zeta(t) + \mathbf{Z} \dot{\mathbf{x}}(s)] ds}_{> 0 \text{ since } \mathbf{Z} > 0} \\ & - \underbrace{\int_{t-\tau_{\max}}^{t-\tau(t)} [\zeta^T(t) \mathbf{S} + \dot{\mathbf{x}}^T(s) \mathbf{Z}] \mathbf{Z}^{-1} [\mathbf{S}^T \zeta(t) + \mathbf{Z} \dot{\mathbf{x}}(s)] ds}_{> 0 \text{ since } \mathbf{Z} > 0} \end{aligned} \quad (16)$$

where

$$\begin{aligned} \zeta(t) = & \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{x}(t - \tau(t)) \\ \mathbf{x}(t - \tau_{\max}) \end{bmatrix}, \mathbf{N} = \begin{bmatrix} \mathbf{N}_1 \\ \mathbf{N}_2 \\ \mathbf{0} \end{bmatrix}, \mathbf{S} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \mathbf{0} \end{bmatrix}, \bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A}^T \\ \mathbf{A}_d^T \\ \mathbf{0} \end{bmatrix}, \\ \Gamma = & \begin{bmatrix} \mathbf{P}\mathbf{A} + \mathbf{A}^T\mathbf{P} + \mathbf{Q} + \mathbf{N}_1 + \mathbf{N}_1^T & \mathbf{P}\mathbf{A}_d + \mathbf{N}_2^T - \mathbf{N}_1 + \mathbf{S}_1 & -\mathbf{S}_1 \\ * & \mathbf{S}_2 + \mathbf{S}_2^T - \mathbf{N}_2 - \mathbf{N}_2^T & -\mathbf{S}_2 \\ * & * & -\mathbf{Q} \end{bmatrix} \end{aligned} \quad (17)$$

The last two integral terms can be dropped, and (16) becomes

$$\begin{aligned} \dot{V}(\mathbf{x}_t) \leq & \zeta^T(t) [\Gamma + \bar{\mathbf{A}}^T \tau_{\max} \mathbf{Z} \bar{\mathbf{A}} + \tau_{\max} \mathbf{N} \mathbf{Z}^{-1} \mathbf{N}^T + \tau_{\max} \mathbf{S} \mathbf{Z}^{-1} \mathbf{S}^T] \zeta(t) \\ & \leq \zeta^T(t) \mathbf{\Omega} \zeta(t) \end{aligned} \quad (18)$$

The condition $\mathbf{\Omega} < 0$ is equivalent to (10) by Schur complements. Thus, if (10) holds, then $\dot{V}(\mathbf{x}_t) < -\varepsilon \|\mathbf{x}_t\|^2$ for a

sufficiently small $\epsilon > 0$, and as a result the system (8) is asymptotically stable.

V. CASE STUDY

Consider the networked control structure from Fig. 1, composed of an electric drive as the controlled process, a PI controller, a TCP/IP network, two Lunberger observers ([14]) and a non-inertial local disturbance compensator. As control objective, the motor speed (y_p) must follow the reference w , despite of the disturbances (load disturbance d and network disturbance d_n) affecting the control system.

A. Numerical setting

The electric drive is composed of a permanent magnet DC motor supplied by an electronic actuator (PWM) and a tachogenerator. The model associated with the electric drive has the form (3) with $\mathbf{x}_p^T = [x_{p1} \ x_{p2}]$ (x_{p1} is the motor speed, x_{p2} is the armature current) and

$$\mathbf{A}_p = \begin{bmatrix} 0 & T_m^{-1} \\ -T_a^{-1} & -T_a^{-1} \end{bmatrix}, \mathbf{b}_p = \begin{bmatrix} 0 \\ T_a^{-1} \end{bmatrix}, \mathbf{b}_{pd} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \mathbf{c}_p^T = [1 \ 0]. \quad (19)$$

The model (4) of the PI controller is defined by

$$\mathbf{A}_c = 0, \mathbf{b}_c = 1, \mathbf{c}_c^T = K_i, \mathbf{d}_c = K_p. \quad (20)$$

The local feedback loop (DCO+DOB) modeled by (5) has

$$\mathbf{A}_{co} = \begin{bmatrix} -\frac{T_m + T_a l_{a1}}{T_a T_m} & l_{a1} \\ -\frac{l_{a2}}{T_m} & l_{a2} \end{bmatrix}, \mathbf{b}_{coc} = \begin{bmatrix} \frac{1}{T_a} \\ 0 \end{bmatrix}, \quad (21)$$

$$\mathbf{b}_{cop} = \begin{bmatrix} \frac{l_{a1}^2 T_a - l_{a1} l_{a2} T_a T_m + (l_{a1} - 1) T_m}{T_a T_m} \\ \frac{l_{a2} l_{a1} - l_{a2}^2 T_m}{T_m} \end{bmatrix},$$

$$\mathbf{c}_{co}^T = [0 \ K_a] \mathbf{d}_{co} = -K_a l_{a2}.$$

The DOB is a reduced order observer ([15]) and the DCO is the proportional element K_a . Because the observer implies the change of variables $\mathbf{x}_{co} = \mathbf{x}_{DOB} + l_a y_p$, with $\mathbf{x}_{DOB}^T = [\hat{x}_{p2} \ \hat{d}]$ and $\mathbf{x}_{co}^T = [x_{co1} \ x_{co2}]$, the output is actually $u_d = K_a \hat{d}$.

The CDOB is a full order observer of type (6) with

$$\mathbf{A}_o = \begin{bmatrix} -l_{n1} & T_m^{-1} & 0 \\ -T_a^{-1} l_{n2} & -T_a^{-1} & T_a^{-1} \\ -l_{n3} & 0 & 0 \end{bmatrix}, \mathbf{b}_{oc} = \begin{bmatrix} 0 \\ T_a^{-1} \\ 0 \end{bmatrix}, \quad (22)$$

$$\mathbf{b}_{op} = \begin{bmatrix} l_{n1} \\ l_{n2} \\ l_{n3} \end{bmatrix}, \mathbf{c}_o^T = [1 \ 0 \ 0], \mathbf{d}_o = 0$$

and the state vector $\mathbf{x}_o^T = [\hat{x}_{p1} \ \hat{x}_{p2} \ \hat{d}_n]$.

All the parameters of the NCS are given in Table I.

The manner in which the design parameters were obtained will be disclosed in the next section.

B. Design principles

This section briefly presents the main design principles regarding the NCS. The design follows the steps explained in detail in [12].

A PI controller was chosen, which ensures a null steady state error, in order to show the robustness (with respect to the network disturbance) of the NCS from Fig. 1 with a simple (classic) controller. The controller parameters were obtained by imposing a time constant $T_i = K_p / K_i$ in order to compensate the inertia of the process and a gain K_p according to a desired settling time.

The design parameters of the local feedback loop composed of DOB and DCO are adopted in order to satisfy the static and dynamic requirements of the local closed loop system. Based on Fig. 1 and the models of the process, DOB and DCO, the estimated disturbance can be obtained as

$$\hat{d}(s) = \frac{T_m l_{a2} (T_a s + 1)}{\underbrace{T_m T_a s^2 + (T_m + l_{a1} T_a - T_a T_m l_{a2}) s - l_{a2} T_m}_{H_d(s)}} d(s). \quad (23)$$

The process output y_p expressed as

$$y_p(s) = H_u(s) u_p(s) + H_d(s) d(s) = \frac{1}{T_a T_m s^2 + T_m s + 1} u_p(s) - \frac{(1 + T_a) T_m}{T_a T_m s^2 + T_m s + 1} d(s), \quad (24)$$

with the aid of $u_p(s) = u_n(s) + K_a \hat{d}(s)$, can be brought to

$$y_p(s) = H_d(s) \left[\underbrace{H_u(s) / H_d(s)}_{H_\alpha(s)} \cdot u_n(s) + \left(\underbrace{H_u(s) K_a H_a(s) / H_d(s)}_{H_\alpha(s)} + 1 \right) d(s) \right]. \quad (25)$$

TABLE I. CONTROL SYSTEM'S PARAMETERS

Parameter	Value
Pwr*	20W
J**	5.18·10 ⁻⁶ kg m ²
y _p ^{max}	4500 rpm
T _m	0.157 sec
T _a	0.039 sec
K _p	0.460
K _i	1.590
K _a	0.157
l _{a1}	-0.300
l _{a2}	-13.260
l _{n1}	102.040
l _{n2}	410.010
l _{n3}	461.430

* Pwr – motor's power, ** J – moment of inertia

The static condition for disturbance rejection is that $H_a(0)=0$, which leads to $K_a=T_m$. The parameters l_{a1} and l_{a2} are further adopted such that H_a acts as a high pass filter (Fig. 2).

Next, a full order CDOB was preferred, instead of a reduced order one, because it leads to better estimation performance in practice, due to its filtering capacity ([12]). Starting from Fig. 1, the models for the process, DOB, DCO and CDOB, the estimated network disturbance can be calculated as

$$\hat{d}_n(s) = H_n(s)d_n(s) + \underbrace{(T_a T_m s + T_m)H_\alpha(s)}_{H_\beta(s)} H_n(s)d(s), \quad (26)$$

with

$$H_n(s) = \frac{-l_{n3}}{T_a T_m s^3 + T_m (1 + l_{n1} T_a) s^2 + (l_{n1} T_m + l_{n2} T_a + 1) s + l_{n3}}. \quad (27)$$

The conditions for which $\hat{d}_n \rightarrow d_n$ are

$$\begin{cases} \text{Condition 1: } |H_n(j\omega)| \rightarrow 1 \\ \text{Condition 2: } |H_\beta(j\omega)H_n(j\omega)| \rightarrow 0 \\ \text{Condition 3: CDOB stable \& faster than the process} \end{cases}.$$

The first condition implies that H_n should have a magnitude close to unity for a frequency domain as large as possible (i.e. \hat{d}_n should be as close as possible to d_n). The second condition implies that across the imaginary axes $H_n H_\beta$ should have a magnitude as small as possible (d should not influence the estimation of d_n). The third condition implies that the poles of the observer have negative real part, with absolute values larger than the process poles.

The design parameters l_{n1} , l_{n2} , l_{n3} were adopted by imposing a cut-off frequency $\omega_n=37$ rad/s for H_n (Fig. 3). Note that a compromise has been made in order to avoid the peaking phenomenon due to large observer gains, respectively large cut-off frequencies (see [12] for discussion).

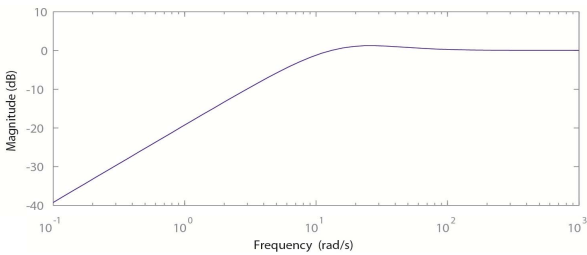


Fig. 2. Magnitude-frequency Bode characteristics for $H_a(s)$

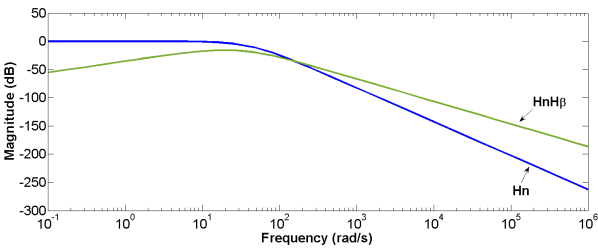


Fig. 3. Magnitude-frequency Bode characteristics for $H_n(s)$ and $H_n(s)H_\beta(s)$

C. Stability assessment

Assessing the stability of the NCS, the LMI condition (10) from Theorem 1 is solved using CVX Toolbox for Matlab ([16]), by formulating the problem as a convex optimization one. Solutions of the optimization problem were found (matrices P , Q , Z , N_1 , N_2 , S_1 and S_2) up to a maximum delay upper bound $\tau_{\max} = 0.35$ s, thus proving according to Theorem 1 that the system is asymptotically stable. Although the stability conditions are less conservative than most presented in the literature ([13]), the result may still be conservative. However it is usually good enough in practice - a network delay range $[0 \text{ s}, 0.35 \text{ s}]$ is consistent with most real life network transmission scenarios over the Internet ([17]).

D. Network characteristics

The network considered as reference for the NCS is a TCP/IP based wide area network (WAN). Under regular conditions a WAN is characterized by time varying delays ranging from a few tens to a few hundreds of milliseconds, the dominant component of the delay being the propagation delay ([17]).

When using an unreliable protocol for data transport (typically the case for a NCS transmission), the communication is also characterized by information loss, mostly because of packet drops due to network equipment saturation. Under such a scenario, there are mainly two approaches for maintaining the quality of control:

- i) adopt a sufficiently small sample period in respect with the maximum packet loss rate and the process dynamics ([1]);
- ii) impose a certain quality of service to the network by using an implementation aware co-design method (e.g. [18]).

The current case study considers the first approach, and as a result the network is characterized only by a time-varying delay defined by the network RTT.

E. Experimental results

The networked control structure from Fig. 1 was developed in a Matlab/Simulink environment and implemented on a dSPACE board connected to the electronic actuator and to the tachogenerator. The values of the time-varying delay (Fig. 4) were generated as uniformly distributed pseudo-random numbers. The delay varies between 0.05 s and 0.58 s with an average of 0.35 s (the scenario was chosen in order to show that the NCS can cope with extreme network transmission delay variations and values, which exceed the delay bound τ_{\max} provided by the sufficient stability conditions).

The experimental results show that the networked induced delays destabilize the system without the CDOB estimation (Fig. 5), leading to oscillations in the system's response (the control objective is no longer met). The extension of the NCS with the CDOB structure manages to eliminate the oscillations induced by the network and assures good tracking performances (Fig. 6), proving the capability of the CDOB to estimate and reject the delay disturbance (Fig. 7).

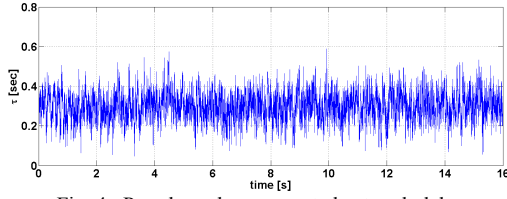


Fig. 4. Pseudorandom generated network delay

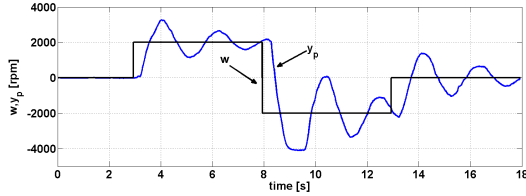


Fig. 5. Experimental results – networked control without CDOB

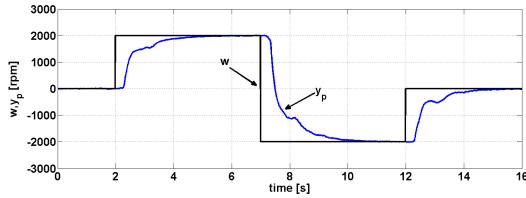


Fig. 6. Experimental results – networked control with CDOB

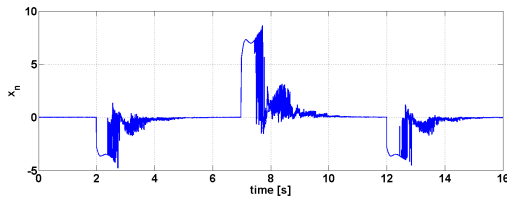


Fig. 7. Experimental results – estimated network delay disturbance

VI. CONCLUSION

The current study addresses the design, analysis, and validation of the generic observer-based delay compensation structure from Fig. 1. First, the design principles are briefly presented, while the case study further particularizes the design. Second, the stability of the observer-based delay compensation structure is addressed for the general case involving time-varying delays. Sufficient stability conditions are given, in terms of maximum delay bounds, using a Lyapunov functional. The stability conditions remain the same, regardless of the design approach used for different types of NCS that can be framed into this structure. Finally, the case study validates through experiments the observer-based NCS for the entire delay range obtained from the stability conditions.

ACKNOWLEDGMENT

This work was partially supported by the strategic grant POSDRU/159/1.5/S/137070 (2014) of the Ministry of National Education, Romania, co-financed by the European Social Fund – Investing in People, within the Sectoral Operational Programme Human Resources Development 2007-2013.

REFERENCES

- [1] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "A Survey of Recent Results in Networked Control Systems," *Proc. IEEE*, vol. 95, no. 1, pp. 138–162, Jan. 2007.
- [2] S. Zampieri, "Trends in Networked Control Systems," in *Proc. 17th IFAC World Congress*, Seoul, 2008, pp. 2886–2894.
- [3] P. F. Hokayem and C.T. Abdallah, "Inherent Issues in Networked Control Systems: A Survey," in *Proc. American Control Conference*, Boston, 2004, pp. 4897–4902.
- [4] H. Gao and T. Chen, "Network-based H^∞ Output Tracking Control," *IEEE Transactions on Automatic Control*, vol. 53, no. 3, pp. 655–667, 2008.
- [5] J. Nilsson, "Real-time control systems with delay," Ph.D. Thesis, Lund Institute of Technology, 1998.
- [6] W. Hu, G. Liu, D. Rees, "Event-Driven Network Predictive Control," *IEEE Trans. Indus. Electron.*, vol. 59, no. 3, pp. 905–913, 2011.
- [7] A. Onat, T. Naskali, E. Parlakay, O. Mutluer, "Control Over Imperfect Networks: Model-Based Predictive Networked Control Systems," *IEEE Trans. Indus. Electron.*, vol. 58, no. 3, pp. 905–913, Mar. 2011.
- [8] O. Stefan, A. Codrean, and T.-L. Dragomir, "A Network Control Structure with a Switched PD Delay Compensator and a Nonlinear Network Model," in *American Control Conference*, Washington, 2013, pp. 758–764.
- [9] C.-L. Lai and P.-L. Hsu, "Design the Remote Control System With the Time-Delay Estimator and the Adaptive Smith Predictor," *IEEE Trans. Indus. Informatics*, vol. 6, no. 1, pp. 73–80, Feb. 2010.
- [10] K. Natori, T. Tsuji, K. Ohnishi, A. Hase, and K. Jezernik, "Time-Delay Compensation by Communication Disturbance Observer for Bilateral Teleoperation Under Time-Varying Delay," *IEEE Trans. Indus. Electron.*, vol. 57, no. 3, pp. 1050–1062, Mar. 2010.
- [11] K. Natori, R. Oboe, and K. Ohnishi, "Stability Analysis and Practical Design Procedure of Time Delayed Control Systems with Communication Disturbance Observer," *IEEE Transactions on Industrial Informatics*, vol. 4, no. 3, pp. 185–197, 2008.
- [12] A. Codrean, O. Stefan, T.-L. Dragomir, "Design, analysis and validation of an observer-based delay compensation structure for a Network Control System," in *Mediterranean Conference on Control & Automation*, Barcelona, 2012, pp. 928–934.
- [13] Y. He, Q.-G. Wang, C. Lin, and M. Wu, "Delay-range-dependent stability for systems with time-varying delay," *Automatica*, vol. 43, pp. 371–376, 2007.
- [14] D. G. Luenberger, "An Introduction to Observers," *IEEE Trans. Autom. Control*, vol. 16, no. 6, pp. 596–602, Dec. 1971.
- [15] G. Franklin, J. Da Powell, A. Emami-Naeini, *Feedback Control of Dynamic Systems*, 7th Ed., Prentice Hall, 2014.
- [16] M. Grant and S. Boyd. (2011, April) CVX: Matlab Software for Disciplined Convex Programming. [Online]. <http://cvxr.com/cvx/>.
- [17] S.-H. Yang, *Internet-based Control Systems*, Springer, 2011.
- [18] O. Stefan, T.-L. Dragomir, "A Control-aware QoS Adaptation Co-design Method for Networked Control Systems," *Studies in Informatics and Control*, vol. 24, no. 1, pp. 33–42, 2015.

Performance Analysis for Operational Optimal Control for Complex Industrial Processes – the Square Impact Principle

Aiping Wang

Anhui University

Institute of Computer Sciences and Technology

Hefei, P R China

e-mail: apwang401@126.com

Hong Wang

The University of Manchester

Control center

Manchester, U K

e-mail: hong.wang@manchester.ac.uk

Ning Sheng

Northeastern University

The State Key Laboratory

of Synthetical Automation for Process Industries

Shenyang, P R China

e-mail: shengning2008@126.com

Xin Yin

The University of Manchester

Control center

Manchester, U K

e-mail: hong.wang@manchester.ac.uk

Abstract—The operation control for complex industrial processes consists of two layer – loop control layer and operational control layer. The former aims at achieving the required loop control for each production unit along production line whilst the later optimizes the set-points to the control loops so that certain performance indexes (such as product quality and energy cost) is optimized when the closed loop controlled variables follow their set-points well. This paper presents a novel method that can be used to analyze the performance deterioration of the optimized operational control, where the impact of tracking errors of loop control to the optimized performance is quantitatively formulated when the tracking errors are small. It has been shown that loop tracking errors would generally deteriorate the optimized performance in a quadratic way – leading to the establishment of the square impact principle (SIP). Moreover, it has been shown that the production infrastructure will also affect the deterioration of the optimized performance. Formulation on the analysis on the flat robustness (FR) and randomness of the optimized performance indexes will also be made and several issues on the future studies are listed.

Keywords—Complex industrial processes; operational optimal control, stochastic distribution control, optimal performance analysis, flat robustness (FR) and square impact principle (SIP)

I. INTRODUCTION

Production and operation infrastructure of most complex industrial processes (such as food processing, pharmaceuticals, paper making and steel making, etc.) were designed many years ago when there were plenty of energy and when the cost of raw materials were low and the quality of the raw materials (i.e., the grade of raw materials) were high. However, with the increased competitiveness of the globalized market and

especially the shortage of resources, energy prices and the cost of the quality raw material have significantly increased over the past decades. This has led to increased production costs and reduced profit for many industrial enterprises, which requires that manufacturers should produce quality and profitable products using less energy with raw materials of low grade. Ideally, the production system can at least produce saleable quality products using raw materials with large variations in composition. For this purpose, plant-wide operational control of complex industrial processes has now become a vitally important issue that is being addressed by both academia and industries [1-11].

On the other hand, most of the above complex industrial processes operate in a “kind of” 2D mode as shown in Fig. 1, where Q1, Q2, Q3 and Q4 are performance indexes that stand for the final product quality, yield, consumption and costs of industrial processes. In the figure, subscripts “min” and “max” represent the range of the indexes. In the horizontal direction (i.e., along production lines) there is normally a number of production units that are largely linked in a series of interconnections to perform the required production. Vertically there are a number of operation layers consisting of either one or several distributed control systems (DCS), planning and scheduling functioning (PS) units and even manufacturing execution systems (MES). Indeed, the multiple layers of the operation in the vertical direction constitute information flow, which manages mass flow and energy flow in the plant-wide operation. Therefore, the purpose of operation control is to manipulate information flows so that mass and energy flows are optimally interacted for an optimal plant-wide operation.

Most importantly, it is a well-known FACT that once the structure of the production operation is determined, the status of the plant-wide operation and ultimately the sustainability are

determined by a number of controlled variables (i.e., variables available to control and manage the process operation). These variables are controlled through control loops via DCSs. On the other hand, the performance such as product quality, production efficiency, costs and sustainability are related to these controlled variables. Therefore, it can be seen that these controlled variables play unique role in realizing plant-wide optimization and improved sustainability as they are the only group of decision variables that can be optimized. However, due to difficulties in establishing usable mathematical models, at present there is no unified optimal operational control approach that can be widely applied to various industrial processes [1, 11].

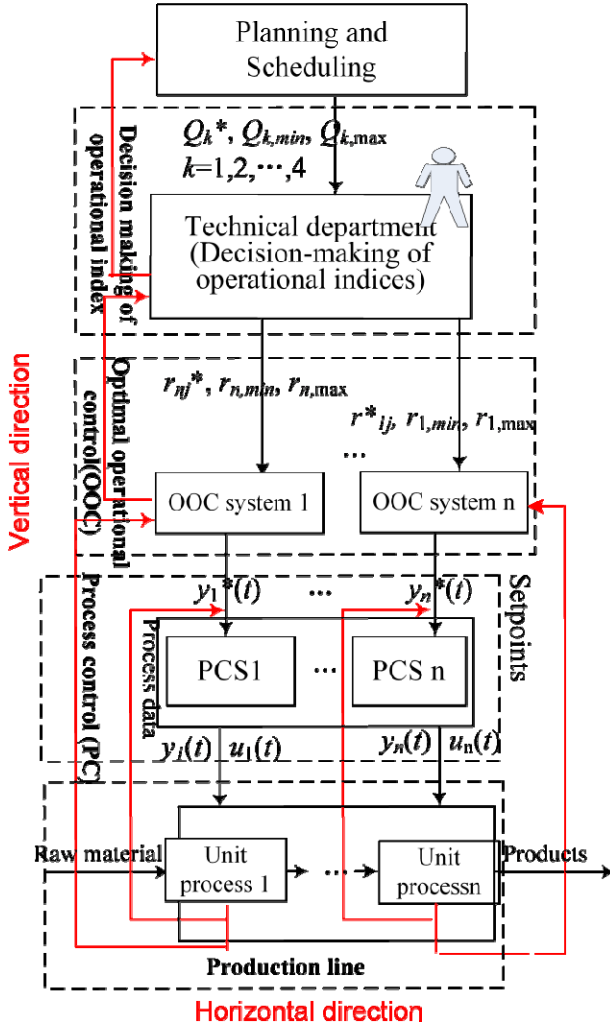


Fig. 1. The 2D operational mode for complex industrial processes

For industrial processes where trustable mathematical models can be established such as chemical processes etc, model based operational optimization and control can be readily applied. In this context, real-time optimization (RTO) uses traditional feedback strategy to form the required control methods for optimal operation. Such methods would select the set points for the controlled variables that correspond to economically optimal steady state for industrial processes. By adjusting relevant set-points of controlled variables and

ensuing that the controlled variables can follow their set-points, the whole process can be made to operate near the economically optimal steady state [1-11]. However, since there are various tracking errors in the loop control level, the actual optimality needs to be quantitatively analyzed. This forms the main purpose of this paper, where through simple mathematical formulation a novel principle, namely the square impact principle (SIP), will be described that shows the fact that the tracking errors at the loop control level affect the optimality of the processes in a square amplitude way.

II. DESCRIPTION OF OPERATIONAL OPTIMAL CONTROL

To simply the representation, let us consider the two-layer structure of plant-wide operational control system as shown in Fig. 2, where a performance function J (say energy) is required to be minimized. Such a performance would be a function of actual controlled variables to give

$$J = f(y)$$

$$y = [y_1, y_2, \dots, y_n]^T \in R^n \quad (1)$$

where $f(\cdot)$ is a nonlinear and assumed differentiable function of controlled variable vector group in y . Each controlled variable is controlled in a closed loop way and these closed loop control systems constitute loop layer as shown in Fig. 2.

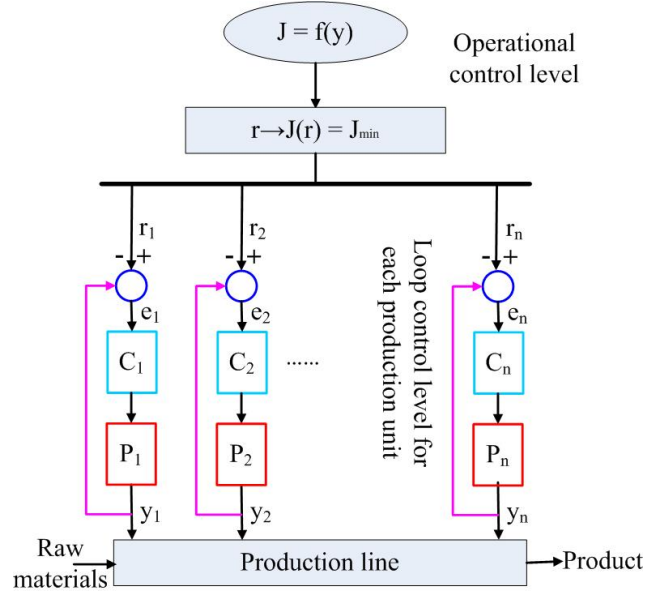


Fig. 2. A simple two-layered operational control scheme

To minimize J , a set of set points $r_i (i=1, \dots, n)$ can be obtained so that

$$\frac{\partial J}{\partial y} = \left[\frac{\partial f(y)}{\partial y} \right]_{y=r} = 0 \quad (2)$$

$$r = [r_1, r_2, \dots, r_n]^T \in R^n$$

where $r_i (i = 1, \dots, n)$ are the set points for controlled variables $y_i (i = 1, \dots, n)$ that minimize the performance index J , i.e.,

$$J_{\min} = f(r) \quad (3)$$

Since the above operational performance index is in fact functions of controlled variables of involved production equipment and units, rather than their set-points which can only be practically selected to optimize the performance indexes, it is necessary to analyze actual optimality of operational optimal control when there are tracking errors between the controlled variables and their set-points. In this regards, qualitative mathematical analysis needs to be established for example to assess the optimality of the optimized operational performance indexes and the functional flatness of the multi-objective index functions after they are optimized. In particular, it is necessary to analyze how the loop tracking errors and disturbances would affect the optimality of the performance index that has been optimized using the above simple method.

III. SQUARE IMPACT PRINCIPLE

To analyze the impact of the tracking errors and disturbances to the optimality, let us define the tracking error of each loop as

$$\begin{aligned} e_i(k) &= r_i(k) - y_i(k) \\ i &= 1, 2, \dots, n \end{aligned} \quad (4)$$

where $k (=1, 2, \dots)$ is the sample number. Then when the closed loop control design is completed and it is assumed that each closed loop exhibit the first order dynamics, the closed loop dynamics for the tracking errors can be expressed in the following discrete-time form,

$$\begin{aligned} e(k+1) &= \pi(e(k), d(k)) \\ e(k) &= [e_1(k), e_2(k), \dots, e_n(k)] \in R^n \\ d(k) &= [d_1(k), d_2(k), \dots, d_n(k)] \in R^n \end{aligned} \quad (5)$$

where $d_i(k) (i = 1, 2, \dots, n)$ are the disturbances of the i th control loop, $\pi(\cdot, \cdot)$ represents the assumed closed loop dynamics. Under the above assumption, the actual performance index J can be further formulated to read

$$J(y(k)) = J(y(k) - r(k) + r(k)) = J(r(k) - e(k)) \quad (6)$$

Assuming that the tracking errors are reasonably small, then the above index can be further formulated to give

$$J(y(k)) = J(r) - \left[\frac{\partial f}{\partial y} \right]_{y=r}^T e(k)$$

$$+ \frac{1}{2} e^T(k) \left[\frac{\partial^2 f}{\partial y^2} \right]_{y=r} e(k) \quad (7)$$

Since when $y = r$, J reaches its minimum, we have

$$\left[\frac{\partial f}{\partial y} \right]_{y=r} = 0, J(r) = J_{\min} \quad (8)$$

Therefore, the actual performance index J is given by

$$J(y(k)) = J_{\min} + \frac{1}{2} e^T(k) \left[\frac{\partial^2 f}{\partial y^2} \right]_{y=r} e(k) \quad (9)$$

$$\Delta J(y(k)) = J(y(k)) - J_{\min} \geq 0 \quad (10)$$

Then when there exist small loop tracking errors and disturbances, the actual performance has been deteriorated by

$$\Delta J(y(k)) = \frac{1}{2} e^T(k) \left[\frac{\partial^2 f}{\partial y^2} \right]_{y=r} e(k) \quad (11)$$

From the above equation it can be seen that for small tracking errors they would affect the performance optimality in a square magnitude way. This reveals a simple principle – the square impact principle (SIP).

From the above analysis it can be seen that there are two groups of factors that can affect or deteriorate the optimality. Apart from the loop tracking errors that affect the optimality of the performance index in a square magnitude way, the infra-structure of the production will also affect the performance. This is simply because the way that the performance index is related to controlled variables is determined by the production infra-structure. This can be

clearly shown by term $\left[\frac{\partial^2 f}{\partial y^2} \right]_{y=r}$ which reveals how y affects

f for a given production infra-structure. Indeed, term

$\left[\frac{\partial^2 f}{\partial y^2} \right]_{y=r}$ reflect industrialization in some way and term

$e^T(k) \left[\frac{\partial^2 f}{\partial y^2} \right]_{y=r} e(k)$ would show the coupling between

information technology and industrialization.

IV. ISSUES NEED TO BE ADDRESSED

Using square impact principle, a number of issues can be studied on the effectiveness of plant-wide operational optimization. Largely speaking, there are two issues that need to be addressed, namely how the shape of the performance

index J would affect the optimality and how the closed loop tracking errors can be minimized.

A. Shape of J – flat robustness concept

As the shape of the performance index J would affect the set-point optimization, and in particular it will affect the value of $\left[\frac{\partial^2 f}{\partial y^2}\right]_{y=r}$, it would be ideal if $\left\|\left[\frac{\partial^2 f}{\partial y^2}\right]_{y=r}\right\|$ is small. It means that we would hope that the shape of the performance index J is somehow ‘flat’ around the optimal point as shown in figure 3. This leads to a new concept on flat robustness which says how flat the performance index is around its optimal point r . Quantitatively, one can define \mathcal{E} -flat robustness as follows

$$\mathcal{E}\text{-flatrobustness} = L\{\|r - y\| \leq L \cup \|J(y) - J(r)\| \leq \mathcal{E}\} \quad (12)$$

Clearly too sharp performance index would not tolerate the possible fluctuation of tracking errors and thus the actual optimality would not be good. On the other hand, too flat performance index would make the search of optimal r difficult. Therefore in practice a good compromise is what needs to be achieved. Indeed, if the closed loop control has been well tuned so that $\|e(k)\| \leq \delta$, then from (11) it can be seen that the conservative estimate of the upper bound of the performance deterioration is given by

$$\|\Delta J(y(k))\| \leq \left\|\left[\frac{\partial^2 f}{\partial y^2}\right]_{y=r}\right\| \delta^2 \quad (13)$$

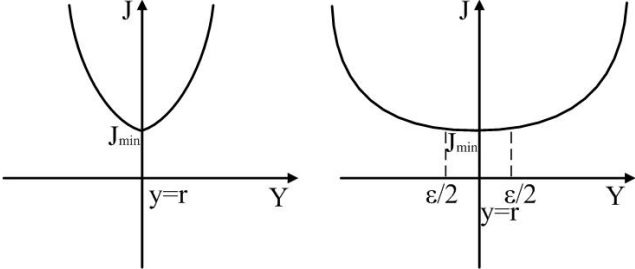


Fig. 3. The shape of the performance function

B. Performance Analysis Using Stochastic Distribution Theory

It is well-known that in practice all the disturbances at loop level are stochastic. Therefore the tracking error $e(k)$ is a random process vector. Using (11), it can be seen that $\Delta J(k)$ is also a random process. In this context, the information of the probability density function (pdf) for $\Delta J(k)$ is vitally important. Indeed, under the assumption that the pdf of $d(k)$ is known, one can formulate the pdf of the tracking error $e(k)$ and thus can use (11) and the structure of J to formulate the pdf of $\Delta J(k)$. To have a good optimality we would hope that the pdf of performance deterioration is narrow and also as close as possible to the

vertical axis as shown in figure 4. This belongs to stochastic distribution control area where the purpose of design to select a good set of loop control strategies so that the pdf of the performance deterioration and the tracking errors are made as narrow as possible [12].

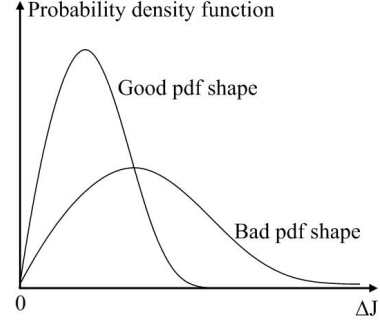


Fig. 4. The probability density function of performance deterioration

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

C. Re-selection of the set point

From (11) it can be seen that another effective way that can reduce the performance deterioration would be to re-select set-points to control loops. For example, by denoting the modifications to the optimal set-points as Δr so that the actual set-points applied to each control loop is given by

$$s(k) = r(k) + \Delta r(k) \quad (14)$$

Then the actual performance index is given by

$$J(y(k)) = J(y(k) - r(k) - \Delta r(k) + s(k)) \quad (15)$$

Assuming that the closed loop control has a small tracking error with respect to the new set-points and the adjustment to the old set-points is also small, then it can be further obtained that

$$\begin{aligned} J(y(k)) &= J(r(k)) + \left[\frac{\partial f}{\partial y}\right]_{y=r}^T \Delta r(k) \\ &\quad - \left[\frac{\partial f}{\partial y}\right]_{y=s}^T e(k) + \frac{1}{2} e^T(k) \left[\frac{\partial^2 f}{\partial y^2}\right]_{y=s} e(k) \\ &= J_m - \left[\frac{\partial f}{\partial y}\right]_{y=s}^T e(k) + \frac{1}{2} e^T(k) \left[\frac{\partial^2 f}{\partial y^2}\right]_{y=s} e(k) \end{aligned} \quad (16)$$

where it has been denoted that $e(k) = s(k) - y(k)$ as the new tracking error. It can be seen that to make the actual performance equal to its minimum, we need to solve the following equation for the set-points modifications

$$-\left[\frac{\partial f}{\partial y}\right]_{y=s}^T e(k) + \frac{1}{2} e^T(k) \left[\frac{\partial^2 f}{\partial y^2}\right]_{y=s} e(k) = 0 \quad (17)$$

This is clearly a difficult issue to be solved. Under the assumption that $e(k) \neq 0$, we can further solve the set-points modification Δr from the following simple equation.

$$2 \left[\frac{\partial f}{\partial y}\right]_{y=s} = \left[\frac{\partial^2 f}{\partial y^2}\right]_{y=s} (r(k) + \Delta r - y(k)) \quad (18)$$

This is a nonlinear equation as the partial derivatives are both calculated at $y = s(k) = r(k) + \Delta r(k)$. This means that

$\left[\frac{\partial f}{\partial y}\right]_{y=s}$ and $\left[\frac{\partial^2 f}{\partial y^2}\right]_{y=s}$ are nonlinear function matrices of

modified set-point $\Delta r(k)$. A possible iterative solution would lead to the recursive solution for $\Delta r(k)$ using equation (18).

V. A SIMULATED EXAMPE

To demonstrate the effectiveness of the proposed analysis method, let us consider the following simple system which contains two control loops as shown in Fig. 5.

In this system, the performance function is selected as

$$J(y) = (y_1 - 2)^2 + 2(y_2 - 3)^2 + 1$$

where it can be seen that the optimal set-points are given by

$$r_1 = 2; r_2 = 3 \text{ with } J_{\min} = 1$$

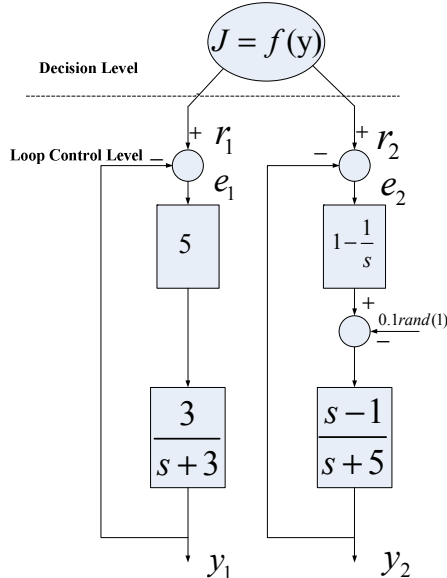


Fig. 5 A simple example

From Fig. 5, it can be seen that the disturbance is applied to the second loop and is selected as a white noise with 0.1 magnitude. From this figure it can be observed that the first loop will always exhibit a non-zero tracking error. This indicates that the actual performance function will be deteriorated in the way as we have analyzed in previous sections. Figures 6-8 show the responses of the tracking errors for each loop and also the response for the variation of the performance function $\Delta J(y(k)) = J(y(k)) - 1$. In the simulation, the sampling interval is selected as 0.01 second.

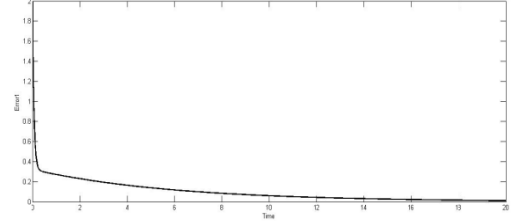


Fig. 6 The tracking error for control loop 1

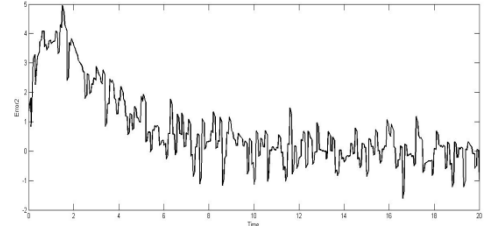


Fig. 7 The tracking error for control loop 2

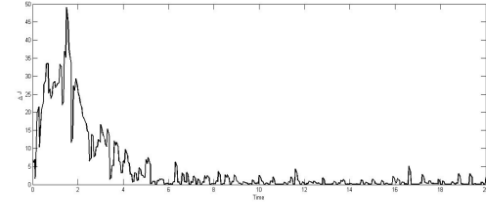


Fig. 8 The deterioration of the performance

VI. CONCLUSIONS

In response to the fast development of process industries and industrial information technology world-wide, new requirements have been imposed on process control. Automation should not only be realized for majority operation of industrial processes so that their control systems can ensure a desired tracking with respect to their set-points for the output of the plant to be controlled, but also the operational optimal control for the whole plant wide should be realized. This paper describes the ways to analyze the optimality of the optimized performance index, where an interesting discovery – square impact principle – has been formulated for modellable complex industrial processes. Future directions have also been discussed in terms of flat-robustness, use of stochastic distribution control and set-point re-selection. It is important to note that the method here is only valid for processes that can be reasonably

modelled and that the loop tracking errors are small. For processes that cannot be modelled, data driven approaches need to be used to develop modelling methods first for the optimization, then the proposed methods can be readily applied to unmodellable plant.

ACKNOWLEDGMENT

The work reported here is supported by NSFC grants 61374128, 61333007 and 61290323, these are gratefully acknowledged.

REFERENCES

- [1] Engell, S., "Feedback control for optimal process operation," *Journal of Process Control*, vol. 17, pp.203-219, 2007.
- [2] Mark L. Darby, "Michael Nikolaou, James Jones, Doug Nicholson. RTO: An overview and assessment of current practice," *Journal of Process Control*, vol. 21, pp. 874-884, 2011.
- [3] Johannes Jaschke, Sigurd Skogestad, "NCO tracking and self-optimizing control in the context of real-time optimization," *Journal of Process Control*, vol.21, pp. 1047-1416, 2011.
- [4] Mehmet Mercangoz, Francis J. Doyle III. "Real-time optimization of the pulp mill benchmark problem," *Computers and Chemical Engineering*, vol.32, pp.789-804, 2008.
- [5] J. Hasikos, H. Sarimveis, P.L. Zervas, N.C. Markatos, "Operational optimization and real-time control of fuel-cell systems," *Journal of Power Sources*, vol.193, pp.258-268, 2009.
- [6] I. P. Tatjewski, "Advanced control and on-line process optimization in multilayer structures," *Annual Reviews in Control*, vol.32, pp.71-85, 2008.
- [7] V. Adetola, M. Guay, "Integration of real-time optimization and model predictive control," *Journal of Process Control*, vol. 20, pp.125-133, 2010.
- [8] Skogestad, S., "Plantwide control: the search for the self-optimizing control structure," *Journal of Process Control*, vol.10, pp.487-507, 2000.
- [9] Nath, R. and Z. Alzein, "On-line dynamic optimization of olefins plants," *Computers & Chemical Engineering*, vol.24, pp.533-538, 2000.
- [10] P. Zhou, T. Y. Chai, and H. Wang, "Intelligent optimal-setting control for grinding circuits of mineral processing process," *IEEE Transactions on Automation Science and Engineering*, vol.6(4), pp.730 -743, 2009.
- [11] J Ding, T Chai, H Wang, "Knowledge-based Plant-Wide Dynamic Operation of Mineral Processing Under Uncertainty," *IEEE Trans on Industry Informatics*, vol.8(4), .pp.849-859, 2012.
- [12] H Wang, "Bounded Dynamic Stochastic Distributions – Modelling and Control," Springer-Verlag, 1999.

Parametric Covariance Assignment using Reduced-order Closed-form Covariance Model

Qichun Zhang

School of Electrical &
Electronic Engineering
The University of Manchester
Manchester, U.K. M13 9PL
qichun.zhang@manchester.ac.uk

Zhuo Wang

Fok Ying Tung Graduate School
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
zwang8381@foxmail.com

Hong Wang

School of Electrical &
Electronic Engineering
The University of Manchester
Manchester, U.K. M13 9PL
hong.wang@manchester.ac.uk

Abstract—In this paper, two novel closed-form covariance models using covariance matrix eigenvalues are presented for continue-time linear stochastic systems and discrete-time linear stochastic systems, respectively, which are subjected to Gaussian noises. Based on these model, the state and output covariance assignment algorithms have been developed with parametric state and output feedback. Due to the simple structure of this model, the low-order controller can be obtained following the proposed algorithms, which reduced computational complexity and the extended free parameters of parametric feedback can supply flexibility to optimization.

Keywords—Covariance assignment, Closed-form covariance model, Parametric eigenstructure.

I. INTRODUCTION

Covariance analysis permeates almost all of systems theory [1]. Based on the development of the stochastic systems control, the covariance control has become one of the most important research fields for multi-outputs stochastic control systems.

During the past two decades, the covariance control theory [1] has made great progresses after proposed by A.Hotz and R.E.Skelton in 1987. The main result of this theory is based on lyapunov equation, and several conditions and controllers [2] [3] [4] [5] have been proposed to control the covariance of the stochastic systems using the determined control signals. But all of the controllers mentioned have no closed-form. Since the closed-form model of state covariance [6] presented in 2007, S.Baromand and H.Khaloozadeh designed different controllers and models [7] [8] to solve state covariance assignment(SCA)problem using the random control signal. All these literatures focus on the states of the systems rather than outputs of the systems and also the closed-form covariance model increases the dimension of the system variables.

Using the closed-form model for covariance assignment problem, there is two key problem remained so far. Firstly, the order of the controller must be more than the order of original systems, in this way, the high-order controller would be obtained and the computational complexity would increase with it. Basically, the order of the controller is limited by applying ,for example, the controller of the Hubble Space Telescope can not be high-order due to the limited space and computational complexity [9]. Secondly, the states cannot be measured in practice, therefore, how to design a controller

using outputs of the systems directly is more significant for application.

Due to the problems mentioned above, novel closed-form state and output covariance models for both continue-time and discrete-time linear stochastic systems have been proposed in this paper. With the models proposed here, the order of the controller can be reduced to original systems order. The computational complexity can be reduced absolutely. The application is much easier using the parametric state and output feedback approaches [10] [11] with the presented models. Meanwhile, the parametric state and output feedback approaches can supply flexibility to optimize covariance controllers as extensions.

This paper is organized as follows: In Section II, the reduced-order closed-form covariance models are presented. Section III presents two parametric feedback algorithms by system state and system output. Finally, the numerical examples and conclusions are drawn in Section IV and Section V, respectively.

II. REDUCED-ORDER CLOSED-FORM COVARIANCE MODEL

In this section, two novel closed-form covariance model are presented using eigen-decomposition for continue-time linear stochastic systems and discrete-time linear stochastic systems, respectively.

A. Discrete-time reduced-order covariance model

Consider the discrete-time linear stochastic systems subjected to Gaussian noises which are represented as

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) + Dw(k) \\ y(k) &= Cx(k) \end{aligned} \quad (1)$$

where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ denote state vector and output vector of the systems. $u \in \mathbb{R}^s$ is the control input vector and $w \in \mathbb{R}^p$ is the Gaussian noise vector. A, B, D and C are real constant matrix with appropriate dimensions.

Assume that the Gaussian noise vector satisfies:

$$\begin{aligned} \underline{H1}: E\{w(k)\} &= 0, E\{x(0)w^T(k)\} = 0 \\ E\{w(i)w^T(j)\} &= Q\delta(i-j) \end{aligned} \quad (2)$$

where $\delta(\cdot)$ is Dirac delta function.

Similar to the assumption of the noise, the control signal can be restricted as:

$$\begin{aligned} E\{u(k)\} &= 0 \\ E\{x(0)u^T(k)\} &= 0, E\{w(i)u^T(j)\} = 0 \\ E\{u(i)u^T(j)\} &= U(i)\delta(i-j) \end{aligned} \quad (3)$$

where $U(i) \triangleq E\{u(i)u^T(i)\}$

Once the mean value of the control signal vector is restricted to zero, we have

$$\lim_{k \rightarrow \infty} E\{x(k)\} = 0, \lim_{k \rightarrow \infty} E\{y(k)\} = 0 \quad (4)$$

Based on the definition of the covariance matrix, the state and output covariance matrix of the given system are given by

$$\begin{aligned} P_x(k) &\triangleq E\{x(k)x^T(k)\} \\ P_y(k) &\triangleq E\{y(k)y^T(k)\} \end{aligned} \quad (5)$$

The covariance matrices can be restated using the system model and the associated dynamic Lyapunov equation is given by

$$\begin{aligned} P_x(k+1) &= AP_x(k)A^T + BU(k)B^T + DQD^T \\ P_y(k) &= CP_x(k)C^T \end{aligned} \quad (6)$$

In [7] [8], the covariance matrices are transformed to vectors using vectorization, which increases the order of the transformed model. To overcome this shortcoming, the eigen-decomposition is used to reduce the order of the model.

The covariance matrices are rewritten as

$$\begin{aligned} P_x(k) &= V_x \Lambda_x(k) V_x^T \\ P_y(k) &= V_y \Lambda_y(k) V_y^T \\ U(k) &= V_u \Lambda_u(k) V_u^T \\ Q &= V_q \Lambda_q V_q^T \end{aligned} \quad (7)$$

where $\Lambda_x, \Lambda_y, \Lambda_u$ and Λ_q are real diagonal matrices. V_x, V_y, V_u and V_q associated orthogonal matrices. All of the matrices are with the same dimensions as the associated vectors.

The new formula of the dynamic Lyapunov equation can be obtained.

$$\begin{aligned} \Lambda_x(k+1) &= A_\Lambda \Lambda_x(k) A_\Lambda^T + B_\Lambda \Lambda_u(k) B_\Lambda^T + D_\Lambda \Lambda_q D_\Lambda^T \\ \Lambda_y(k) &= C_\Lambda \Lambda_x(k) C_\Lambda^T \end{aligned} \quad (8)$$

where $A_\Lambda = V_x^T A V_x, B_\Lambda = V_x^T B V_u, D_\Lambda = V_x^T D V_q$ and $C_\Lambda = V_y^T C V_x$, respectively.

Furthermore, Eq. (8) can be transformed as standard state space model:

$$\begin{aligned} \lambda_x(k+1) &= A_{\text{cov}} \lambda_x(k) + B_{\text{cov}} \lambda_u(k) + D_{\text{cov}} \lambda_q \\ \lambda_y(k) &= C_{\text{cov}} \lambda_x(k) \end{aligned} \quad (9)$$

where $\lambda_x, \lambda_y, \lambda_u$ and λ_q denote eigenvalue vectors. The coefficient matrices $A_{\text{cov}}, B_{\text{cov}}, D_{\text{cov}}, C_{\text{cov}}$ can be calculated using $A_\Lambda, B_\Lambda, D_\Lambda, C_\Lambda$ by nonlinear mapping.

Remark 1: It has been shown that the Eq.(9) satisfies the Dynamic Lyapunov equation when the time trends to be infinite and the order of the model equals to the original system model.

B. Continue-time reduced-order covariance model

Similar to the case of discrete-time system, the model for continue-time systems is presented briefly.

Consider the continue-time linear stochastic system subjected to Gaussian noise.

$$\begin{aligned} dx &= (Ax + Bu)dt + Dd\beta_t \\ dy &= Cxd t \end{aligned} \quad (10)$$

where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ denote state vector and output vector of the systems. $u \in \mathbb{R}^s$ is the control input vector and β_t is the p-dimensional Brownian Motion. A, B, D and C are real constant matrix with appropriate dimensions.

Notice that a Wiener process which is used to represent the integral of a Gaussian white noise process can describe a Brownian motion in random process and it yields

$$d\beta_t = wdt \quad (11)$$

where w is a standard Gaussian white noise.

Therefore, the system can be rewritten as

$$\begin{aligned} dx &= (Ax + Bu + Dw)dt \\ dy &= Cxd t \end{aligned} \quad (12)$$

Compared with the assumption for the discrete-time system, the following assumption is given.

$$\underline{\text{H2:}} \quad E\{w(t)\} = 0, E\{x(0)w^T(t)\} = 0$$

$$E\{w(t)w^T(\tau)\} = Q\delta(t-\tau) \quad (13)$$

where $\delta(\cdot)$ is Dirac delta function.

Also, all the discrete-time variables defined above can be redefined as continue-time variables by replacing k to t , and the covariance matrix for continue-time system satisfies the following dynamic Lyapunov function:

$$\begin{aligned} \dot{P}_x(t) &= AP_x(t) + P_x(t)A^T + BU(t)B^T + DQD^T \\ P_y(t) &= CP_x(t)C^T \end{aligned} \quad (14)$$

Using the similar approach presented above, the eigenvalue matrix equation is given by

$$\begin{aligned} \dot{\Lambda}_x(t) &= A_\Lambda \Lambda_x(t) + \Lambda_x(t) A_\Lambda^T + B_\Lambda \Lambda_u(t) B_\Lambda^T + D_\Lambda \Lambda_q D_\Lambda^T \\ \Lambda_y(t) &= C_\Lambda \Lambda_x(t) C_\Lambda^T \end{aligned} \quad (15)$$

where $A_\Lambda = V_x^T A V_x, B_\Lambda = V_x^T B V_u, D_\Lambda = V_x^T D V_q$ and $C_\Lambda = V_y^T C V_x$, respectively.

Finally, Eq.(15) can be further expressed as standard state space model:

$$\begin{aligned} \dot{\lambda}_x(t) &= A_{\text{cov}} \lambda_x(t) + B_{\text{cov}} \lambda_u(t) + D_{\text{cov}} \lambda_q \\ \lambda_y(t) &= C_{\text{cov}} \lambda_x(t) \end{aligned} \quad (16)$$

Remark 2: The new models presented in this section can be considered as the coordinate transformation of the original dynamic Lyapunov equations associated to covariance matrices.

III. PARAMETRIC COVARIANCE ASSIGNMENT ALGORITHMS

Based on the reduced-order closed-form covariance models, the parametric state and output feedback control algorithms are developed to assign the covariance values. To simplify the contents of the paper, the control algorithms are proposed using discrete-time model, on the other hand, the similar algorithms using continue-time model are omitted.

H3 : In this section, assume that the reduced-order closed-form covariance models are controllable

A. state covariance controller design

For the state covariance assignment, the reference covariance matrix can be rewritten using eigen-decomposition as

$$R = V_r \Lambda_r V_r^T \quad (17)$$

Since the diagonal matrix Λ_r can be arranged as vector λ_r , the covariance assignment problem transfer to state tracking problem using the presented reduced-order covariance model if we set $V_x = V_r$.

To track the desired state covariance vector, the integrator should be considered in the control scheme. The error vector $e_x(k+1) = e_x(k) + \lambda_r - \lambda_x$ is treated as the extended state and substitute the error into the closed-loop system.

Then, the closed-loop system in the state space form can be obtained as

$$\begin{bmatrix} e_x(k+1) \\ \lambda_x(k+1) \end{bmatrix} = \bar{A} \begin{bmatrix} e_x(k) \\ \lambda_x(k) \end{bmatrix} + \bar{B} \lambda_u(k) + \begin{bmatrix} 0 \\ D_{cov} \end{bmatrix} \lambda_q + \begin{bmatrix} \lambda_r \\ 0 \end{bmatrix} \quad (18)$$

where

$$\bar{A} = \begin{bmatrix} I & -I \\ 0 & A_{cov} \end{bmatrix}, \bar{B} = \begin{bmatrix} 0 \\ B_{cov} \end{bmatrix}$$

For this control system with error vector, a full-state feedback can be designed using parametric state feedback approach [10]

$$\lambda_u(k) = K \begin{bmatrix} e_x(k) \\ \lambda_x(k) \end{bmatrix} \quad (19)$$

and the feedback gain can be obtained by

$$K = [W_1 f_1, \dots, W_n f_n] \times [(\lambda_1^* I - \bar{A}_1)^{-1} \bar{B}_1 f_1, \dots, (\lambda_n^* I - \bar{A}_n)^{-1} \bar{B}_n f_n]^{-1} \quad (20)$$

where the modified parameter vectors are denoted by f_1, \dots, f_n as free parameters.

In the case of a common open-loop and closed-loop eigenvalue, then, other parameters in Eq. (20) can be determined as below.

$$\begin{cases} \bar{A}_i = \bar{A} + v_j^0 w_j^{0T} \\ W_i = I - \frac{e_k w_j^{0T} \bar{B}}{w_j^{0T} b_k} \\ \bar{B}_i = \bar{B} W_i + v_j^0 e_k^T \end{cases} \quad (21)$$

where v_j^0 and w_j^0 ($j = 1, \dots, n$) denote the open-loop eigenvectors and eigenrows of the model (18). \bar{A} and \bar{B} are given by model (18). b_k is the k -th column of the matrix \bar{B} . e_k is a unit vector where the k -th element is 1. In the other case, there is no common eigenvalue, $w_j^{0T} b_k = 0$, so that the parameters of Eq. (20) can be selected by

$$\begin{cases} \bar{A}_i = \bar{A} \\ W_i = I \\ \bar{B}_i = \bar{B} \end{cases} \quad (22)$$

To backwards the transformation, Λ_u can be obtained by λ_u and the actual control signal for original model can be given by

$$u(k) = [V_u \Lambda_u(k) V_u^T]^{\frac{1}{2}} \xi(k) \quad (23)$$

where $\xi(k)$ denotes the standard Gaussian white noise.

Remark 3: The actual control law is nonlinear though the original system model is linear.

Controller designing procedure can be summarized as Algorithm I:

- Step 1, choose the reference covariance matrix and calculating the eigenvalues and eigenvectors of the desired covariance matrix.
- Step 2, transform the stochastic systems from the original model to reduced-order closed-form covariance model.
- Step 3, transform the closed-form model to reference tracking model by adding the error vector.
- Step 4, choose poles and free parameters for closed-loop covariance control model and design the state covariance controller via parametric feedback approaches.
- Step 5, calculate the control signal for original systems using control law of covariance model.
- Step 6, substitute the control signal into the original systems.

B. output covariance controller design

The output feedback is widely used when the system state cannot be measured. Similar to the state covariance controller design, the error vector $e_y(k+1) = e_y(k) + \lambda_r - \lambda_y$ should be introduced to this control systems, and the new state space model with output equation can be described by

$$\begin{bmatrix} e_y(k+1) \\ \lambda_x(k+1) \end{bmatrix} = F \begin{bmatrix} e_y(k) \\ \lambda_x(k) \end{bmatrix} + \bar{B} \lambda_u(k) + \begin{bmatrix} 0 \\ D_{cov} \end{bmatrix} \lambda_q + \begin{bmatrix} \lambda_r \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} e_y(k) \\ \lambda_y(k) \end{bmatrix} = \bar{C} \begin{bmatrix} e_y(k) \\ \lambda_x(k) \end{bmatrix} \quad (24)$$

where

$$F = \begin{bmatrix} I & -C_{cov} \\ 0 & A_{cov} \end{bmatrix}, \bar{C} = \begin{bmatrix} I & 0 \\ 0 & C_{cov} \end{bmatrix}$$

For this extended system, assume the following conditions hold.

H4: (Kimura's condition [11]) $m + s \geq n + 1$

The output feedback control law can be designed by parametric output feedback approach [11]

$$\lambda_u(k) = G \begin{bmatrix} e_y(k) \\ \lambda_y(k) \end{bmatrix} \quad (25)$$

and the feedback gain G can be obtained by

$$G = K_0 + K_2 U_1' + U_2 K_3 U_1' \quad (26)$$

where the K_0, K_2, U_1', U_2 can be calculated by kernel space and K_3 can be calculated by exterior algebra. (see [11] for calculation)

Once the control input λ_u is obtained, the actual control law also can be calculated by Eq.(23)

The procedure of the controller design can be summarized as Algorithm II:

- Step 1, choose the reference covariance matrix and calculating the eigenvalues and eigenvectors of the desired covariance matrix.
- Step 2, transform the stochastic systems from the original model to reduced-order closed-form covariance model.
- Step 3, transform the closed-form model to reference tracking model by adding the error vector.
- Step 4, choose poles and free parameters for closed-loop covariance control model and design the output covariance controller via parametric feedback approaches.
- Step 5, calculate the control signal for original systems using control law of covariance model.
- Step 6, substitute the control signal into the original systems.

IV. NUMERICAL EXAMPLE

To verify the new model and the controller proposed in this paper, one numerical example is presented in this section.

The original model can be showed below:

$$A = \begin{bmatrix} -0.5 & -0.3 \\ 0.1 & -0.2 \end{bmatrix}, B = \begin{bmatrix} 2 & 0.3 \\ 0.1 & 4 \end{bmatrix} \\ C = \begin{bmatrix} 0.7 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}, D = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$$

The covariance matrix of disturbance white noise is

$$\begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}$$

To assign the covariance matrix, we choose the reference covariance matrix as

$$\begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}$$

A. Case for state covariance

From the reference covariance matrix, the reference signal for covariance matrix eigenvalue model can be obtained as

$$\lambda_r = \begin{bmatrix} 0.1382 \\ 0.3618 \end{bmatrix} \\ V_r = V_x = V_u = V_q = \begin{bmatrix} -0.8507 & 0.5257 \\ 0.5257 & 0.8507 \end{bmatrix}$$

Simply, the reduced-order covariance model can be described.

$$A_{\text{cov}} = \begin{bmatrix} 0.1073 & 0.1436 \\ 0.0004 & 0.1387 \end{bmatrix} \\ B_{\text{cov}} = \begin{bmatrix} 5.6354 & 0.4970 \\ 0.8190 & 13.1486 \end{bmatrix} \\ D_{\text{cov}} = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$$

Based on this model and parametric state feedback approach, the state covariance can track the given reference matrix by transformation and choosing different free parameter matrix.

$$\lambda^* = \{ 0.1 \quad 0.5 \quad 0.3 \quad 0.2 \} \\ F_1 = \begin{bmatrix} -3 & 1 & -2 & -1 \\ 1 & 3 & 1 & 2 \end{bmatrix} \\ F_2 = \begin{bmatrix} -3 & 1 & -2 & -1 \\ 1 & -1 & 1 & 2 \end{bmatrix}$$

The state covariance feedback gain can be obtained respectively as follows:

$$K1_x = \begin{bmatrix} -0.007 & -0.0054 & -0.0671 & 0.0083 \\ -0.0013 & -0.0055 & -0.0125 & -0.0328 \end{bmatrix} \\ K2_x = \begin{bmatrix} -0.0044 & 0.0013 & -0.0427 & 0.0313 \\ -0.0035 & -0.0112 & -0.0331 & -0.0522 \end{bmatrix}$$

The results have been showed below. In Fig. (1) and Fig. (2), the state covariance curves have been given using different feedback gain K_1 and K_2 . Both of the control laws can assign the covariance to reference covariance matrix, however, different free parameters of the controller lead to different performance. Fig. (3) and In Fig. (4) show that the curves of the eigenvalues of the state covariance matrix, i.e. the state vector of the reduced-order closed-form covariance model. From the curves, it is obvious that K_1 is better than K_2 for this example.

B. Case for output covariance

As the approach mention above, the eigenvectors can also choose in the same way by

$$V_r = V_y = V_x = V_u = V_q = \begin{bmatrix} -0.8507 & 0.5257 \\ 0.5257 & 0.8507 \end{bmatrix}$$

and then we have

$$C_{\text{cov}} = \begin{bmatrix} 0.3522 & 0.0008 \\ 0.0052 & 0.8218 \end{bmatrix}$$

Based on this model and parametric output feedback approach, the output covariance can track the given reference

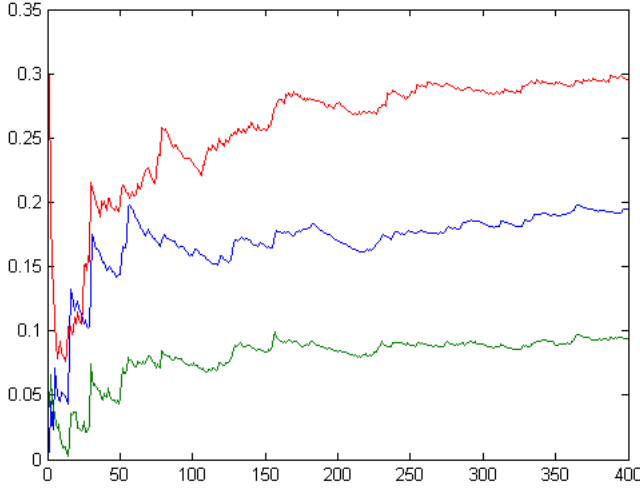


Fig. 1. State Covariance using K_1

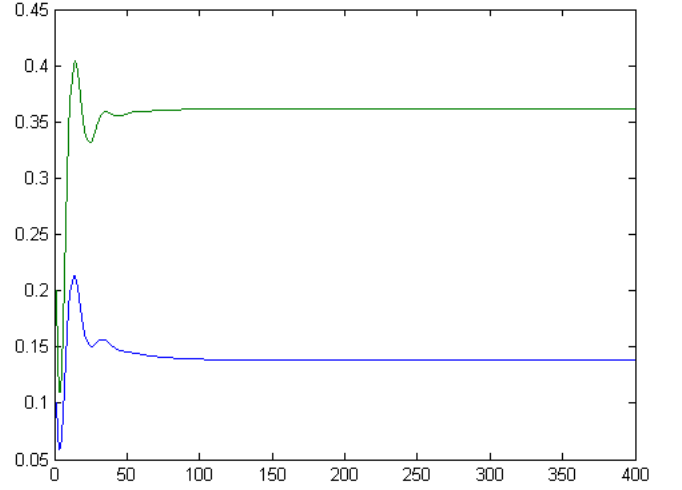


Fig. 3. Eigenvalues of the State Covariance Matrix using K_1

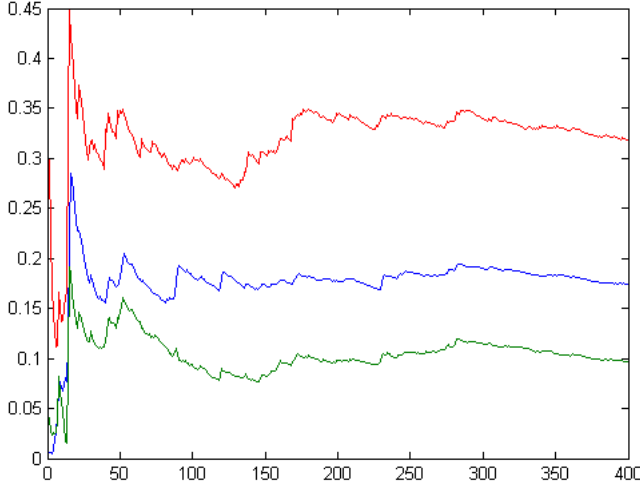


Fig. 2. State Covariance using K_2

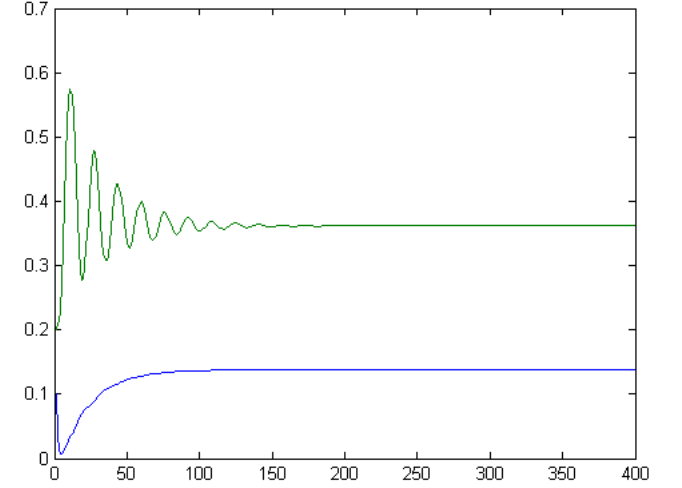


Fig. 4. Eigenvalues of the State Covariance Matrix using K_2

matrix by transformation and choosing different free parameter matrix.

$$q = \begin{bmatrix} 1 & -1 & 2 \\ 1 & 1 & 1 \end{bmatrix}, z = 1, K_3 = 5$$

Finally, the feedback gain can be calculated as

$$K_y = \begin{bmatrix} -1.9180 & 1.3353 & -2.8218 & 3.3578 \\ -0.2034 & 0.1415 & -0.3448 & 0.3850 \end{bmatrix}$$

To avoid repeat the results which is similar to the state covariance controller design, the curves for output covariance controller have been omitted.

Remark 4: From the results of the simulation, the different free parameters associated with the feedback gain bring the different performance to the covariance model. It means that optimal free parameters can be obtained for a given

performance criterion, such as minimum sensitivity, minimum control energy, etc. The control algorithms presented in this paper can be extended simply which is helpful to apply in practice.

V. CONCLUSION

In this paper, novel models for covariance assignment have been proposed named reduced-order closed-form covariance model using eigen-decomposition for discrete-time and continue-time linear stochastic systems subjected to Gaussian white noise. Based on these models, the controller design is simplified and the low-order controller can be obtained. Two nonlinear control laws have been formulated following two presented algorithms by parametric state feedback approach and parametric output feedback approach, where the free parameters can be further used to optimize for other control criterion. Generally, the covariance assignment problem is

solved by parametric approach using novel covariance control model.

REFERENCES

- [1] A. Hotz and R. E. Skelton, "Covariance control theory," *International Journal of Control*, vol. 46, no. 1, pp. 13–32, 1987.
- [2] K. Yasuda, R. Skelton, and K. Grigoriadis, "Covariance controllers: A new parametrization of the class of all stabilizing controllers," *Automatica*, vol. 29, no. 3, pp. 785–788, 1993.
- [3] K. Grigoriadis and R. Skelton, "Minimum-energy covariance controllers," *Automatica*, vol. 33, no. 4, pp. 569–578, 1997.
- [4] E. G. Collins Jr. and R. E. Skelton, "Theory of state covariance assignment for discrete systems," *IEEE Transactions on Automatic Control*, vol. AC-32, no. 1, pp. 35–41, 1987.
- [5] E. Yaz and R. Skelton, "Continuous and discrete state estimation with error covariance assignment," vol. 3, 1991, pp. 3091–3092.
- [6] H. Khaloozadeh and S. Baromand, "State covariance assignment problem," *IET Control Theory and Applications*, vol. 4, no. 3, pp. 391–402, 2010.
- [7] S. Baromand and H. Khaloozadeh, "On the closed-form model for state covariance assignment problem," *IET Control Theory and Applications*, vol. 4, no. 9, pp. 1678–1686, 2010.
- [8] S. Baromand and B. Labibi, "Covariance control for stochastic uncertain multivariable systems via sliding mode control strategy," *IET Control Theory and Applications*, vol. 6, no. 3, pp. 349–356, 2012.
- [9] G. Zhu, K. M. Grigoriadis, and R. E. Skelton, "Covariance control design for hubble space telescope," *Journal of Guidance, Control, and Dynamics*, vol. 18, no. 2, pp. 230–236, 1995.
- [10] G. Roppenecker, "On parametric state feedback design," *International Journal of Control*, vol. 43, no. 3, pp. 793–804, 1986.
- [11] U. Konigorski, "Pole placement by parametric output feedback," *Systems and Control Letters*, vol. 61, no. 2, pp. 292–297, 2012.

Stabilizing slug flow at large valve opening using active feedback control

Adegboyega Bolu Ehinmowo*, Yi Cao

Oil & Gas Engineering Centre
Cranfield University
Bedfordshire, UK

*Corresponding Author email: a.ehinmowo@cranfield.ac.uk

Abstract— The threat of slugging to production facilities has been known since the 70's. This undesirable flow phenomenon continues to attract the attention of researchers and operators alike. The most common method for slug mitigation is by choking the valve at the exit of the riser which unfortunately could negatively affect production. The focus therefore is to satisfy the need for system stability and to maximize production simultaneously. Active feedback control is a promising way to achieve this. However, due to the complexity of multiphase flow systems, it is a challenge to develop a robust slug control system to achieve the desired performance using existing design tools. In this paper, a new general method for multiphase flow system stability analysis was proposed. Active feedback control was observed to optimize slug attenuation compared with manual choking.

Keywords- Choking, active feedback, slugging, stability analysis, multiphase flow

I. INTRODUCTION

Oil and gas activities in many oil producing nations of the world have shifted to deep offshore. Many of the fields are too small to accommodate a standalone offshore processing facilities. Also many of the existing fields are either in plateau production phase thus it becomes very popular tying production pipelines from satellite fields to an existing pipeline to use common offshore processing facility. The transportation of the produced crude is usually done in multiphase pipelines. In so doing, one of the challenges encountered is slugging. Slugging in oil and gas pipelines is a cardinal problem for all oil and gas producers. It is characterized by large pressure and production fluctuations.

One of the ways of suppressing or eliminating fluctuation due to slugging is by choking. In practice, oil and gas industry have used this method for many years to eliminate severe slugging by manipulating the valve opening at the exit of the riser which unfortunately could negatively affect production [1; 2]. The use of controller however has been reported to be able to help alleviate this problem by stabilizing the system at larger valve opening [3]. Significant efforts have been concentrated on modelling and understanding the slug attenuation mechanism for choking [2; 4] and active slug control [5-9]. These models can be used to gain insight into the mechanism and control design. Nevertheless, these models may not accurately represent real systems due to the complexity of multiphase flow [10-12]. This leaves the robustness of slug control systems designed based on these models questionable. There is therefore the need for

a simple but yet robust methodology that can be used for system analysis and controller design. The aim of this study is to develop an approach to slug flow stability using feedback control. To achieve this aim, we propose a new method that can be used for slug flow stability analysis and design a controller for stabilizing the unstable slug flow. A theoretical analysis was attempted to show the slugging mitigation potential of active feedback control at larger valve opening compared with traditional manual choking. This paper is organized as follows: in section II, the new approach to slug flow stability analysis is presented; in section III numerical case study was performed and the work is concluded in section IV.

II. STABILITY ANALYSIS OF SLUG ATTENUATION WITH FEEDBACK CONTROL

In the co-current flow of gas-liquid mixtures through pipeline-riser system, slugging is frequently encountered for wide range of pipe inclinations and flow rates. Severe slugging is a flow regime which can be described as a four stage transient cyclic phenomenon shown in Fig.1. At low flow rate, the liquid accumulates at the riserbase blocking the gas flow while the riser column get filled with liquid (slug formation stage), the gas is compressed in the upstream pipeline causing a pressure buildup which later becomes sufficient to overcome the hydrostatic head in the riser thereby forcing the liquid slug out of the riser (slug production stage). This is followed by a gas surge (gas blow out) and the remaining liquid in the riser fall back to the riserbase (liquid fall back stage) which again starts the cycle [2]. Severe slugging usually manifests in significant fluctuation of flow and pressure. This instability is as a result of the upward multiphase flow in the riser and compressibility of gas. Due to these two factors, any increment of gas flow can cause two opposite effects on the riserbase pressure, positive and negative. The negative effect can make the system unstable if it is dominant.

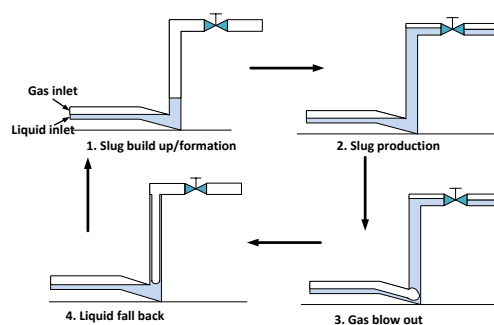


Figure 1. Severe slugging mechanism

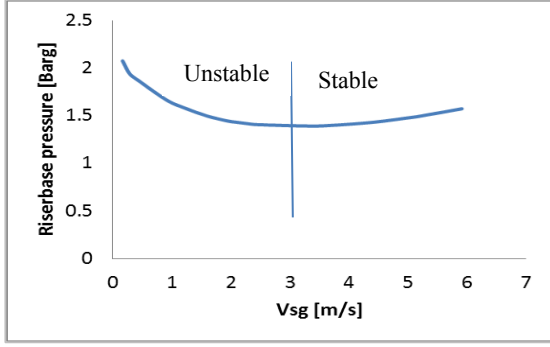


Figure 2. Stable and unstable regions for the riserbase pressure as a function of gas flow rate

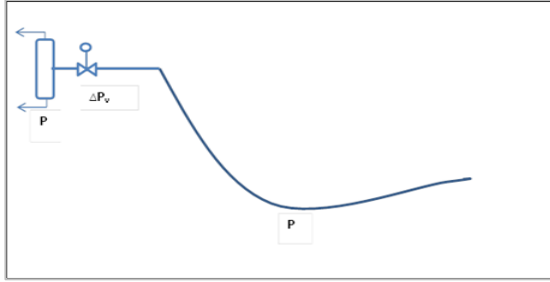


Figure 3. schematic of pipeline-riser system

Figure 2 shows the general relationship between the riserbase pressure and the gas flow rate for a given constant liquid flow rate. When the gas flow rate is low, which corresponds to a low friction loss, any increment in the gas flow rate will cause an increase in the gas-liquid ratio within the riser, hence results in a decrease in the riserbase pressure. Conversely, when the gas flow rate is large enough (on the right side of the vertical line in Fig.2) the friction loss becomes dominant; hence any increase in the gas flow rate will increase the friction loss and results in the riserbase pressure increase. The region to the left and right of the minimum value represent the unstable flow and stable flow regimes as shown in Fig.2. Figure 2 shows clearly that the system will be stable only at considerably high gas flow rates. This is the bane of gas injection as a method for slug attenuation [13]. Alternative method is therefore required for stabilizing the unstable system.

Considering a pipeline-riser system shown in Fig. 3, the riserbase pressure depends on the liquid head, frictional head, acceleration head, and pressure drop across the valve and the separator pressure. This can be shown mathematically as (1).

$$P = \Delta P_h + \Delta P_f + \Delta P_a + \Delta P_v + P_s \quad (1)$$

Where P is the riserbase pressure, ΔP_h , ΔP_f , ΔP_a , P_s and ΔP_v are the hydrostatic head, frictional head, acceleration head, separator pressure and pressure drop across the valve respectively

Assuming a constant liquid flow rate with small perturbation in gas flow rate, the riserbase response can be given as (2)

$$\frac{dP}{dQ} = \frac{d\Delta P_h}{dQ} + \frac{d\Delta P_f}{dQ} + \frac{d\Delta P_a}{dQ} + \frac{d\Delta P_v}{dQ} + \frac{dP_s}{dQ} \quad (2)$$

For the system to be stable the riserbase pressure response to the change in gas flow rate must have a positive slope as shown in (3).

$$\frac{dP}{dQ} > 0 \quad (3)$$

$$\frac{dP}{dQ} < 0 \quad (4)$$

The system will be unstable when the riserbase pressure slope is negative. The condition is given as (4).

A. Stabilizing the unstable slug flow regime with manual choking

Considering the pipeline-riser system shown in Fig.3, under unstable behaviour, the system can be stabilized by choking the topside valve. This can be achieved by increasing the pressure drop across the valve.

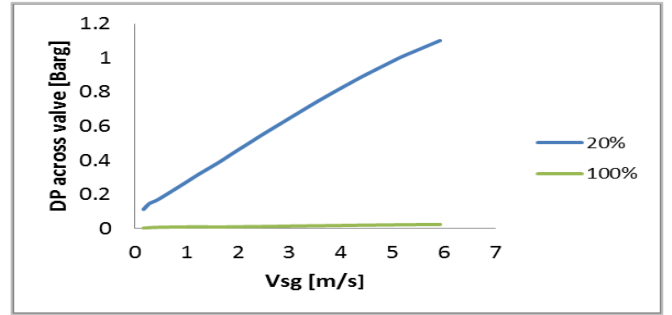


Figure 4. Pressure drop across valve as a function of gas flow rate

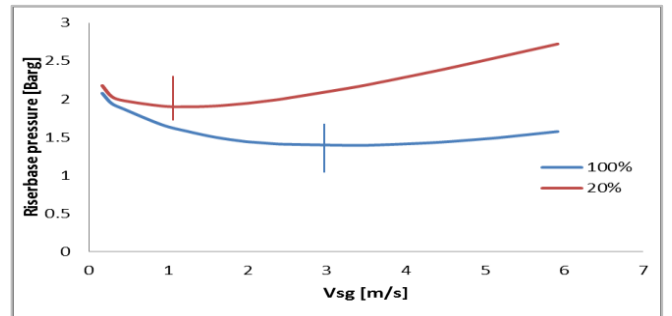


Figure 5. Use of choking to obtain stable flow

Fig.4 show a plot of pressure drop across the valve against the gas flow rate at constant liquid flow rate. The pressure drop across the valve was shown to increase as the gas flow increases for a constant valve opening. This relationship is shown in (5). When the pressure drop across the valve is sufficiently large, the region of negative slope can be sufficiently made positive as shown in Fig. 5

The pressure drop across the valve in (2) can be estimated using valve equation. Assuming linear valve

characteristics, for a given liquid flow rate, the pressure drop across the valve can be given as (5)

$$\Delta P_v = \frac{Q^2}{C_v^2 u^2 \rho} \quad (5)$$

ρ is the density of fluid flowing through the valve (mixture density), C_v is the valve coefficient, u is the valve opening with values ranging between 0 and 1 and Q is the flow across the valve. The pressure drop across the valve is a function of the flow and the valve opening as shown in (5). At constant flow rate, the only variable that can be manipulated is the valve opening. This has been previously explored for slug attenuation by many authors [1; 14][15], others developed bifurcation maps based on this concept and further designed controllers for slug attenuation[3; 16-18].

If (5) is differentiated with respect to Q keeping valve opening u constant (typical of manual choking), we have (6)

$$\frac{d\Delta P_v}{dQ} = \frac{2Q}{C_v^2 u^2 \rho} \quad (6)$$

Substituting (6) into 2 and on rearrangement, we have (7).

$$\frac{2Q}{C_v^2 u^2 \rho} > - \left[\frac{d\Delta P_r}{dQ} + \frac{d\Delta P_f}{dQ} + \frac{d\Delta P_a}{dQ} + \frac{dP_s}{dQ} \right] \quad (7)$$

This shows the condition under which manual choke valve would stabilize the unstable slug flow when the gas flow is perturbed. For this condition to hold, the pressure drop across the valve must be sufficiently large that is, the valve opening must be considerably small which means low flow through the valve.

The riserbase pressure increases as the pressure drop across the valve increases. Choking causes restriction to flow which result in flow deceleration thereby reducing the acceleration term during transient conditions. The use of choking to achieve stability contributes largely to the system pressure and this considerably reduces production rate. This is the bane of choking as a method for slug control and has been reported by many authors including [1]. Thus reducing the pressure drop across the valve would be desirable as this would lead to increase in production. One of the ways to achieve this is to use a controller. Ogazi [3] has reported the ability of controller to help stabilise an open loop unstable system however no robust stability analysis was given for this benefit. We attempt to show this next.

B. Stabilizing the unstable slug flow regime with feedback controller

The production of system is directly associated with the riserbase pressure (1), while the stability is related to the pressure gradient, dP/dQ in (2). Therefore, the aim of a slug control system can be translated as to achieve positive dP/dQ for certain flow rate with relatively low riserbase pressure P . Under feedback control, in (5) the valve opening u is not constant but varying as gas flow rate Q changes although the specific relationship between u and Q depends on the feedback law designed. Differentiating (5) therefore yields:

$$\frac{d\Delta P_v}{dQ} = \frac{2Q}{C_v^2 u^2 \rho} + \left(\frac{Q^2}{C_v^2 \rho} \right) \frac{d}{dQ} \left(\frac{1}{u^2} \right) \quad (8)$$

Comparing (6) and (8), the second term of (8) provides extra gradient to satisfy stable condition (3). In other words, active slug control can achieve oil production higher than manual choking when severe slugging is eliminated. Equation (8) also suggests that to maximise oil production of a slug control system, the second term of (8) should be maximised. This confirms that slug attenuation using choking can be more effective with the aid of controller compared with manual choking [3].

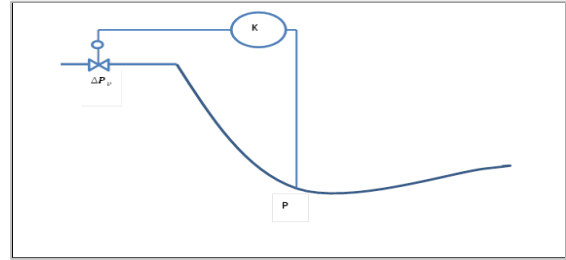


Figure 6. Pipeline-riser configuration with controlled choking

1) Design of Active feedback controller

a) Bifurcation map

The first step in the design procedure is to establish the critical point after which a controller will be designed to stabilise the system in the open loop unstable region. The bifurcation map can be generated by keeping Q constant and varying u . The pressure gradient contributed by the valve to stabilize the system can be estimated at the critical valve opening using (5).

b) Controller design

Considering a simple pipeline-riser system with feedback controller in Fig. 6, our goal is to control system response at larger valve opening. To achieve this, an extra pressure gradient must be introduced through feedback control to compensate for the gradient loss due to increased valve opening. Assuming the parameter of interest is the flow rate Q , for a slight perturbation in the gas flow rate, Q will deviate from set point Q_0 . We propose that Q can be driven to Q_0 with a feedback controller of the form:

$$u = K(Q_0 - Q) + u_0 \quad (9)$$

$$\frac{d}{dQ} \left(\frac{1}{u^2} \right) = \frac{2K}{u^3} \quad (10)$$

Therefore (8) becomes:

$$\frac{d\Delta P_v}{dQ} = \frac{2Q}{C_v^2 u^2 \rho} + \left[\left(\frac{Q^2}{C_v^2 \rho} \right) \frac{2K}{u^3} \right] \quad (11)$$

Therefore the stability condition for feedback

control is given as (12).

$$\frac{2Q}{C_v^2 u^2 \rho} + \left[\left(\frac{Q^2}{C_v^2 \rho} \right) \frac{2K}{u^3} \right] > - \frac{d\Delta P_o}{dQ} \quad (12)$$

$$\text{Where } \frac{d\Delta P_o}{dQ} = \left[\frac{d\Delta P_h}{dQ} + \frac{d\Delta P_f}{dQ} + \frac{d\Delta P_a}{dQ} + \frac{dP_s}{dQ} \right]$$

For a desired $\frac{d\Delta P_v}{dQ}$ and a given valve opening u , there exists a value of K , which stabilises the system. It is shown in (9) that large value of K will lead to increased oil production.

III. NUMERICAL CASE STUDY

In order to meet the objective of this study, numerical study on the stabilization of an unstable slug flow in pipeline-riser system was attempted.

A. Modelling and simulation of slug flow in a pipeline-riser system

LedaFlow (an industrial code) have been used for the numerical investigation of a pipeline-riser system shown in Fig. 7. The system is a 17'' pipeline-riser system with 3.7 km horizontal pipeline leading to a 0.13km riser. The fluid file used for the simulation was generated in PVTsim using the fluid compositions and properties shown in Table I.

TABLE I. FLUID PROPERTIES

Component	Gas	Oil	water
Density [Kg/m ³]	23	780	1000
Viscosity [Kg/m-s]	1.3×10^{-5}	1.1×10^{-3}	3.5×10^{-4}

TABLE II. PROPERTIES OF PIPE AND INSULATION MATERIALS

Material	Density [kg/m ³]	Specific heat [J/kgC]	Thermal conductivity [W/mC]
Steel pipe	7850	500	50
Insulation	2500	880	1

TABLE III. CASE STUDY OF TYPICAL SLUG FLOW CONDITION

Total mass flow [kg/s]	120
Gas mass fraction [-]	0.007
Oil mass fraction [-]	0.239
Water mass fraction [-]	0.754
Inlet Temperature [°C]	90
Outlet Temperature [°C]	40
PR outlet Pressure [bar]	22.5

The geometry was discretized and all the properties of material used for the pipes specified. Table II shows the properties of materials used for the pipe in this study. The heat transfer coefficient and pipe roughness values of 10W/m²-K and 4.572e-5 m were used respectively.

1) Mesh sensitivity studies

The accuracy and convergence of the solution depend largely on the mesh. The mesh size is also a very crucial factor in determining the computational time which is a major issue in numerical simulations. Therefore, mesh sensitivity studies were carried out to identify the optimum mesh size required to obtain solution which is not mesh dependent and at lowest possible computational cost. A case with 1800 cells was found to be the optimum mesh. This is in consonance with the suggestion in the online LedaFlow user manual that a mesh size of less than 5ID is fine enough for hydrodynamic slug study.

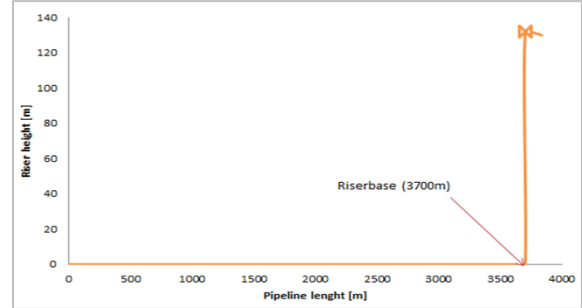


Figure 7. Geometry of pipeline-riser system

2) Stability curve

Fig. 8 shows the average riserbase pressure against the gas flow rate. The system was simulated for various gas flow rates at constant liquid flow rate of 119.16 kg/s.

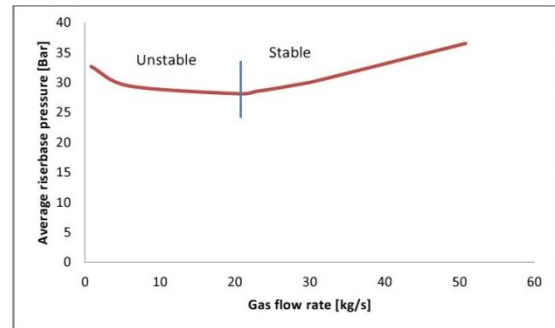


Figure 8. Riserbase pressure against gas flow rate

It is shown that at this constant liquid flow a gas flow rate (of 0.84kg/s, corresponding to a gas mass fraction at 0.007) is in unstable region. It is shown from the map that about 20kg/s gas flow rate will be needed to stabilize the system without choking. This is the bane of using gas injection as a slug mitigation technique [4]. Following (3), (6) and (7), it is proposed that when sufficient dP/dQ is added to the system such that total gradient is greater than zero, the system will become stable.

For a close look, Fig. 9 shows that without any choking, i.e. at 100% valve opening, around the operating point defined in Table III, the local gradient (dP/dQ) was estimated as - 14.29 bar/kgs⁻¹. This is in consonance with (4), thus the system is unstable. In this study, it is desired to stabilize the system around this operating condition. From (6), it was estimated that at least 14.29 bar/kgs⁻¹

gradient must be supplied by the choke in order to stabilize the system at this operating condition such that (7) is satisfied. This was achieved by choking and the corresponding bifurcation map is shown in Fig. 10.

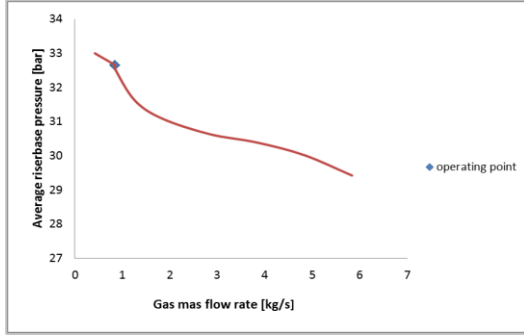


Figure 9. Stability curve showing the operating condition

1) Bifurcation map

The system was simulated for various valve openings and bifurcation map was generated for a typical slug flow for the boundary conditions shown in Table III.

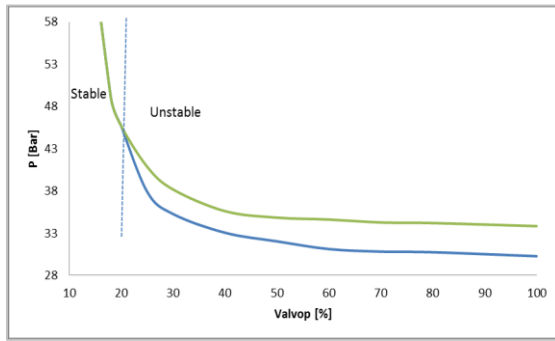


Figure 10. Riserbase pressure against valve opening

The valve was significantly choked to 20% opening to achieve stability by adding the required gradient to the system. This gradient was supplied by the pressure drop across the valve which added about 14 bar pressure to the system. It is desired to reduce the magnitude of this pressure so that the system pressure can be lowered for higher production.

B. Stabilizing the pipeline-riser system with feedback control

Having established the bifurcation point with manual choking and the pressure gradient contributed by the valve, the next goal is to control the system response at larger valve opening.

1) Slug controller design

It has been shown in (8) that with the help of active control, a system can be stabilized at larger valve opening. In this study, we attempt to control the gas flow rate using a simple proportional controller. At 22% valve opening for example, the gradient supplied at this valve opening was $10.71 \text{ bar/kgs}^{-1}$ which was less than the required $14.29 \text{ bar/kgs}^{-1}$. From (8), it was shown that a controller can provide this shortfall. The gain required to meet this

shortfall gradient was estimated from (11). The minimum required gain required to stabilize our system at 22% valve opening was obtained as 0.0794.

2) Implementation of the active controller

The estimated controller gain 0.0794 was implemented using the inbuilt proportional controller structure in LedaFlow.

Fig.11 shows the system response to the application of control designed using the new method. The simulation was run for about 5000seconds before the controller was introduced. The reference valve opening u_0 was initially set at 20% valve opening, the controller was able to stabilise the system, and after 11000s, the reference valve opening was changed to 22% and the controller was still able to stabilise the system, but when u_0 was opened beyond this value at 23% from 16000s, the system became closed loop unstable. A benefit of 5% reduction in the riser base pressure from 45.73 bar to 43.4 bar was recorded.

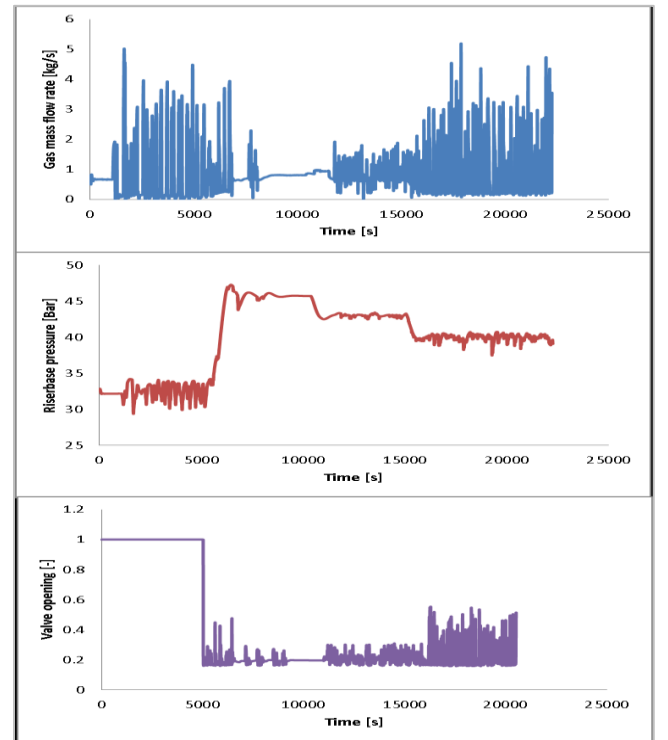


Figure 11. System response to active feedback control

IV. CONCLUSION

The theoretical understanding of slug attenuation potential of active feedback control at large valve opening has been investigated. The following conclusions can be drawn from the study.

- A new methodology for slug flow stability analysis has been reported.
- Active feedback control helps to maximise slug attenuation by optimising the pressure drop across the valve compared with manual choking.

- For the specific case study, additional 2% valve opening translating into 3% reduction in riserbase pressure was achieved. This practically implies increase in oil production for the system.
- With the help of a more robust controller, greater benefit might be achieved using the proposed method.

ACKNOWLEDGMENT

The authors thank the Niger Delta Development Commission (NDDC), Nigeria who supports the PhD of Adegboyega Ehinmowo

REFERENCES

- [1] Yocum, B. (1973), "Offshore riser slug flow avoidance: mathematical models for design and optimization", *SPE European Meeting*, April 2-3, London, England, .
- [2] Taitel, Y. (1986), "Stability of severe slugging", *International Journal of Multiphase Flow*, vol. 12, no. 2, pp. 203-217.
- [3] Ogazi, A. I. (2011), *Multiphase severe slug flow control* (PhD thesis), Cranfield University, Bedfordshire, UK.
- [4] Jansen, F. and Shoham, O. (1994), "Methods for eliminating pipeline-riser flow instabilities", *SPE Western Regional Meeting*, 23-25 March, Long Beach, California, USA, .
- [5] Hedne, P. and Linga, H. (1990), "Suppression of terrain slugging with automatic and manual riser choking", *Advances in Gas-Liquid Flows*, vol. 155, no. 19, pp. 453-460.
- [6] Storkaas, E., Skogestad, S. and Godhavn, J. (2003), "A low-dimensional dynamic model of severe slugging for control design and analysis", *11th International Conference on Multiphase flow (Multiphase03)*, 11-13 June, San Remo, Italy, pp. 117-133.
- [7] Storkaas, E. and Skogestad, S. (2003), "Cascade control of unstable systems with application to stabilization of slug flow", *International Symposium on Advanced Control of Chemical Processes AdChem'03*, 11-14 January, Hong Kong, .
- [8] Storkaas, E. and Godhavn, J. (2005), "Extended slug control for pipeline-riser systems", *12th International Conference on Multiphase Production Technology*, 25-27 May, Barcelona, Spain, BHR Group, Cranfield, Bedfordshire, UK, .
- [9] Storkaas, E. and Skogestad, S. (2007), "Controllability analysis of two-phase pipeline-riser systems at riser slugging conditions", *Control Engineering Practice*, vol. 15, no. 5, pp. 567-581.
- [10] Di Meglio, F., Kaasa, G. and Petit, N. (2009), "A first principle model for multiphase slugging flow in vertical risers", *Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference*, 16-18 December, Shanghai, P.R China, IEEE, pp. 8244.
- [11] Di Meglio, F., Kaasa, G., Petit, N. and Alstad, V. (2010), "Model-based control of slugging flow: an experimental case study", *American Control Conference (ACC)*, 30 June - 02 July, Marriott Waterfront, Baltimore, MD, USA, IEEE, pp. 2995.
- [12] Di Meglio, F., Petit, N., Alstad, V. and Kaasa, G. (2012), "Stabilization of slugging in oil production facilities with or without upstream pressure sensors", *Journal of Process Control*, vol. 22, pp. 809-822.
- [13] Hill, T. (1990), "Gas injection at riser base solves slugging flow problems", *Oil and Gas Journal*, vol. 26, pp. 88-92.
- [14] Farghaly, M. (1987), "Study of severe slugging in real offshore pipeline riser-pipe system", *5th SPE Middle East Oil Show*, 7-10 March, Manama, Bahrain, .
- [15] Schmidt, Z., Brill, J. and Beggs, H. (1979), "Choking can eliminate severe pipeline slugging", *Oil and Gas Journal*, vol. 12, pp. 230-238.
- [16] Storkaas, E. (2005), *Stabilizing control and controllability: control solutions to avoid slug flow in pipeline-riser systems* (PhD thesis), Norwegian University of Science and Technology, NTNU, Norway.
- [17] Ogazi, A., Ogunkolade, S., Cao, Y., Lao, L. and Yeung, H. (2009), "Severe slugging control through open loop unstable PID tuning to increase oil production", *14th International Conference on Multiphase Technology*, 17-19 June, Cannes, France, pp. 17.
- [18] Ogazi, A., Cao, Y., Yeung, H. and Lao, L. (2010), "Slug control with large valve openings to maximize oil production", *SPE Journal*, vol. 15, no. 3, pp. 812-821.

Transmissibility Damage Indicator for Wind Turbine Blade Condition Monitoring

Long Zhang*, Ziqiang Lang*, Wen-Xian Yang[†]

*Department of Automatic Control and System Engineering, University of Sheffield, UK

Email: {long.zhang, z.lang}@sheffield.ac.uk

[†]School of Marine Science&Technology, Newcastle University, UK

Email: wenxian.yang@ncl.ac.uk

Abstract—Wind turbine blade has a high failure rate as it has to constantly work under varying loads and occasionally suffer extreme weather like storm and sleet. Online condition monitoring plays an important role in finding early or minor damage to avoid catastrophic failure. Due to the large size of blade, condition monitoring systems often use multi-sensors to obtain the blade conditions in different locations. To deal with the multi-sensor data, most conventional frequency methods deal with the multiple sensor data separately without consideration of their correlations. Further, they have to use multiple indicators to represent conditions of different locations. Third, most of frequency methods are dependent on the dynamic loadings and therefore different frequency features and thresholds have to be determined under different loading conditions. To address these problems, this paper employs transmissibility analysis for multi-sensor based wind turbine blade condition monitoring. Transmissibility analysis considers the relations between different sensors in the frequency domain, and only produces one single transmissibility damage indicator to represent the condition of the whole wind turbine blade. This is independent of loading conditions and computationally efficient as the main computations only involve Fast Fourier Transform (FFT). The effectiveness of the transmissibility analysis is demonstrated using both simulation example and experimental data analysis.

Index Terms—Transmissibility analysis, Wind turbine blade, Condition monitoring, Frequency methods

I. INTRODUCTION

Wind turbines have been installed increasingly over the past decade and this trend will continue in the future as countries from all over the world set up ambitious targets of emission reduction by using renewable energy. The wind power is generated by converting wind kinetic energy to electrical energy. A big challenge is how to monitor health condition of wind turbine systems and components including blades to improve their reliability in a cost-effective and time-efficient manner, especially for these expensive and multi-MW turbines [1]. Surveys show wind turbine blade is on the top of maintenance list and therefore its condition has to be evaluated carefully during its long service lifetime [2]. Conventional visual inspections and manual tapping tests at regular intervals may not find the minor damages at an early stage. Further, such detection accuracy is highly dependent on inspectors' experiences. Finally, it is not economic-effective method due to economic loss caused by the down-time for inspections [3]. More recently, sensor based condition monitoring systems provide online and real-time monitoring without affecting wind

turbine operation, which is attracting more attentions from both academics and industry. Due to the large size of wind turbine blade, multi-sensor based systems are needed to monitor its condition. Such a system produces a large amount of multi-dimensional data and therefore efficient and effective multi-sensor data processing methods are highly desirable for wind turbine blade condition monitoring.

Frequency domain analysis methods have been widely applied to process multi-sensor data measurements, such as spectrum analysis, frequency spike detection and enveloped spectrum methods [4]. However, most of these methods deal with multiple dimensional data separately and need to use multiple thresholds. Alternatively, the combination of frequency methods and artificial intelligent algorithms, such as artificial neural networks, fuzzy logic and support vector machine, have been widely studied [1]. In general, these methods first map the frequency features to nonlinear space, and then carry out detection or diagnosis tasks in a linear space. A common drawback is that they may need large computations to optimize nonlinear model parameters. The existing industrial condition monitoring system may not afford to take the time-consuming tasks due to its low-level computing ability. Further, these frequency domain methods are loading or working condition dependant. In other words, the frequency characteristics vary with the wind turbine blade loadings. To cope with this issue, different frequency features under various loading conditions have to be considered. This often increases the difficulties of real applications.

Frequency transmissibility analysis is well studied in structure dynamics, especially for multiple degree of freedom (MDOF) structure systems [5], [6]. Transmissibility function (TF) is defined as the spectra ratio between two different response measurements and therefore it naturally represents the frequency characteristic relations between different sensors. This is different from most existing frequency methods that do not consider of the relationship of different responses. In theory, TF can be derived using the dynamic properties of the system structure for a general MDOF system no matter the inputs are harmonic or random [5], [7], [8]. In other words, TF can be independent of different working conditions as it has been shown that, given the MDOF structure system, TF is solely dependent on physical parameters, such as mass, stiffness and damping [9], [10]. Recently, a transmissibility damage

indicator (TDI) is proposed for structure health monitoring in [11], where the correlation between reference and in-service TFs is used to estimate the structural conditions. The main principal is that physical damages, like crack, can result in the changes in TFs. TDI is capable to deal with multiple sensor data and only produces one index to represent the condition. Further, it is computationally extremely fast as its main computation is from the calculation of Fourier transform of multiple sensor data via well-known Fast Fourier Transform (FFT). In this paper, the TDI is used for wind turbine blade condition monitoring and its effectiveness is evaluated by both simulation and experimental data analysis.

II. TRANSMISSIBILITY DAMAGE INDICATOR

In this section, the MDOF structure model of wind turbine blade is first introduced, and then its transmissibility functions that represents the blade dynamic properties, like stiffness and damping ratios, are given. Finally, the transmissibility damage indicator is applied for wind turbine blade condition monitoring.

Wind turbine blade is generally composed of composite materials and has sandwich structures. It can be simplified to be a linear MDOF system that is given by [12]

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{C}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{F} \quad (1)$$

where the applied force \mathbf{F} is the system input vector and displacement \mathbf{x} is system output vector, while

$$\mathbf{M} = \begin{bmatrix} m_1 & 0 & 0 & \dots & 0 \\ 0 & m_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & m_n \end{bmatrix} \quad (2)$$

$$\mathbf{C} = \begin{bmatrix} c_1 + c_2 & -c_2 & 0 & 0 & \dots & 0 \\ 0 & -c_2 & c_2 + c_3 & -c_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -c_n & c_n \end{bmatrix} \quad (3)$$

$$\mathbf{K} = \begin{bmatrix} k_1 + k_2 & -k_2 & 0 & 0 & \dots & 0 \\ 0 & -k_2 & k_2 + k_3 & -k_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -k_n & k_n \end{bmatrix} \quad (4)$$

are system structure parameters representing the system mass, damping and stiffness matrices, respectively.

The TF is defined as the ratio of spectra of two different output measurements. Therefore the spectra have to be obtained first. All the spectra of these responses are given by $F = [X_1, \dots, X_i, \dots, X_n]^T$ where

$$\begin{aligned} X_i &= [x_{i1}e^{-jw_1}, \dots, x_{ir}e^{-jw_r}, \dots, x_{iN}e^{-jw_N}] \\ &= [x_{i1}(w_1), \dots, x_{ir}(w_r), \dots, x_{iN}(w_N)] \end{aligned} \quad (5)$$

with x_{ir} is a complex number.

The TF is the spectra ratio between two adjacent responses. Let i th and $(i+1)$ th denote the two adjacent measurement indexes, where $i = 1, \dots, n-1$, and the transmissibility function for neighboring measurements can be written as

$$T_{i(i+1)} = [t_{i(i+1)}(w_1), \dots, t_{i(i+1)}(w_r), \dots, t_{i(i+1)}(w_N)] \quad (6)$$

where

$$t_{i(i+1)}(w_r) = \frac{x_{ir}(w_r)}{x_{(i+1)r}(w_r)} = \frac{x_{ir}e^{-jw_r}}{x_{(i+1)r}e^{-jw_r}} = \frac{x_{ir}}{x_{(i+1)r}} \quad (7)$$

The total number of such transmissibility functions is $L = (n-1)$. To make the expression $t_{i(i+1)}(w_r)$ simpler, the combinations of the two index variables $i, i+1$ can be rewritten as one single vector, which is shown as follows:

$$\{t_{i(i+1)}(w_r), i = 1, \dots, n-1\} = \{\tau_l(w_r), l = 1, \dots, L\} \quad (8)$$

Therefore, the total spectra of all the transmissibility functions can be written as

$$\mathbf{\Gamma} = \begin{bmatrix} \tau_1(w_1) & \tau_1(w_2) & \dots & \tau_1(w_N) \\ \tau_2(w_1) & \tau_2(w_2) & \dots & \tau_2(w_N) \\ \vdots & \vdots & \ddots & \vdots \\ \tau_L(w_1) & \tau_L(w_2) & \dots & \tau_L(w_N) \end{bmatrix} \quad (9)$$

To detect damage, the transmissibility correlation (TC) between healthy ${}^h\tau$ and in-service τ corresponding to one frequency w_r is used, which is defined as

$$TC(w_r) = \frac{|\sum_{l=1}^L \tau_l(w_r) {}^h\bar{\tau}_l(w_r)|^2}{[\sum_{l=1}^L \tau_l(w_r) \bar{\tau}_l(w_r)][\sum_{l=1}^L {}^h\tau_l(w_r) {}^h\bar{\tau}_l(w_r)]} \quad (10)$$

where the upper bar represents the conjugate operator and h represent the healthy condition. TDI is the average of transmissibility correlations at all frequencies w_1, \dots, w_N , which is given by

$$TDI = \frac{1}{N} \sum_{r=1}^N TC(w_r) \quad (11)$$

It is worth mentioning that the range of TDI values is $[0, 1]$. TDI value is 1 or near 1 for healthy case and it is smaller than 1 for damaged cases. The more serious the damage is, the smaller the TDI value is.

III. NUMERICAL EXAMPLE

In this section, a 10-DOF simulation system is used to test the performance of TDI. For the healthy condition, the system structure parameters are chosen as $m_1 = m_2 = m_3 = m_4 = m_5 = m_6 = m_7 = m_8 = m_9 = m_{10} = 0.8 \times 10^5$, $k_1 = k_2 = k_3 = k_4 = k_5 = k_6 = k_7 = k_8 = k_9 = k_{10} = 4 \times 10^7$ and $c_1 = c_2 = c_3 = c_4 = c_5 = c_6 = c_7 = c_8 = c_9 = c_{10} = 1.5 \times 10^6$ [13]. For damaged conditions, 10 damage levels are simulated where the stiffness parameter of the 5th coordinate is reduced to the [90% 80% 70% 60% 50% 40% 30% 20% 10%] of its original value, respectively. The input signal is chosen to be a multiple sine wave and its frequency range is from 1 Hz to 20 Hz, with step change being 1 Hz.

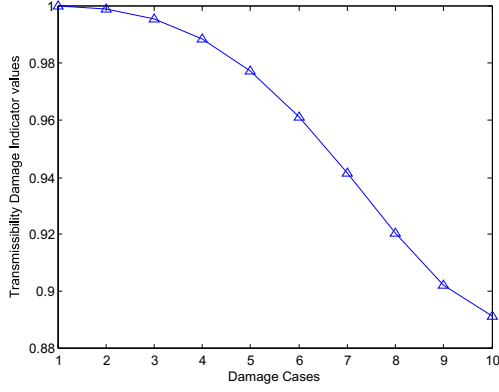


Fig. 1: TDI with full responses

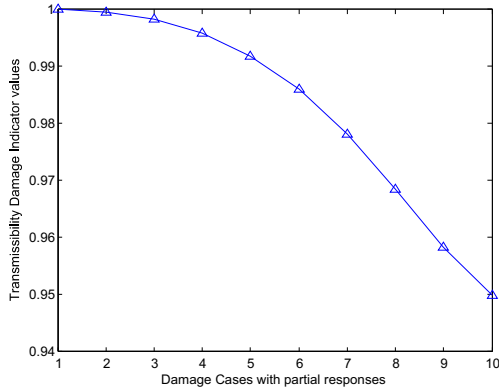


Fig. 2: TDI with partial responses

In order to test the performance of the proposed TDI method, two different cases are considered. The first one is an ideal case where the system inputs are applied in the 10 coordinates with full frequency range and all the system outputs are measured. The second case considers the partial responses where only seven outputs are measured and the 3rd, 6th and 9th coordinates outputs are not used. Further, to test the robustness of TDI, random noise with 25dB signal-to-noise ratio is added on the response data. To carry out condition monitoring, the Fourier transforms of responses data is first computed. Using the these Fourier transform results, the TFs are calculated using the Equ. (6). Finally, the healthy case without any changes in stiffness parameters is chosen as a reference and then the TDI values for the ten cases of different damage levels are computed using Equ. (11). All the TDI results for these two cases are shown in Fig. 1 and Fig. 2. It is clearly shown that in all cases TDI can detect the system parameter changes and their values decrease with the increased damage severity levels. TDI can clearly distinguish different damage severities levels and therefore it is a sensitive indicator of damage.

IV. CASE STUDY

In this section, the performance of TDI will be evaluated using real experimental data from blade testing. The real blade has to first experience a long fatigue test with 10 million cycles of a sinusoidal loading profile, then it is under the

TABLE I: Data groups of different damage levels

case	number of groups
health	4
damage level 1	4
damage level 2	4
damage level 3	4
damage level 4	4

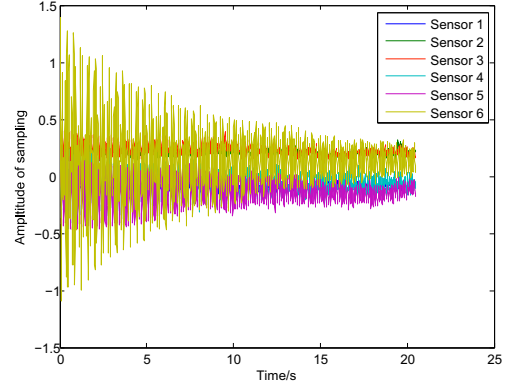


Fig. 3: Blade hammer response data

static test, where heavy loads are applied to the blade statically in flap direction. The static test generates huge stresses to the blade using forced displacement, resulting in a joint debonding between the the aero-shell and spar beam. After the damage occurs, the hammer-striking inputs are applied to the blade again under the static test, the blade responses are simultaneously measured by 6 accelerometer sensors along the blade. The sampling rate is 100 Hz and each sampling lasts for 20.48 seconds, leading to 2048 data points. In total, 20 groups of data were collected during healthy and crack damage severities, which is shown in Table I. A representative group of data is plotted in Fig. 3 and it can be seen the responses from six sensors are different in their amplitudes due to their different locations.

Once again, the TFs and TDI of these response data are computed using Equ. (6) and Equ. (11), respectively, and the final values are shown in Fig. 4. The first 4 groups of data were collected in healthy condition. Therefore, the TDI values are 1. The following 4 groups were collected when damage level is 1. Their TDI values are about 0.85, which indicates there were some damages. For the reminding groups, TDI values can correctly show their damage levels. In a word, TDI values can estimate the condition accurately and very clearly indicate different severities of damage.

The present study has demonstrated the effectiveness of the TDI method using both simulation and experimental data. Future work will investigate the application of TDI method to real world data and compare with other state-of-art methods. Further, the information of damage location is also important for fault diagnosis. A potential work on damage localization based on the TDI method can be carried out.

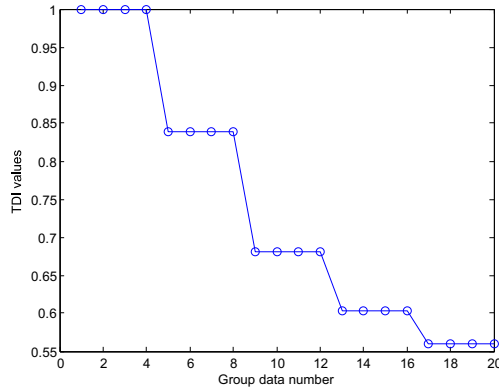


Fig. 4: TDI values for wind turbine blade accelerometer data

V. CONCLUSION

This paper has introduced the transmissibility analysis method for wind turbine blade condition monitoring. The blade is modelled by the multiple degrees of freedom system and then the transmissibility functions were computed in the frequency domain. Finally, the correlation based transmissibility damage indicator was used to examine the wind turbine blade conditions. Good results from both simulation and experimental data analysis have shown the transmissibility damage indicator can not only estimate the damage condition accurately, but also can indicate the damage severities well.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support of EU FP7 optimus grant for this research work.

REFERENCES

- [1] Z. Hameed, Y. S. Hong, Y. M. Cho, S. H. Ahn, and C. K. Song, "Condition monitoring and fault detection of wind turbines and related algorithms: A review," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 1, pp. 1–39, 2009.
- [2] A. Ghoshal, M. J. S., M. J. Schulz, and P. F. Pai, "Structural health monitoring techniques for wind turbine blades," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 85, no. 3, pp. 309–324, 2000.
- [3] A. Jüngert, "Damage detection in wind turbine blades using two different acoustic techniques," *Journal of Nondestructive Testing*, no. 12, pp. 1–10, 2008.
- [4] M. P. Norton and D. G. Karczub, *Fundamentals of noise and vibration analysis for engineers*. Cambridge university press, 2003.
- [5] M. J. Schulz, A. S. Naser, P. F. Pai, M. S. Linville, and J. Chung, "Detecting structural damage using transmittance functions," in *Proceedings SPIE the Intertional Society for Optical Engineering*. SPIE the International Society for OpticalL, 1997, pp. 638–644.
- [6] H. F. Zhang, M. J. Schulz, A. Naser, F. Ferguson, and P. F. Pai, "Structural health monitoring using transmittance functions," *Mechanical Systems and Signal Processing*, vol. 13, no. 5, pp. 765–787, 1999.
- [7] A. M. R. Ribeiro, J. M. M. Silva, and N. M. M. Maia, "On the generalisation of the transmissibility concept," *Mechanical Systems and Signal Processing*, vol. 14, no. 1, pp. 29–35, 2000.
- [8] M. Fontul, A. M. R. Ribeiro, J. M. M. Silva, and N. M. M. Maia, "Transmissibility matrix in harmonic and random processes," *Shock and Vibration*, vol. 11, no. 5, pp. 563–571, 2004.
- [9] T. J. Johnson and D. E. Adams, "Transmissibility as a differential indicator of structural damage," *Journal of Vibration and Acoustics*, vol. 124, no. 4, pp. 634–641, 2002.
- [10] S. N. Ganeriwala, J. Yang, and M. Richardson, "Using modal analysis for detecting cracks in wind turbine blades," *Sound and Vibration*, vol. 45, no. 5, p. 10, 2011.
- [11] N. M. M. Maia, R. A. B. Almeida, A. P. V. Urgueira, and R. P. C. Sampaio, "Damage detection and quantification using transmissibility," *Mechanical Systems and Signal Processing*, vol. 25, no. 7, pp. 2475–2483, 2011.
- [12] —, "Damage detection and quantification using transmissibility," *Mechanical Systems and Signal Processing*, vol. 25, no. 7, pp. 2475–2483, 2011.
- [13] Z. Q. Lang, P. F. Guo, and I. Takewaki, "Output frequency response function based design of additional nonlinear viscous dampers for vibration control of multi-degree-of-freedom systems," *Journal of Sound and Vibration*, vol. 332, no. 19, pp. 4461–4481, 2013.

Condition monitoring of wind turbines based on extreme learning machine

Peng Qian, Xiandong Ma, Yifei Wang

Engineering Department
Lancaster University
Lancaster, UK LA1 4YR
p.qian@lancaster.ac.uk

Abstract—nowadays, wind turbines have been widely installed in many areas, especially in remote locations on land or offshore. Routine inspection and maintenance of wind turbines has become a challenge in order to improve reliability and reduce the energy of cost; thus adopting an efficient condition monitoring approach of wind turbines is desirable. This paper adopts extreme learning machine (ELM) algorithms to achieve condition monitoring of wind turbines based on a model-based condition monitoring approach. Compared with the traditional gradient-based training algorithm widely used in the single-hidden layer feed forward neural network, ELM can randomly choose the input weights and hidden biases and need not be tuned in the training process. Therefore, ELM algorithm can dramatically reduce learning time. Models are identified using supervisory control and data acquisition (SCADA) data acquired from an operational wind farm, which contains data of the temperature of gearbox oil sump, gearbox oil exchange and generator winding. The results show that the proposed method can efficiently identify faults of wind turbines.

Keywords—component; Wind turbines, condition monitoring, SCADA data, model-based approach, artificial neural network, extreme learning machine.

I. INTRODUCTION

Nowadays, wind power has been considered as one of most sustainable and eco-friendly energy sources. Wind turbines are widely installed in different places, especially in remote locations on land or offshore, because in these locations the wind resource is stronger and more reliable, and visual and noise impacts can be reduced [1]. However, routine inspection and maintenance of wind turbines has become a challenge in these remote areas. The maintenance cost for wind turbines usually accounts for a considerable proportion of income [2-3]. Consequently, it is essential to develop efficient condition monitoring techniques for wind turbines, providing information about the past and current conditions of the turbines and enabling optimal scheduling of maintenance activities. Supervisory control and data acquisition (SCADA) data have been considered to be the commonly used monitoring data system that are applied to monitor and control devices in many industrial applications [4], such as space flight and aviation, transportation, biological medicine, and power energy sectors. These systems can generate monitoring data that helps to build models of a process operating under different conditions.

The schematic diagram of the model-based condition monitoring approach is illustrated in Fig. 1 where the data generated from condition monitoring system or equivalents are used as inputs using models to predict the output signals of a physical process. Then, actual output signals are compared with the signals that are predicted by the model for given input signals. Differences between actual output signals and the model predicting signals could be caused by changes in the process, possibly due to the occurrence of faults [5]. The residual signal can be an important indicator to provide an early warning of the impending component failure.

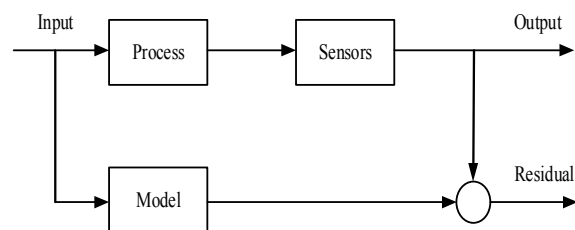


Fig. 1 Schematic diagram of model-based condition monitoring system [5]

A number of supervised learning methods have been applied in model-based condition monitoring system of wind turbines, such as artificial neural networks (ANNS) [6]. However, ANNS have suffered a drawback in the real-time implementation [7], which limits their engineering applications. Extreme learning machine (ELM) is considered in this paper due to its extremely fast learning speed [8-9]. The paper is organised as follows. The principle of ELM is described in Section 2; SCADA data including the temperatures of gearbox oil exchange, gearbox oil sump and generator winding are then employed to verify the effectiveness of the proposed method and the results are presented in Section 3. Conclusions and suggestions for future work are given in Section 4.

II. EXTREME LEARNING MACHINE

A single-hidden layer feed forward neural network (SLFN) has been widely used in many fields such as mode recognition and state prediction, because of its efficient learning skills [10-12]. However, SLFN always selects a gradient-based neural network algorithm as its training algorithm. The traditional gradient-based training algorithms have some disadvantages such as trapping at local minima, the overtraining, and the high computing burdens, which causes longer training time of the SLFN during the learning process [7]. As a relatively new

technique, the ELM can avoid this drawback [8-9]. Compared with the traditional gradient-based training algorithms, ELM randomly chooses the input weights and hidden biases and needs not be tuned in the training process. Thus, ELM algorithm features an extremely faster learning speed than most popular learning algorithms such as back-propagation, and thus dramatically reduce learning time. Furthermore, if the chosen activation function is infinitely differentiable, the ELM can identify distinct samples exactly with zero error under the condition of equal number of hidden-layer neurons and distinct samples processed in the ANN.

The schematic diagram of a single-hidden layer feed forward neural network is shown in Fig.2, which is consisted of an input layer, a hidden layer and an output layer. It assumes that the input layer and the hidden layer have n and L neurons, respectively, while the output layer has m neurons.

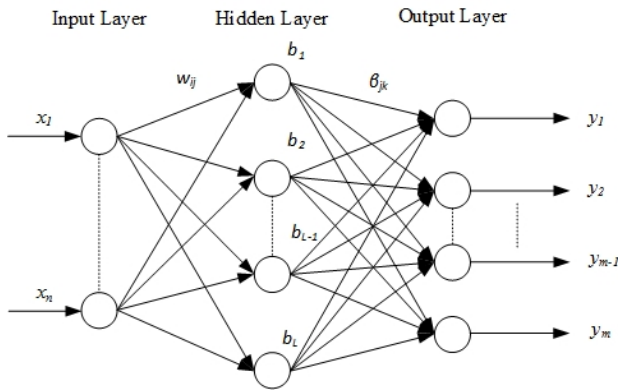


Fig. 2 Schematic diagram of single-hidden layer feed forward neural network (SLFN)

The coefficients ω_{ij} represent the input weight matrix between the input layer and the hidden layer; β_{jk} denotes the output weight matrix connecting the hidden layer and the output layer; b represents the bias of the hidden layer matrix. These parameters can be defined

$$\omega = \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2n} \\ \vdots & \vdots & & \vdots \\ \omega_{L1} & \omega_{L2} & \cdots & \omega_{Ln} \end{bmatrix}_{L \times n} \quad (1)$$

$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & & \vdots \\ \beta_{L1} & \beta_{L2} & \cdots & \beta_{Lm} \end{bmatrix}_{L \times m} \quad (2)$$

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_L \end{bmatrix}_{L \times 1} \quad (3)$$

Given datasets with Q distinct training samples, X and Y are input matrix and output matrix, respectively.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1Q} \\ x_{21} & x_{22} & \cdots & x_{2Q} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nQ} \end{bmatrix}_{n \times Q} \quad (4)$$

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1Q} \\ y_{21} & y_{22} & \cdots & y_{2Q} \\ \vdots & \vdots & & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mQ} \end{bmatrix}_{m \times Q} \quad (5)$$

If there exists an ELM with L neurons in the hidden layer and an activation function $g(\cdot)$ can approximate the Q samples with zero error, this ELM can be represented by

$$T = \begin{bmatrix} \sum_{i=1}^L \beta_{i1} g(\omega_i x_j + b_i) \\ \sum_{i=1}^L \beta_{i2} g(\omega_i x_j + b_i) \\ \vdots \\ \sum_{i=1}^L \beta_{im} g(\omega_i x_j + b_i) \end{bmatrix}_{m \times 1} \quad \left(\begin{array}{l} j = 1, 2, \dots, Q \\ i = 1, 2, \dots, L \end{array} \right) \quad (6)$$

$$\omega_i = [\omega_{i1} \quad \omega_{i2} \quad \cdots \quad \omega_{in}]$$

$$x_j = [x_{1j} \quad x_{2j} \quad \cdots \quad x_{nj}]^T$$

For simplicity, eq. (6) can be compactly described as,

$$H \beta = T' \quad (7)$$

where T' is the transpose matrix of T and H is the output matrix of the hidden layer of the ELM. The matrix H can be expressed as,

$$H = \begin{bmatrix} g(\omega_1 x_1 + b_1) & g(\omega_2 x_1 + b_2) & \cdots & g(\omega_L x_1 + b_L) \\ g(\omega_1 x_2 + b_1) & g(\omega_2 x_2 + b_2) & \cdots & g(\omega_L x_2 + b_L) \\ \vdots & \vdots & & \vdots \\ g(\omega_1 x_Q + b_1) & g(\omega_2 x_Q + b_2) & \cdots & g(\omega_L x_Q + b_L) \end{bmatrix}_{Q \times L} \quad (8)$$

When the input weight matrix ω and the hidden layer bias matrix b are initialized, the hidden layer output matrix H can be uniquely determined. The output weight matrix β can be calculate as follow,

$$\min_{\beta} \|H \beta - T'\| \quad (9)$$

During this process, the input weight matrix ω and the hidden layer bias matrix b do not need to be changed and the solution can be expressed,

$$\hat{\beta} = H^+ T' \quad (10)$$

This process is equivalent to finding a unique smallest norm least-squares solution of the linear system in eq. (7). The matrix H^+ is the Moore-Penrose generalized inverse of the hidden layer output matrix H , which can be derived through the singular value decomposition (SVD) method.

The Fig. 3 demonstrates the flowchart of ELM algorithm. There are four steps to implement the ELM algorithm, including (i) design of the SLFN structure, (ii) random choice of the input weights ω and hidden biases b , (iii) acquisition of the initial hidden layer output matrix H and the output weights β , and finally (iv) improvement and updating of the hidden layer output matrix H and output weights β .

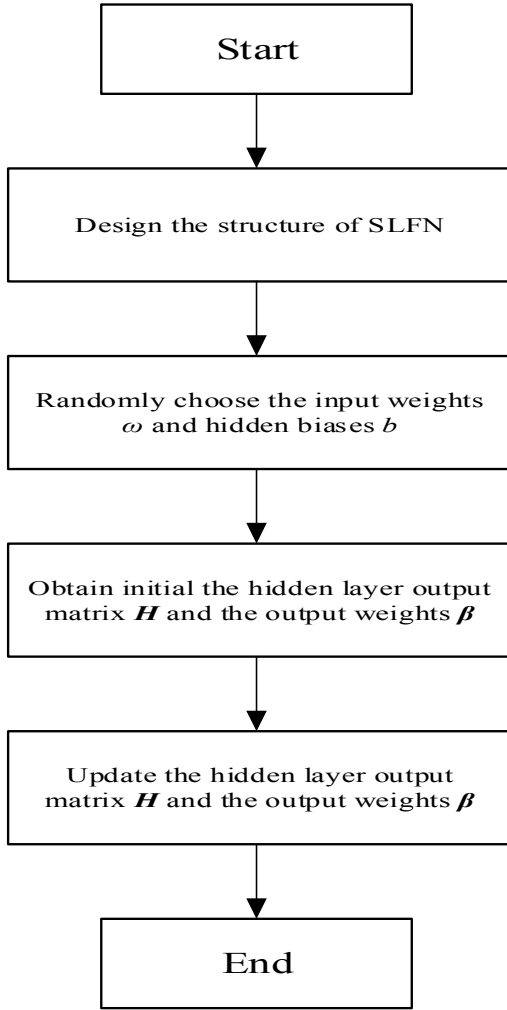


Fig. 3 The flowchart of extreme learning machine (ELM)

III. RESULTS AND ANALYSIS

Supervisory control and data acquisition (SCADA) system is an industrial automation control system, which has been widely used in many industries, including oil and gas, metallurgy, manufacturing, railway system, transportation, energy, and power systems. Modern SCADA systems work relying on multiple hardware and software elements and IT technologies to monitor, gather, and process data. In power systems, SCADA is a mature technology and it is used for data collection, device

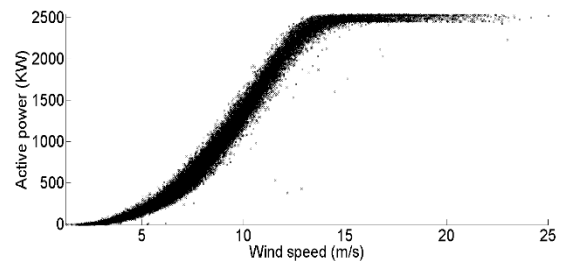
control, parameter adjustment, signal alarm generation and other functions.

SCADA data used in this paper were obtained from an operational wind farm. The use of actual operational data of wind turbines is a good method in order to demonstrate the effectiveness of the proposed algorithms. The data cover 12 months' duration and consist of 128 parameters that contain various temperatures, pressures, vibrations, power outputs, wind speed and digital control signals. In order to reduce the amount of data gathered from the operating wind turbines, SCADA data are usually sampled at 10 minute interval.

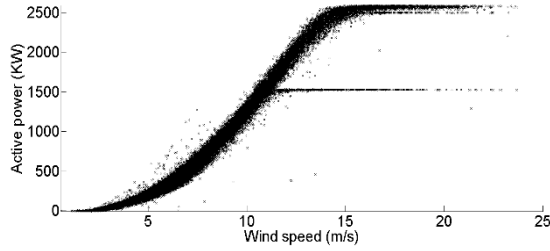
A. Model identification

Examples of the power curve of turbines are illustrated in Fig. 4. Fig. 4 (a) shows a normal power curve of the reference wind turbine; turbine power varies cubically with wind speed, and wind speed varies continuously on time-scales. When the wind speed is lower than the cut-in speed (4 m/s in this case), the turbine does not produce any power because the rotor torque is too low; while the wind speed is above the cut-out speed (15 m/s in this case), the turbine does not produce any power either because it is shut down to protect the turbine. If the wind speed is above the rated wind speed (15 m/s in this case) but below the cut-out speed, the turbine's output power is capped at the rated power. Fig. 4 (b) shows the power curve of a wind turbine operating for a period with reduced power output following a gearbox fault at some points. Other points shown in the curves might be because of the occurrence of maintenance periods or indicate the turbines were inactive when the wind speed was too high or too slow.

It has been found that gearbox oil sump temperature, gearbox oil exchange temperature and generator winding temperature change with both power output and cooling air temperature. Therefore in order to achieve an appropriate model identification, wind speed and power output are selected as the inputs while the temperature is considered as the output. This multiple-input and single-output (MISO) approach allows a more sensitive detection, which naturally mean that temperatures have a close relationship with both wind speed and power output. Furthermore, it is essential to choose the fault-free wind turbine as the reference turbine to train the model with the ELM method as proposed above.



a. Power curve of the fault-free turbine



b. Power curve of the turbine with a gearbox fault

Fig. 4 Examples of power curve of the turbines

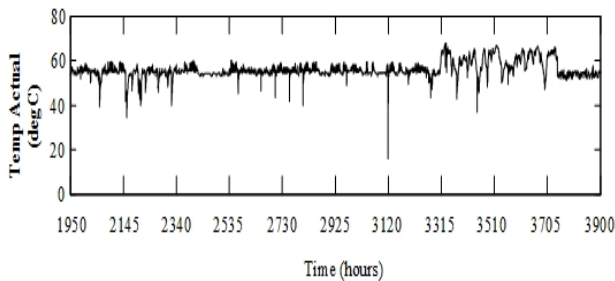
B. Detection of faults

The wind speed and the power output are selected as the input for the model identification to test whether the proposed ELM approach is able to detect faults of the wind turbine by comparing corresponding model predicting signals with actual temperatures of gearbox oil sump, gearbox oil exchange and generator winding from SCADA system.

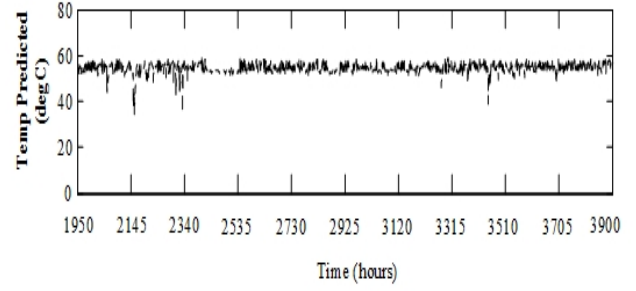
Let us first look at model prediction for the temperature of gearbox oil sump. Fig. 5 (a) illustrates the temperature of gearbox oil sump obtained from the SCADA system and Fig. 5 (b) shows the temperature of gearbox oil sump predicted by the ELM model. It can be seen from the residual signal as shown in Fig. 5 (c), the temperature of gearbox oil sump deviates from the model prediction around at hour 3315, which indicates the beginning of the gearbox fault.

Fig. 6 is also a good example to verify the effectiveness of the proposed ELM model. It shows the plots of actual SCADA data, the model predicting signal, and the residual signal of the temperature of gearbox oil exchange, respectively. Residual signal is shown in Fig. 6 (c). The temperature of gearbox oil exchange deviates from the model prediction also around at hour 3315, which happens at the same time as the temperature of gearbox oil sump. This means the gearbox fault of the wind turbine can be detected effectively.

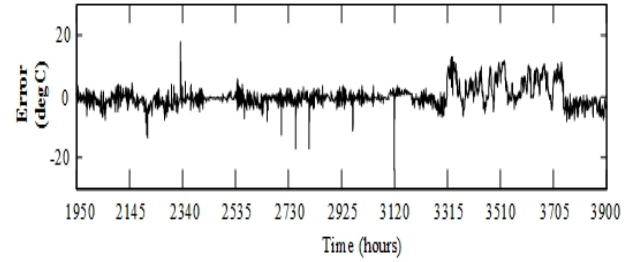
Apart from the detection of gearbox fault, the ELM model is also used to detect the generator winding faults. As can be seen from Fig. 7, the temperature of generator winding starts abnormal at around hour 3120.



a. Actual gearbox oil sump temperature from SCADA system

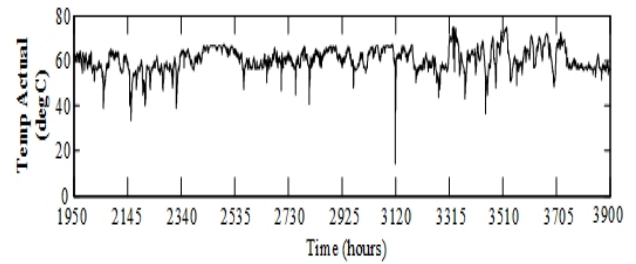


b. Model output of the gearbox oil sump temperature

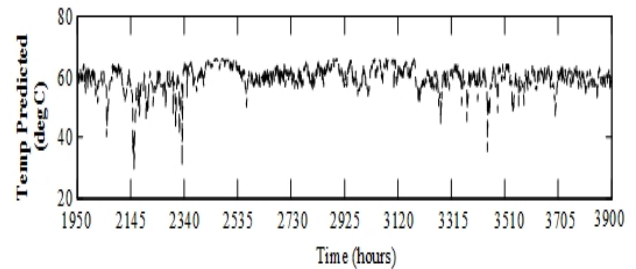


c. Residual signal of gearbox oil sump temperature between SCADA data and model prediction

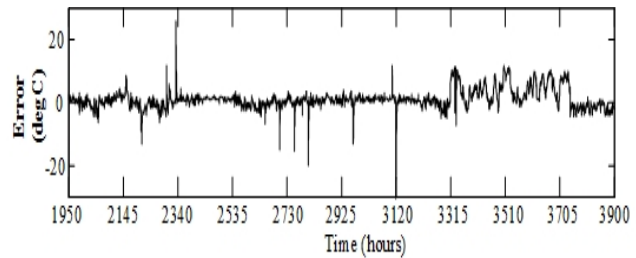
Fig. 5 ELM model response compared to SCADA data for gearbox oil sump temperature



a. Actual gearbox oil exchange temperature from SCADA system



b. Model output of the gearbox oil exchange temperature



c. Residual signal of gearbox oil exchange temperature between SCADA data and model prediction

Fig. 6 ELM model response compared to SCADA data for gearbox oil exchange temperature

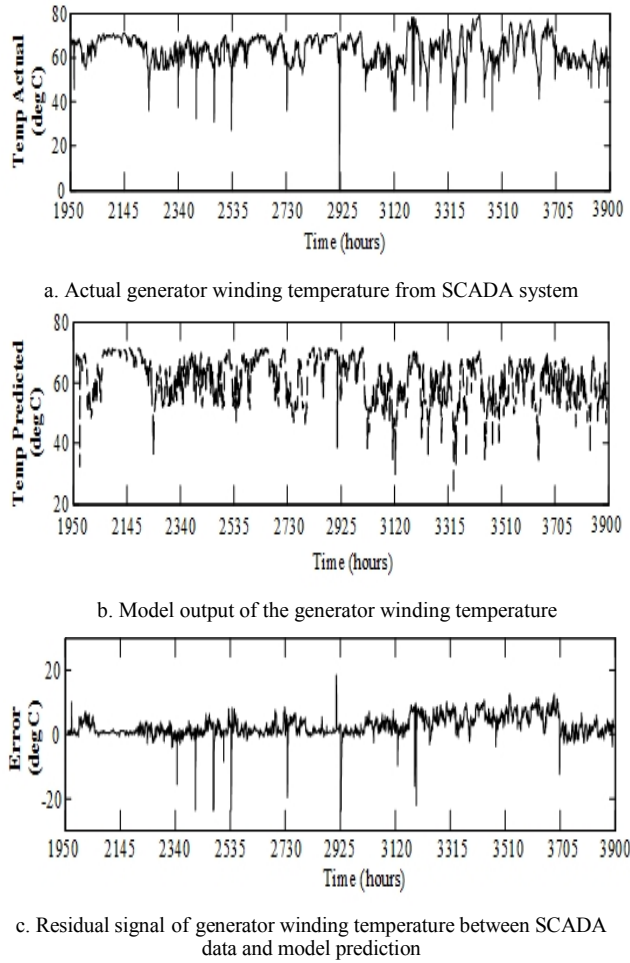


Fig. 7 ELM model response compared to SCADA data for generator winding temperature

The ELM algorithm was implemented on an ordinary PC platform equipped with a Xeon E3-1271 v3 3.6GHz CPU and 16GB RAM. The execution time was compared with the traditional BP neural network. As can be seen from Table 1, the ELM algorithm significantly outperform the BP neural network in terms of the computation time required.

Table 1 Performance comparison between the ELM and the BP method

Type	Algorithm	Time(s)
Gearbox oil sump temperature	ELM	0.161s
Gearbox oil exchange temperature	ELM	0.158s
Generator winding temperature	ELM	0.165s
Gearbox oil sump temperature	BP	23.376s
Gearbox oil exchange temperature	BP	26.14s
Generator winding temperature	BP	25.384s

IV. CONCLUSION

This paper has presented ELM algorithm that has been used in condition monitoring of wind turbines. SCADA data obtained from an operational wind farm, including the temperatures of gearbox oil exchange, gearbox oil sump and generator winding t, are used to verify the effectiveness of the proposed method. Models developed from SCADA data have been used to identify faults in gearbox and generator winding in the turbines. The results have shown that differences between actual output signals and the model predicting signals are caused by a gearbox fault and a generator winding fault. Consequently, the proposed method can provide an early warning of the impending component failure.

It is worth emphasising that the SCADA data used in the paper are mostly indicative of normal operations of the wind turbines and contain less fault information; therefore static models are only used in the paper. Dynamic models will be considered in our future work to investigate the effect of more values of the past inputs on the model output. Future work will also focus upon the development of the early warning system employing online sequential ELM that can predict faults of the wind turbines in real-time. In addition, the research will consider using the experimental platform in our lab to further validate the models in real-time implementation.

ACKNOWLEDGMENT

Permission to use the SCADA data obtained from Wind Prospect Ltd. is gratefully acknowledged.

REFERENCES

- [1] B. Snyder, M. J. Kaiser. A comparison of offshore wind power development in Europe and the US: Patterns and drivers of development. *Applied Energy*, vol. 86, no. 10, pp. 1845-1856, 2009.
- [2] C. A. Walford. *Wind Turbine Reliability: Understanding and Minimizing Wind Turbine Operation and Maintenance Costs*, Technical Report SAND 2006-1100, Sandia National Laboratories, USA, 2006.
- [3] G. J. W. van Bussel, C. Schontag. Operation and maintenance aspects of large offshore wind farms. In *Proceedings of European Wind Energy Conference*, Dublin, Ireland, 1997.
- [4] Communication Technologies Inc. *Supervisory control and data acquisition systems*. USA: National Communications System; 2004.
- [5] Dvorak D, Kuipers B. Model-based monitoring of dynamic systems. In: *International Joint Conference on Artificial Intelligence*, vol. 2; 1989. pp. 1238-43.
- [6] Schlechtingen M, Ferreira Santos I. Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mech Syst Signal Process* 2011;25:1849-75
- [7] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489-501, Dec. 2006.
- [8] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 946-955, Aug. 2008.
- [9] G. B. Huang, X. J. Ding, and H. M. Zhou, "Optimization method based extreme learning machine for classification," *Neurocomputing*, vol. 74, no. 1-3, pp. 155-163, Dec. 2010.
- [10] G. Karniotakis, G. S. Stavrakakis, and E. F. Nogaret, "Wind power forecasting using advanced neural network models," *IEEE Trans. Energy Convers.*, vol. 11, no. 4, pp. 762-767, Dec. 1996.

- [11] T. G. Barbounis, J. B. Theocharis, M. C. Alexiadis, and P. S. Dokopoulos, "Long-term wind speed and power forecasting using local recurrent neural network models," *IEEE Trans. Energy Convers.*, vol. 21, no. 1, pp. 273–284, Mar. 2006.
- [12] K. Bhaskar and S. N. Singh, "AWNN-assisted wind power forecasting using feed-forward neural network," *IEEE Trans. Sustain. Energy*, vol. 3, no. 2, pp. 306–315, Apr. 2012.

Detection and Diagnosis of Motor Stator Faults using Electric Signals from Variable Speed Drives

Abdulkarim Shaeboub¹, Samieh Abusaad¹, Niaoqing Hu², Fengshou Gu¹ and Andrew D. Ball¹

¹Centre of Efficiency and Performance Engineering, University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK

²College of Mechatronics and Automation, National University of Defense Technology, Changsha, 410073, China

E-mail: Abdulkarim.Shaeboub@hud.ac.uk

Abstract—Motor current signature analysis has been investigated widely for diagnosing faults of induction motors. However, most of these studies are based on open loop drives. This paper examines the performance of diagnosing motor stator faults under both open and closed loop operation modes. It examines the effectiveness of conventional diagnosis features in both motor current and voltage signals using spectrum analysis. Evaluation results show that the stator fault causes an increase in the sideband amplitude of motor current signature only when the motor is under the open loop control. However, the increase in sidebands can be observed in both the current and voltage signals under the sensorless control mode, showing that it is more promising in diagnosing the stator faults under the sensorless control operation.

Keywords- Induction motor; Stator faults; Variable speed drive; Motor current and voltage signatures analysis.

I. INTRODUCTION

Induction motors are commonly mentioned as the workhorse of industries, mainly because of its simple yet powerful architectural construction, ergonomically adaptable structure, rugged and highly robust and offering high value of reliability. However, they are prone to various faults related to its functionalities and operational environments. Such faults can cause not only the loss of production, but also even catastrophic incidents and additional costs. Therefore, efficient and effective condition monitoring techniques are actively studied to detect the faults at early stage in order to prevent any major failures on motors [1, 2].

Of many different techniques in developing, motor current signature analysis (MCSA) has been found more effective and efficient in monitoring different motor faults including air-gap eccentricity, broken rotor bar and turn to turn fault in the stator. It is centered on using popular frequency analysis methods to diagnose current harmonics and sidebands at such frequencies that are uniquely identifying the features of relative faults. Moreover, it does not require any additional systems for measurements[3, 4] and can be implemented remotely at low investment.

Studies show that 35–40% of induction motor breakdowns are because of stator winding breakages [5, 6]. This has motivated more works on diagnosing this type of faults. Sharifi and Ebrahimi [7] developed a method for the diagnosis of inter-turn short circuits faults in the stator

windings of induction motors. The technique is based on MCSA and utilizes three phase current spectra to overcome the problem of supply voltage unbalance. Adaptive Neuro-fuzzy Inference system (ANFIS) was explored to diagnose stator turn-to-turn and stator voltage unbalance faults in an induction motor [8]. Another novel technique is proposed for the diagnosis of inter-turn short circuit fault in induction motor by [9] based on the analysis of external magnetic field in the surrounding of the machine.. Flux and vibration analysis for the detection of stator winding faults in induction motor was presented [10]. Park-Hilbert (P-H) was introduced to diagnose stator faults in induction motors using grouping between the Hilbert transform and the Extended Park's Vector methodology [11]. Spectral analysis techniques also applicable under variable speed drive condition have been proposed in [12].

Less work have been found to explore the diagnosis performance of voltage and motor current signals upon motors with variable speed drives (VSDs) which are increasingly used in industry for obtaining better dynamic response, higher efficiency and lower energy consumption. However, VSD systems can induce strong noise to voltage and current measurements. Fault detection and diagnostics for such systems have been gaining more attentions for many years at the Centre for Efficiency and Performance Engineering (CEPE), the University of Huddersfield. Djoni Ashari et al [4] discussed a method for the diagnosis of broken rotor bar based on the analysis of signals from a variable speed drive. Mark Lane et al [13] also used signals from a variable speed drive but for detecting unbalanced motor windings of an Induction Motor. Ahmed Alwodi et al [14] provides the details of applying a modulation signal bispectrum (MSB) analysis to current signals to diagnose and enhance feature components for the detection and diagnosis of the stator faults without use close loop control modes. Although these works represents key progress in the direction, the effect of close loop VSD systems on the power supply parameters, i.e. both the current and voltage has not examined in the case of stator faults. Voltage signature analysis investigated with respect to the bandwidth variations for induction motor stator faults. Although, there have been some very strong arguments and literatures about the effectiveness of this technique with respect to closed-loop but when it comes to open-loop, there aren't any significant results that can be used in order to analyze the faults [15, 16].

This paper presents comparative results between the performance of current and voltage spectra for detecting stator faults with different degrees of severities and under both the open loop and sensorless control (close) modes. The results are obtained from common spectrum analysis applied to signals from on a laboratory experimental setup operating under different loads.

II. FAULT FEATURES AND EFFECTS

Stator windings may have different types of faults. However, there are mainly four typical stator faults: turn-to-turn fault, Phase-to-Phase short fault, Phase-to-Earth short fault and open circuit coil fault [17]. These are asymmetric faults, which are typically associated to insulation failures, caused by various reasons such as; poor connection, overloading and overheating, Fig. 1 shows these main four faults. An open circuit in the stator winding will affect the distribution of the stator MMF in the air gap. Early detection and mitigation of such faults can bring a great value to induction motor condition monitoring. [18].

Studies [19, 20] showed that stator faults generate unbalanced flux waveforms that cause the motor to draw asymmetrical phase currents and unbalanced air gap flux. Additionally, distortion caused by the fault can be attributed to the changes in the air gap flux caused by the resistive and inductive change across the stator windings. Due to the fact that the rotor magnetic field includes harmonics related to the rotor slots, this field induces frequency components in the stator windings which modulate with supply frequency. Hence when fault occurs, harmonics around the base frequency is generated and modulated by rotor slots harmonics. The feature frequency f_{sf} for the stator fault is identified according to [14, 21]:

$$f_{sf} = f_s \left[1 \pm mN_b \left[\frac{1-s}{p} \right] \right] \quad (1)$$

where f_s denotes the frequency of supply, f_r is the speed of rotor, N_b is the number of rotor bars, p the number of pole-pairs, s is the motor per unit slip, and $m = 1, 2, 3 \dots$ is the harmonic order.

As the rotor frequency f_r is calculated by

$$f_r = \frac{1-s}{p} f_s \quad (2)$$

Equation (1) therefore can be rewritten as:

$$f_{sf} = f_s \pm mN_b f_r \quad (3)$$

The per unit slip s is calculated as follows:

$$s = \frac{f_{slip}}{f_{sync}} = \frac{f_s/p - f_r}{f_s/p} \quad (4)$$

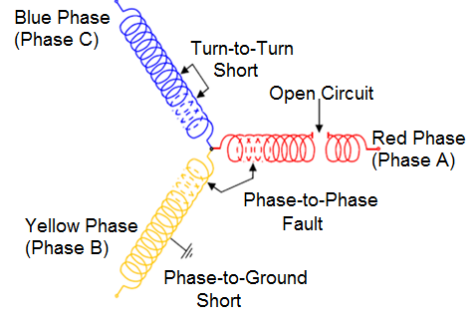


Figure 1. Different three-phase stator winding faults.

Equations (1-3) show that stator faults frequency components are a function of slip and number of rotor slots. As the slip varies with load, the feature frequency will be different under different motor loads.

III. EFFECT OF THE FAULT ON THE VSD FED MOTORS:

Different speed control schemes are offered in the market. Options vary depending on the requirements of each application. For instance, the basic V/Hz (open-loop) scheme which applied when precise speed control or/and full torque at high speeds is not required; while on the other hand closed-loop schemes with speed feedback measuring devices provide accurate speed regulation, whereas vector control schemes which allow for optimum dynamic performance when necessary for some applications, this can be with or without speed measuring, (sensorless), feedback devices. In this paper the open loop (V/Hz) and sensorless drives are considered as they are very commonly used in industries. However, closed loop and field oriented control, either with speed feedback sensors or sensorless, are all based on feedback regulation and stator faults can have nearly the same effects on them [22].

A. Influence of the open loop (V/Hz)

The V/Hz induction motors drives maintain the air-gap flux constant by keeping the ratio V/f_s constant. This aims at maintaining the torque at a value given by the slip frequency, $f_s - f_r$ for any supply frequency value f_s . Induction motor drives normally employ pulse width modulation inverters that vary the magnitude and frequency of the output voltages. The drive continually feeds the motor with a constant V/Hz ratio by the inverter output keeping the motor air-gap flux constant [22, 23]. Fig. 2 shows a simplified structure of an open-loop induction motor speed drive [23]. The air gap flux is held constant based on the following formula [22]

$$\varphi_m = L_m |i_s + i_r| = \frac{v_s}{\omega_s} \quad (5)$$

Where: φ_m is the air gap flux, L_m is the mutual inductance, i_s and i_r are the stator and rotor currents respectively, v_s is the stator voltage and $\omega_s = 2\pi f_s$ the supply angular speed.

By holding a constant air gap flux as in (5), the developed electromagnetic torque remains the same for a

given slip value. That is the electromagnetic torque is mainly depending on slip frequency and stator flux. However it is worth mentioning that the required terminal voltage is a function of both frequency, as in (5), and the load. The developed electromagnetic torque T_{em} is defined as [22]:

$$T_{em} = \frac{3p}{2(\omega_s - \omega_r)} \left(\frac{v_s^2}{2(\frac{\omega_b R_r}{\omega_s - \omega_r})^2 - x_{lr}^2} \right) R_r \quad (6)$$

Where ω_b and ω_r is the motor base angular frequency and the rotor angular frequency respectively, and R_r and x_{lr} rotor resistance and impedance respectively.

When a fault occurs in the stator windings, the air gap flux distribution is not balanced anymore.

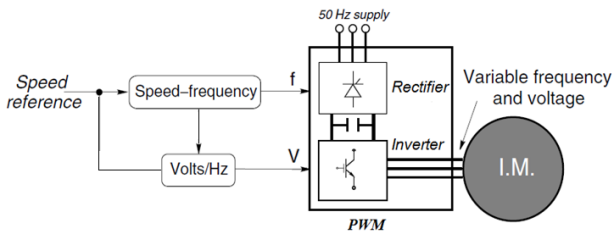


Figure 2. A simplified structure of the V/Hz drive .

This unbalanced flux oscillates the induced rotor current around the fault frequency component defined in (3). Hence this oscillation will also modulated by the electromagnetic torque as shown in (5) and (6). When the drive is in open loop operational mode, there is no feedback to the drive and such torque oscillations will modulate the motor current. The drive only keeps the V/Hz ratio constant based on the reference speed, as shown in Fig. 2. To conclude, in the case of the open loop mode, the drive keeps the voltage fixed base on the frequency requirements. The air flux will be oscillated due to the fault and hence the torque is also oscillated causing changes in the slip frequency as clear from equation (4) and (5). Oscillations in the front side of the system are not compensated and hence the feature frequencies most likely appear in the current signals, rather than the voltage. Also, the slip frequency will be influenced as it is sensitive to the changes in the electromagnetic torque, as depicted in (5) and (6). Changes in the slip can be more dominant as the load increases.

B. Influence of the closed loop drive

Many different schemes are used for this drive mode. However, the general arrangements are the same and typically as shown in Fig. 3. [23]. In this drive mode a feedback loop is involved for providing better speed regulation and enhanced dynamic response. The output voltage is separately regulated utilizing the knowledge of phase angle, while the frequency is controlled by switching in time of the inverter [22].

Additionally, closed loop mode drives include two parallel branches. The first is constructed from at least four cascade PI control loops. The outer is speed PI

control loop which compares the reference speed with the feedback speed and generates the speed error.

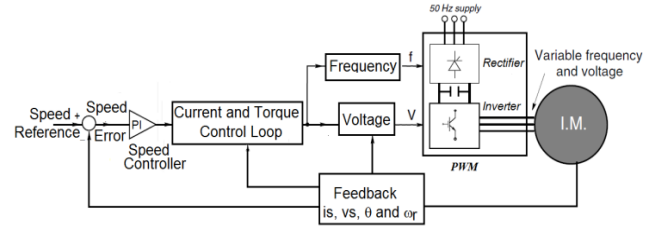


Figure 3. A general structure of a closed loop variable speed drive

The speed error sets the reference value of the inner control loops. The inner control loops are two PI control loops in series, i.e. the current control loop that sets the reference torque signal based on the difference between the actual current and the reference from the speed loop; and the torque control loop that sets the reference value for the voltage output based on the difference between the feedback torque and reference from the current loop. The fourth is the voltage control loop that gives the control action to the PWM to feed the motor with the required voltages and frequencies. The other branch is the field PI control loop which keeps the field at the rated value when the speed is less than the base motor speed. When the speed is at or over the rated speed, the flux is reduced by $1/\omega_r$.

For stability reasons, the speed control loop is normally tuned with a bandwidth that is three times less than the inner current and torque loops. This means that the current loop is at least three times faster than the speed loop and the torque loop is also three time higher than the current loop [24].

Equation (3) shows that stator faults frequency components are modulated by the rotor slits' components and slip frequency. The speed control loop is not influenced by this frequency as it is out of its bandwidth. On the other hand, this frequency can be within the bandwidth of the current and torque control loops and will be treated by their controllers as a disturbances. Hence the torque reference will include such frequency component which will be directed to the voltage regulators as an output voltage to the motor supply. It is also worth mentioning that when the fault is not big enough, the drive regulator actions and the noise from the PWM switches masks such fault features in the current signal make it more difficult to be detected in the current signal. Therefore, in closed loop systems the voltage signals are likely to be more sensitive to stator faults than the current.

IV. TEST FACILITY

To evaluate the analysis in previous section, an experimental study was conducted based on a three-phase induction motor with rated output power of 4 kW at speed 1420 rpm (two-pole pairs), as shown in Fig.4. A digital variable speed drive is employed to control the motor speed. The controller can be set in either open loop (V/Hz) or sensorless modes which is very common in industry applications. When sensorless mode is used, the drive estimates the system speed based on the built-in Model Reference Adaptive system (MRAS). The induction motor

is coupled with a loading DC generator using a flexible spider coupling. The DC load generator is controlled by a DC variable drive that varies the armature current in the DC load generator to provide the required load to the AC motor. The operating speeds and loads are set by the operator via a touch screen on the control panel.

A power supply measurement box is employed to measure the AC voltages, currents and power using Hall effect voltage and current transducers and a universal power cell. During the experimental work all the data was acquired using a YE6232B high speed data acquisition system. This system has 16 channels; each channel is equipped with a 24 bit analogue-digital converter. The maximum sampling frequency is 96 kHz. A speed encoder is mounted on the motor shaft that produces 100 pulses per revolution for measuring the motor speed.

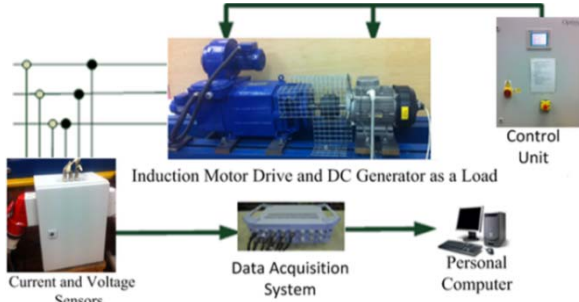


Figure 4. A photograph of the test rig facility

In this work, current and voltage signals were collected for three diverse winding formations; healthy motor, one coil removal, and two coil removal, with equal load increments: 0%, 20%, 40%, 60%, and 80% load, which lets the investigative performance to be inspected at variable loads and avoid any probable damages of the test system when faults are simulated at the full load. As illustrated in Fig. 5, there are three concurrent coils in individual phase and by rearranging the terminals from the phase terminals, it permits rewriting the three coils in phase B at three multiple ways, which are: supply to B1-B2-B3 in a healthy case, moreover, supply B1-B2 in a case of one coil removal, which illustrates a smaller fault, and supply to only B1 for two conductor removal that presents a bigger fault. Clearly, removals of these coils could be a simulation of an asymmetric stator that is going to enhance the motor equivalent impedance and therefore damage its overall performance.

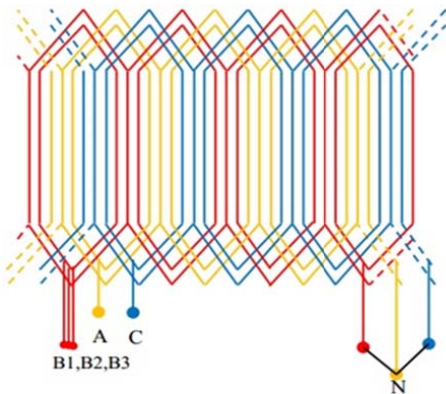


Figure 5. Achematic of stator fault simulation in phase B

V. RESULTS AND DISCUSSION

A. Stator Fault During Open Loop Control with Full Speed and Variable Loads.

Fig. 6 shows one sideband current spectrum of stator current which was obtained by applying the fast Fourier transform (FFT) to signals measured under healthy and the two faulty cases at full speed and under different loads with respect to open loop (O.P.) control mode. It can be seen that there is a small visible sideband for stator faults under 0% motor load since the slip is small. However, the amplitude of sideband increases as the load and fault severity increase. This shows that the amplitude of the sideband at the frequency component defined in (3) is sensitive to load changes, in other words slip frequency, and fault severity and the fault can be best detected under higher load. Additionally, Fig. 6 indicates that the fault enlarges the slip frequency. This is can be explained as the stator faults resulted in additional oscillations in the magnetic flux that transferred into the electromagnetic torque, which in turn has a direct effect on the slip frequency.

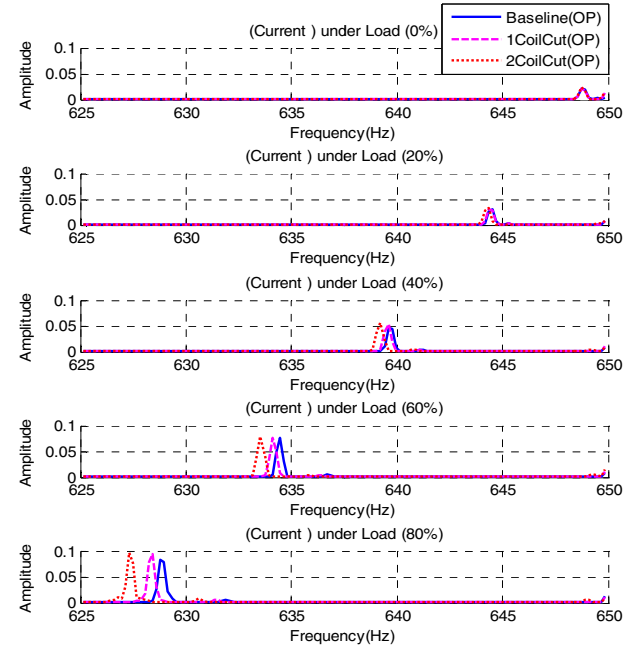


Figure 6. Phase current spectra under O.P. and different loads.

On the other hand, there are no changes found that can be used to analyze the faults by using motor voltage signature analysis with respect to open loop control mode. This is due to the fact that in the open loop mode the drive feeds the motor with a supply of a constant V/Hz ratio regardless the changes in the motor electromagnetic torque and current. Oscillations in the motor torque due to the fault are not seen by the drive as there is no feedback. The V/Hz ratio is kept constant even when the slip changes either due to the load or the fault. No compensation action is taken by the drive as long as the speed reference stays the same.

B. Stator Fault During Sensorless Control with Full Speed and Variable Load

Fig. 7 shows the spectrum of stator current under faulty and healthy motor conditions under loads 0%, 20%, 40%, 60% and 80% for the sensorless (S.L.) control mode. It can be seen that there are small visible sidebands for stator fault under 0% and 20% motor load since the slip is small. It is clear that the amplitude of sideband increases as the severity of the fault and load increases and the fault can be best detected under higher load. These changes exhibit more prominently when the load reaches 80% on the sensorless graph, where the controller performed less accurately in maintaining the desired speed because of the fault effects. However not much of a difference can be seen in the open loop graph when the load touches 80%.

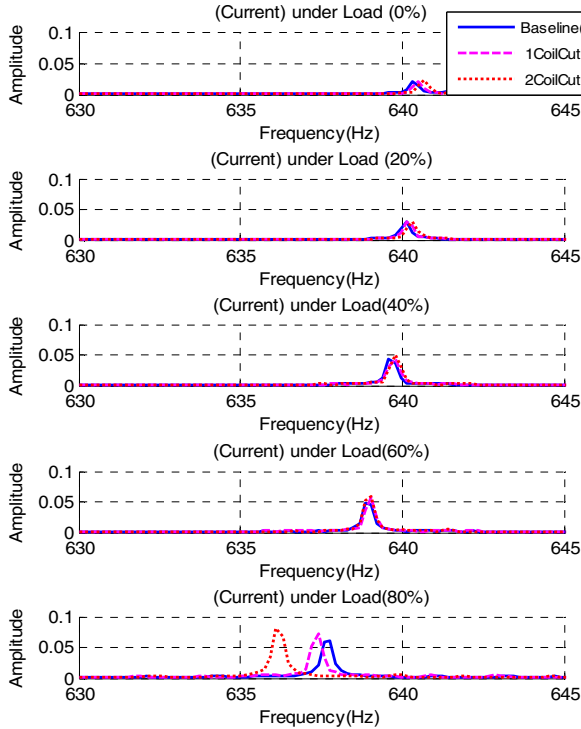


Figure 7. Current spectra under different loads for the S.L. mode

Fig. 8 depicts the output voltage spectra from the drive to the motor terminals. The sidebands of voltage increase in amplitude with load and severity of the fault. The main concern of this domain is that how effective motor voltage signature analysis (MVSA) is when it comes to analyzing and detecting faults that occur in an induction motor. Although there have been some very strong arguments and references about the effectiveness of this technique with respect to sensorless but when it comes to open-loop, there aren't any inevitable results that can be used in order to analyze the faults. Experiments and results that are presented in this paper claim for the same thing.

In particular, MVSA with respect to open-loop doesn't provide a clear value even if the load is raised to 80%. On the other hand in the Sensorless control mode gives some significant indication to the fault as soon as load reaches 40% and at 80% load the graph shows prominent increases in sidebands. This proves that sensorless technique allows more accurate and efficient outputs as compare to open loop technique.

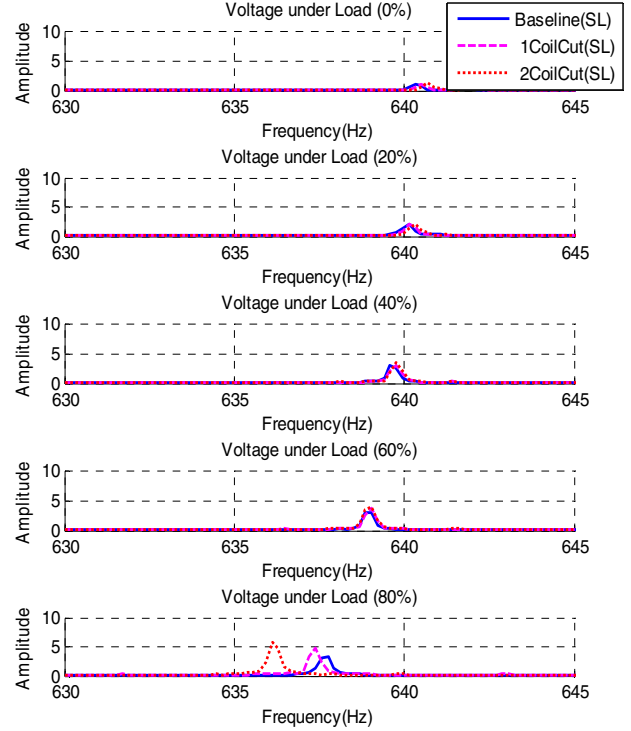


Figure 8. Voltage spectra under different loads for the S.L. mode

VI. COMPARISON BETWEEN TECHNIQUES

The comparative study of different condition monitoring techniques which include MCSA and MVSA has been made for healthy and stator fault conditions. Fig. 9 shows the diagnostic performance comparison of the current signal for different fault cases and the two control modes under the 80% of the full motor load. It is significant that during open loop operation by use MCSA the amplitude of sideband increases as the severity of the fault and load increase.

However, MVSA provides effective diagnostic features under the sensorless control mode as shown in Fig 10. In addition, the voltage spectrum demonstrates slightly better performance than the motor current spectrum because the VSD regulates the voltage to adapt changes in the electromagnetic torque caused by the fault.

VII. CONCLUSION

This paper compares the effectiveness of motor current signature analysis and voltage signature analysis for detection and diagnostics of motor stator faults under Volt/Hz and sensorless control mode. Their comparison was done on the relevant basis and on same ground. The spectrum of stator current shows that the amplitude of sideband increases with fault severity and load in both open loop and sensorless operating modes. Similarly, the sideband of voltage also increase in amplitude with the load and fault severity increase but they appear just at sensorless control operation mode. Additionally, a significant increase in slip frequency has been noted as the fault severity and load increases in both sensorless and open loop modes. In addition, it also shows that the sensorless control gives more reliable and accurate diagnosis result.

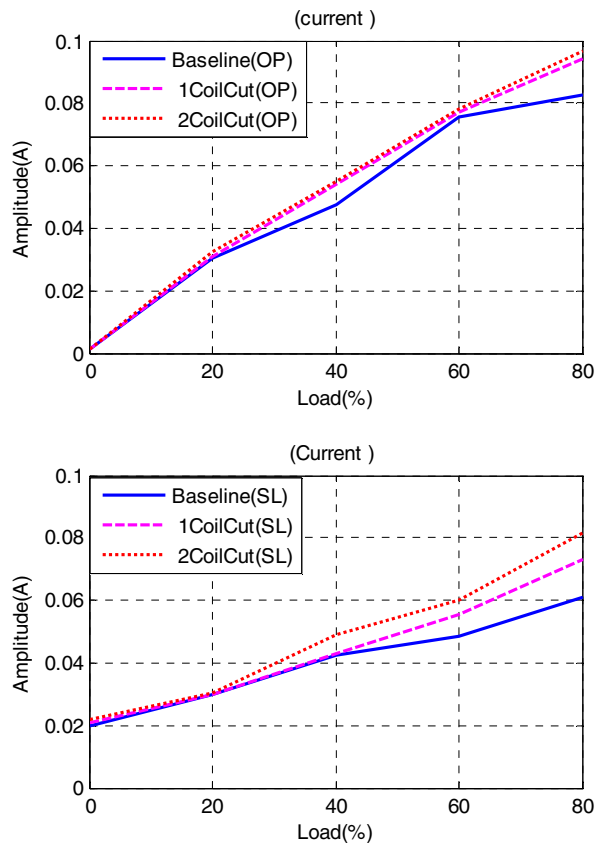


Figure 9. Current signal diagnostic performance comparison between O.P. and S.L.Modes

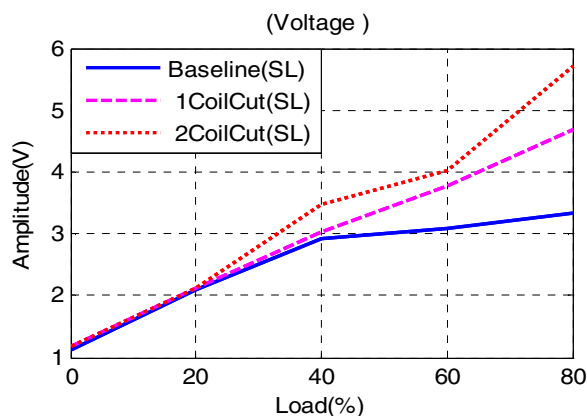


Figure 10. Voltage diagnostic results from S.L.mode.

REFERENCES

- [1] N. A. Hussein, D.Y.M., and I. M. Abdulbaqi, 3-phase Induction Motor Bearing Fault Detection and Isolation using MCSA Technique based on Neural Network Algorithm. *Int. J. Appl. Eng. Res.*, no. 5, , 2011. 6: p. 581-591.
- [2] Alwodai, A., F. Gu, and A. Ball, A Comparison of Different Techniques for Induction Motor Rotor Fault Diagnosis. 2012.
- [3] El Hachemi Benbouzid, M., A review of induction motors signature analysis as a medium for faults detection. *Industrial Electronics, IEEE Transactions on*, 2000. 47(5): p. 984-993.
- [4] Ashari, D., et al. Detection and Diagnosis of Broken Rotor Bar Based on the Analysis of Signals from a Variable Speed Drive.
- [5] Stone, G.C., A perspective on online partial discharge monitoring for assessment of the condition of rotating machine stator winding insulation. *IEEE Electrical Insulation Magazine*, 2012. 28(5): p. 8-13.
- [6] Report of Large Motor Reliability Survey of Industrial and Commercial Installations, Part I. *IEEE Transactions on Industry Applications*, 1985. IA-21(4): p. 853-864.
- [7] Sharifi, R. and M. Ebrahimi, Detection of stator winding faults in induction motors using three-phase current monitoring. *ISA transactions*, 2011. 50(1): p. 14-20.
- [8] Ahmed, S.M., et al., Diagnosis of Stator Turn-to-Turn Fault and Stator Voltage Unbalance Fault Using ANFIS. *International Journal of Electrical and Computer Engineering (IJECE)*, 2013. 3(1): p. 129-135.
- [9] CEBAN, A., et al., Diagnosis of inter-turn short circuit fault in induction machine. *Annals of the University of Craiova, Electrical Engineering series*, 2011. 35: p. 103-110.
- [10] Lamim Filho, P., R. Pederiva, and J. Brito, Detection of stator winding faults in induction machines using flux and vibration analysis. *Mechanical Systems and Signal Processing*, 2014. 42(1): p. 377-387.
- [11] Sahraoui, M., et al. A new method to detect inter-turn short-circuit in induction motors. in *Electrical Machines (ICEM), 2010 XIX International Conference on*. 2010. IEEE.
- [12] Wieser, R.S., C. Kral, and F. Pirker. The Vienna induction machine monitoring method; on the impact of the field oriented control structure on real operational behavior of a faulty machine. in *Industrial Electronics Society, 1998. IECON'98. Proceedings of the 24th Annual Conference of the IEEE. 1998. IEEE*.
- [13] Lane, M., et al., Investigation of Motor Current Signature Analysis in Detecting Unbalanced Motor Windings of an Induction Motor with Sensorless Vector Control Drive, in *Vibration Engineering and Technology of Machinery 2015*, Springer. p. 801-810.
- [14] Alwodai, A., et al. Modulation signal bispectrum analysis of motor current signals for stator fault diagnosis. in *Automation and Computing (ICAC), 2012 18th International Conference on*. 2012. IEEE.
- [15] Gritli, Y., et al. Closed-loop bandwidth impact on MVSA for rotor broken bar diagnosis in IRFOC double squirrel cage induction motor drives. in *Clean Electrical Power (ICCEP), 2013 International Conference on*. 2013. IEEE.
- [16] Bose, B.K., *Modern power electronics and AC drives*. Vol. 123. 2002: Prentice Hall USA.
- [17] Mehala, N., Condition monitoring and fault diagnosis of induction motor using motor current signature analysis, 2010, NATIONAL INSTITUTE OF TECHNOLOGY KURUKSHETRA, INDIA.
- [18] Henao, H., C. Martis, and G.-A. Capolino, An equivalent internal circuit of the induction machine for advanced spectral analysis. *Industry Applications, IEEE Transactions on*, 2004. 40(3): p. 726-734.
- [19] Nandi, S., H.A. Toliyat, and X. Li, Condition monitoring and fault diagnosis of electrical motors-a review. *Energy Conversion, IEEE Transactions on*, 2005. 20(4): p. 719-729.
- [20] Alwodai, A., et al., Inter-Turn Short Circuit Detection Based on Modulation Signal Bispectrum Analysis of Motor Current Signals, 2013, Brunel University.
- [21] Cusidó, J., et al., Signal injection as a fault detection technique. *Sensors*, 2011. 11(3): p. 3356-3380.
- [22] Ong, C.-M., *Dynamic simulation of electric machinery: using MATLAB/SIMULINK*. Vol. 5. 1998: Prentice hall PTR Upper Saddle River, NJ.
- [23] Hughes, A. and B. Drury, *Electric motors and drives: fundamentals, types and applications 2013*: Newnes.
- [24] ABB, Technical Guide No. 100, High Performance Drives-speed and torque regulation, in *High Performance Drives-speed and torque regulation*, I. ABB Industrial Systems, Editor 1996, ABB Industrial Systems, Inc.

Investigation of Motor Current Signature Analysis to Detect Motor Resistance Imbalances

M. Lane D. Ashari F. Gu A.D. Ball
Centre for Efficiency and Performance Engineering
University of Huddersfield
Huddersfield, UK
mark.lane@hud.ac.uk

Abstract— The trend to use inverter drives in industry is well established. It is desirable to monitor the condition of the motor/drive combination with the minimum of system intervention and at the same time retaining compatibility with the latest generation of AC PWM vector drives. This paper studies the effect that an increase in motor stator resistance has on the motor performance, efficiency and voltage/current characteristics during operation of a latest-generation unmodified AC PWM drive under varying speed conditions. The increased resistance is intended to simulate the onset of a failing connection between drive and motor but one that is non-critical and will remain undetected in use because the resistance increase is small and does not appear to affect the motor operation. Performance of the motor/drive combination is measured against baseline motor data for the resistance increase. Measurements are also taken following an autotune on the drive to observe the effects that motor stator resistance imbalance has on the sensorless vector control algorithms. All data collection signals are post-processed using data analysis methods developed for MATLAB. Initial results from the motor tests clearly show a difference in values measured from the motor current and voltage signals post-processed under MATLAB and the asymmetry values equally show the 0.1Ω resistance increase. The test results are presented herein and future research work is identified.

Keywords— *Efficiency, Unbalanced, Stator resistance increase, MCSA, PWM, Random switching pattern, IAS.*

I. INTRODUCTION (HEADING 1)

Motor efficiency in industrial applications is becoming more critical but some traditional motor condition monitoring techniques have been limited to constant speed applications. This series of tests simulates the motor operation over a range of speeds and with a small resistance increase in the motor stator windings that would not normally cause faults to be detected in the inverter drive system itself but could lead to a reduction in motor efficiency and unbalanced running of the inverter drive system that may result in premature failure of the motor and the drive system.

Existing research in this area has covered standard inverter drives without sensorless vector control operation and random-pattern PWM techniques.

The purpose of this research is to determine if the resistance increases can be detected for a motor that is running at variable

speeds and also the effect that a motor autotune has on performance following a resistance increase.

Although stator resistance increases are not the only cause of motor failures, the nature of the fault is such that it can remain undetected for an extended period of time because once a motor system is installed and commissioned, the drive system is unlikely to be checked for correct operation until the system has failed completely. For companies interested in maintaining the efficiency of plant and equipment there are more favourable non-intrusive methods to be used such as external vibration measurements to monitor bearings or thermal imaging techniques for overload monitoring. It is unlikely that electrical checks on motor stator resistances will be carried out due to the equipment having to be brought off-line, the requirement to disturb the motor connections, risk element in isolating connections and the potential of a future consequential failure of plant equipment following intervention.

Bin Lu et al. [1] recorded that more than 92% of North American paper mills use periodic vibration measurement techniques and reported an increasing take up of MCSA methods in industry. A. Bellini et. al [2] reported on the study of mechanical failures detected in industrial applications by the use of MCSA, but on non-inverter fed motors. Pedro Vicente et al. [3] have studied a simplified scheme for motor condition monitoring on inverter-fed systems and in particular they were used to detect inter-turn short circuit faults, eccentricity and broken rotor bars. Lucia Frosini et al. [4] investigate external monitoring methods such as stator currents and external flux leakage analysis to detect inter-turn faults. However, it is noted that inter-turn winding insulation failures develop within 30 to 60 seconds before the iron core is melted [1]. At this stage of motor failure, there is probably little that can be done in terms of rectification of the fault from the point of failure and detection to before the unit fails catastrophically.

The effects of unbalanced supply voltage feeds to an AC motor are well understood and result in the motor operating efficiency being reduced. The voltage imbalances are documented in the NEMA standards for AC induction motor performance

The purpose of this paper is to focus on detecting gradual failures in an inverter-fed motor system before total failure occurs. With early detection and subsequent intervention to correct the gradual faults, the system can be restored to the

installed condition and efficiency thus preserving optimum operation of the equipment. To this end, this research adopts an experimental study based on the latest motor drive technology with manually introduced unbalanced conditions.

II. TEST FACILITIES AND FAULT SIMULATION

A. Test Rig

A test rig consisting of a PWM inverter with 3 kHz switching frequency and a carrier frequency selected by a random pattern generator enabled by default was used. For the research to be valid with future drive systems, it is important that valid test results can be obtained from equipment that utilizes the very latest motor control technology and with these advanced switching techniques enabled. A standard inverter-rated 690 V AC motor was used, connected in Delta to the 415 V output inverter.

A shunt-wound DC motor is used to apply different loading to the AC motor drive system. The speed, load and test duration are all programmable and repeatable. The DC motor regenerates to the mains supply through a four-quadrant two phase DC drive.

B. Healthy operating conditions

In order that a meaningful and consistent baseline data set was available, which the simulated faults would later be compared to, healthy baseline data was measured in three separate tests, each with the same operating data applied. The test run data detailing test run speeds, duration of test and AC motor loading is given Table I. All test result plots use the x-axis as the "time domain", with each total period of 4 units corresponding to one test run. A complete test run comprises a total of three sets of 4 tests contiguously, making 12 units in all. A sampling rate of 96kHz with a 40 second sample time was used for each test run. The table below details the time values and load settings applied at each set:

TABLE I. TEST RUN DATA

Test	Speed (RPM)	Test duration(s)	Load (% of motor FLC)
1	367.5 (25%)	40	100
2	735 (50%)	40	100
3	1102.5	40	100
4	1470	40	100

A diagrammatical representation of the total test cycle is shown in figure 1.

So that tests can be compared between drive operating modes, the drive would be run for each of the three test cycles as follows:

- Healthy motor non-autotuned
- Healthy motor autotuned
- Motor with simulated faults

The drive operating mode for each of the above test runs would be in sensorless vector control.

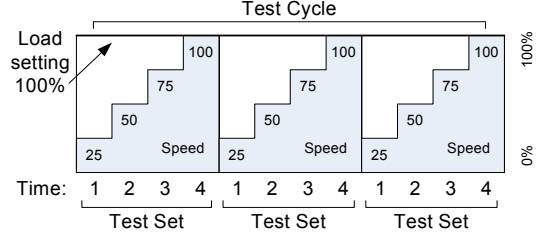


Fig. 1 Test cycle for variable speed test

III. MOTOR IMBALANCE SIMULATION TEST RESULTS

This section details and compares the test results obtained from each of the baseline and simulated faulty data sets.

A sampling rate of 96kHz was used for all tests across all data channels. Random pattern PWM switching was enabled for these test results, because this was considered the worst-case condition for trying to extract data from motor current signals due to the random noise floor.

Motor speed is measured from an encoder mounted on the rear of the AC motor. This encoder is purely for speed measurement purposes and is not connected to the inverter drive.

A. Voltage, Current and Speed Plots

Figure 2 presents a plot of motor voltage measured at each set motor speed with 100% load setting applied on the test rig it can be observed that there is a difference in motor voltage measured from the baseline data compared to the increase in resistance and the autotuned motor. This becomes more apparent when the scale is increased

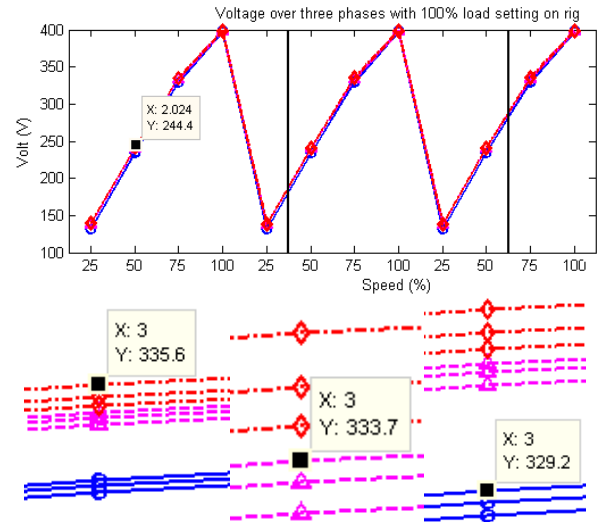


Fig. 2 Plot of voltage over 3 phases with variable speed and 100% load setting.

The measured voltage difference between the measured voltage with under fault conditions compared to the baseline data at 75% speed is 6.4V.

Figure 3 details the test results for motor voltage, speed and current measurements for the same tests.

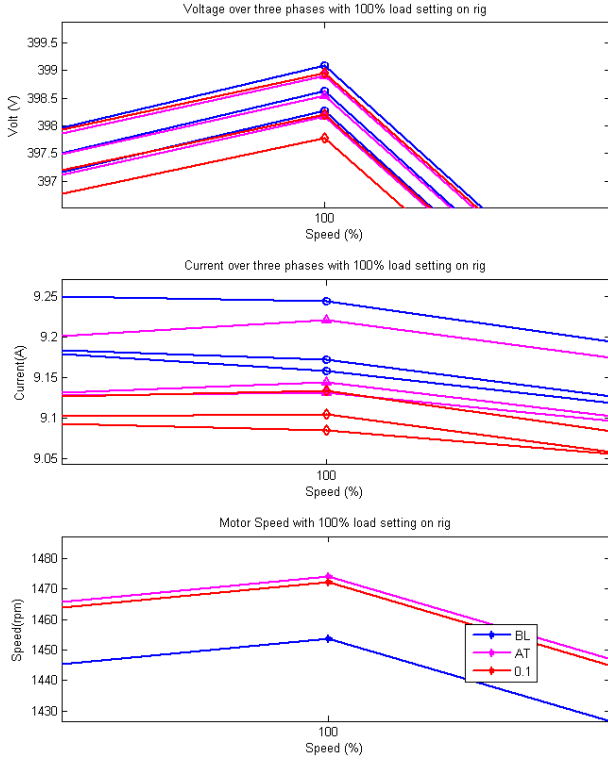


Fig. 3 Plot of motor voltage, current and speed with variable speed and 100% load setting. BL = Baseline (Healthy Motor); AT = Baseline Motor post-autotune; 0.1 = Stator Fault Resistance of 0.1Ω introduced

At the first 100% speed setpoint, the three separate tests show an increase in motor voltage after each test run. This can be explained by an increase in the motor winding resistance as the motor temperature increases. To provide more consistent tests it may be necessary to run the motor at full load for longer periods so that the motor is brought up to operating temperature before the tests commence. Even with this taken into consideration, there is still a consistent set of results obtained because the effect of the 0.1Ω resistance increase can be observed at each 100% speed setpoint.

A reduction in motor current is seen markedly from baseline, through to the resistance increase, which is to be expected.

The motor speed is increased after autotune and resistance increase compared to the baseline test, but the motor speed with 0.1Ω is slightly reduced because the higher resistance would correspond to a reduction in available motor torque and also affect the motor model calculated by the drive.

B. Motor Current Plots

If the current in three phases is observed over the three consecutive test runs, the gradual effect of an increase in motor resistance during the tests is clear and Figure 4 presents the data plot for this.

The relationship of a reduction in motor current during subsequent tests can possibly be explained by the motor

resistance gradually increasing as motor temperature rises during the course of 15 minutes of continuous running.

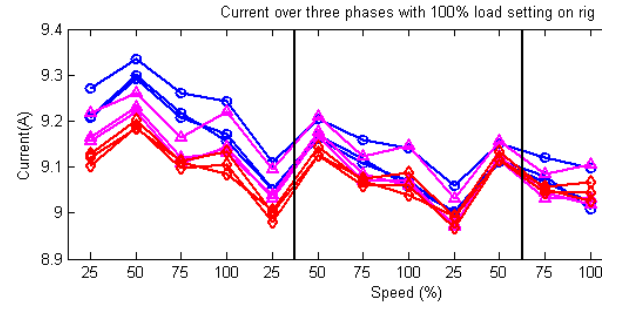


Fig. 4 Plot of motor current in consecutive tests

It is important to accurately measure the motor temperature so that the tests can be started at a consistent motor temperature and some uniformity of measurement is obtained. The temperature measurements are described in section 3.4 of this report.

C. Motor current and voltage asymmetry

The comparison of motor voltage and current imbalances on each phase gives a more meaningful indication of the actual fault occurring. In Figure 5 it can be seen that there is a marked difference between healthy and induced fault motor voltages (upper graph) and current (lower graph).

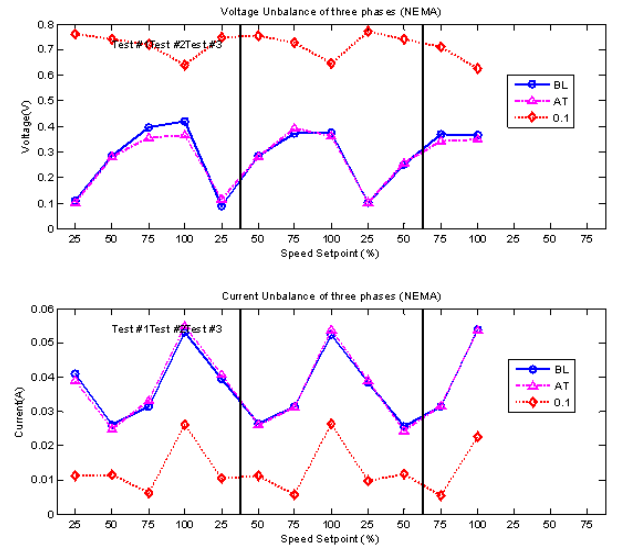


Fig. 5 Plot of motor current in consecutive tests

By calculating the difference between motor voltages and currents, the effect that a gradual increase in motor resistance that naturally occurs during the test has on the displayed test results is reduced and a consistency is restored to the readings. The readings in Figure 5 show that there is little difference between the baseline and autotuned test runs. The major difference occurs when the resistance is introduced. This is a positive step for the test results – over three tests, consistent results are achieved.

D. Motor temperature

Figure 6 shows the plot of motor temperature during each of the three tests. As can be seen from the baseline data, the motor was not fully warmed-up before the tests commenced. After the baseline tests, an autotune was performed on the motor. This would have caused further heating of the motor before the post-autotune tests were run. A more prolonged period of running at higher load to heat the motor up to normal operating temperature would be beneficial, as each test could be started at the same temperature. The immediate temperature increase for the 0.1 Ω resistance tests cannot be correlated to the increased resistance. The motor thermal time constant is too large for this. The motor temperature will have to be measured again and preferably with a separate temperature probe.

E. Efficiency measurements

The motor efficiency calculations are presented in Figure 7. The motor efficiency calculations do not present any clear evidence that efficiency is reduced with the 0.1 Ω resistance added. From baseline to autotuned data, the results back up what is to be expected, that is efficiency is increased as a result of the motor being autotuned. However, after autotuning and adding the 0.1 Ω resistance, there was no reduction in efficiency shown.

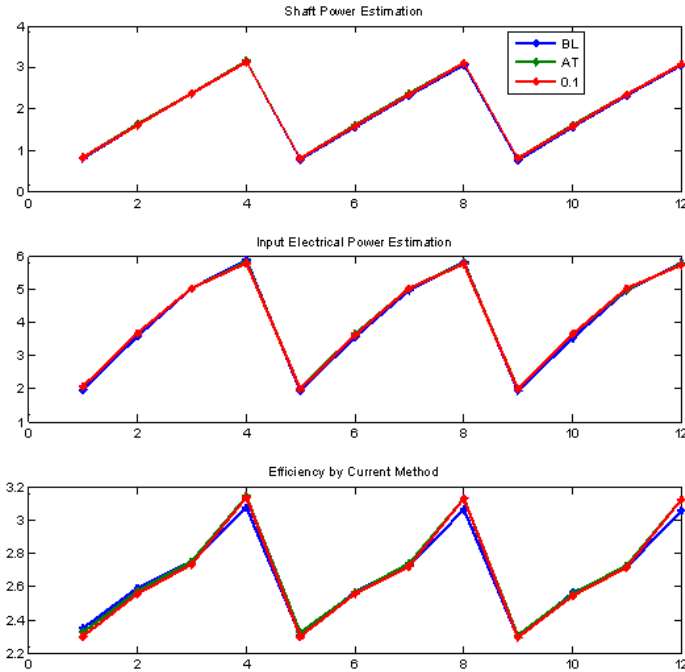


Fig. 7 Motor efficiency calculations

F. Motor parameters

The motor parameters pre and post-autotune are shown in Table 2. The data is taken from the Parker SSD ConfigED Lite software package.

From Table 2, it can be seen that the increase in stator resistance post-autotune is 0.823 Ω . Prior to running the drive on the healthy motor, the drive was previously autotuned on the same design of a healthy 4.0kW motor from the same manufacturer.

TABLE 2. AUTOTUNE DATA

Parameter	Pre-Autotune	Post-Autotune	Difference
Magnetising current (A)	5.25	5.13	0.12
Stator resistance (\bullet)	0.949	1.772	0.823
Leakage inductance (mH)	17.15	18.99	1.84

The difference in stator resistance can therefore be attributed to manufacturing variances. The difference in value is significantly more than the resistance increase applied of 0.1 Ω , so some confidence is gained that the unbalanced resistance applied is of a small enough value and less than the manufacturing variances between two motors of the same type and manufacturer.

IV. CONCLUSIONS

Initial results from the motor tests clearly show a difference in values measured from the motor current and voltage signals post-processed under MATLAB and the asymmetry values equally show the 0.1 Ω resistance increase.

The motor current plots show a more marked difference than the voltage measurements when comparing the healthy and faulty motor data. It was observed that a re-autotune from one healthy motor to another of the same batch did not affect the test results when compared to the imbalanced test results. However, this is to be expected since the increase is balanced across all motor phases, whereas the motor current difference measurements are measured from phase-to-phase.

These initial results are positive, indicating that motor imbalances can be observed without the need for spectral analysis of the motor current signals.

The test results will be extended to cover an autotune following the imbalanced motor resistance to determine if the inverter can compensate for stator resistance imbalances.

REFERENCES

- [1] Bin Lu, David B. Durocher, Peter Stemper (2008) Online And Nonintrusive Continuous Motor Energy And Condition Monitoring In Process Industries. Pulp and Paper Industry Technical Conference, 2008. PPIC 2008. Conference Record of 2008 54th Annual
- [2] Bellini, A; Filippetti, F; Franceschini, G; Tassoni, C; Passaglia, R; Saottini, M; Giovannini, M (2003) Mechanical failures detection by means of induction machine current analysis: a case history. 4th IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics and Drives
- [3] Pedro Vicente, Jover Rodn'iguez, Marian Negrea, Antero Arkkio (2007) A simplified scheme for induction motor condition monitoring. Laboratory of Electromechanics, Department of Electrical and Communication Engineering, Finland.
- [4] Austin HB (1999) The impact that voltage and frequency variations have on AC induction motor performance and life in accordance with NEMA MG-1 standards. In: Conference record of 1999 annual pulp and paper industry technical conference

Survey of Greener Ignition and Combustion Systems for Internal Combustion Engines

Wuqiao Luo, Yun Li

School of Engineering, University of Glasgow,
Glasgow G12 8LT, U.K.

w.luo.1@research.gla.ac.uk; Yun.Li@glasgow.ac.uk

Zhong Tian, Bo Gao, Ling Tong, Houjun Wang,
Baoqing Zeng

University of Electronic Science and Technology of
China, Chengdu, China

Abstract— The spark and compression ignition principles of petrol and diesel internal combustion engines (ICEs) have not advanced for a century. These do not lead to complete combustion and hence result in high exhaust emission and low energy efficiency. This paper presents a comprehensive survey on the attempts and developments of greener ignition and combustion systems for ICEs and points out that homogeneous charge microwave ignition (HCMI) holds the key to a perfect solution. Increasing the ignition volume has become a trend in research for high-performance, lean-burn and low-emission petrol engines. It started with increasing ignition points by adopting multi-point spark and lasers. The ignition volume is future increased by high energy and long endurance ignition as method of zone-based ignition. Production of transient plasma in ignition stage is the key point because combustion performance and flame front speed is related to volume of transient plasma. But the volume of ignition is still limited for high ignition energy requirement. The volume-based ignition methods, like HCMI, with bigger ignition volume which leads to better efficiency and low emission are assessed and compared. Many tests on physical engines have proved that HCMI offers significant performance, but the problem lies in a production-oriented design. Virtual prototyping through Computer-Automated Design would help in this regard and could lead to novel processes for Industry 4.0.

Keywords- internal combustion engine; ignition system; HCCI; HCMI; Industry 4.0

I. INTRODUCTION

With the rapid growth of automotive vehicles on the road worldwide [1], exhaust emissions from internal combustion engines (ICEs) have raised urgent concerns. The search has been on for new technologies to reduce the emissions by addressing this problem at root, i.e., by increasing the fuel efficiency through more thorough combustion, and not just by filtering unburnt mixture through catalytic conversion.

At present, the fuel efficiency of an ICE is typically between 25% and 40%, and 40% of the energy released is lost through the exhaust gas [2, 3]. Certain technologies have been developed to address these concerns of low fuel efficiency. Electrical and hydrogen fuel cell vehicles have gained much attention recently for their high energy efficiency of vehicle and zero emissions on the road [4-7]. However, their total energy efficiency is even lower than ICEs after accounting for multiple energy conversions; higher emission is expected is electricity generation and hydrogen generation [8]. Electric vehicles (HEVs) have been proven with economic benefits by the market, but

they have a slow market penetration, which reinforces that conventional vehicles will remain dominant in the next few decades. Other resorts like increasing the voltage of the electrical system of conventional vehicles from 12 volts to 42 volts, fuel economy can be improved with an estimated benefit between 9% and 12% [9]. But this change will affect the entire electronic supply chain, which is reluctant to those suppliers. Further, improvements by HEVs or 42-volts vehicles target at the system efficiency after the energy having been released and transformed as storable energy, and hence room still exists to address emission and efficiency at the root of energy transform.

For example, a lean combustion technology, which is desirable for its economic effect and low nitrogen oxide (NO_x) emission, as exhaust gas recirculation (EGR) has been studied for modern engines to meet emissions targets. Both NO_x emission and unburnt fuel are reduced for low burning temperature and re-burning of exhaust gas. However, high EGR rates may cause misfire at spark ignition (SI) [10]. Further, it is reported that ignition not only initiates combustion but also influences subsequent combustion [11]. For petrol engines with SI, ignition occurs between two electrodes about 2 mm apart. Only a negligible fraction (10^{-3}) of the mass is ignited [12]. It takes 0.2-0.4 millisecond for the initial flame ball to grow to 1-1.5 mm after the spark onset [13]. The combustion is incomplete as the flame front propagates in a low speed, which is called the laminar combustion phase. Increasing the ignition volume has become a trend in research for high-performance, lean-burn and low-emission petrol engines. For diesel engines, compression ignition (CI) gives higher fuel-efficiency, but also higher NO_x emissions due to a higher temperature. The drawbacks of SI and CI lead to the study of homogeneous charge compression ignition (HCCI) system, first proposed by Najt in 1983 [14]. However, challenge lays in HCCI is to change an open-loop control problem of the ignition timing to closed-loop, which requires accurate sensing and control of in-cylinder temperature and/or pressure. This has proved theoretically simple but practically impossible so far. An electromagnetic (EM) field based ignition is brought out in an U.S. Patent issued in 1952 [15]. Not only a large volume of fuel can be ignited at the same time, but also the ignition timing control can be realised by timing control of EM field formation. This EM field based ignition is named as microwave ignition (MI), homogeneous charge microwave ignition (HCMI) if the gas is homogeneous before ignition, because the EM field is formed by microwave resonant. HCMI is a promising

Wuqiao Luo is grateful to the China Scholarship Council and the University of Glasgow for a CSC scholarship.

solution for high fuel efficiency and low emission for the ignition happens in the whole space of cylinder head theoretically.

Through the comprehensive survey on efforts made so far to improve ignition and combustion of ICEs presented in this paper, a trend of increasing ignition volume is pointed out. Based on this aspect, the next section presents an overview of current ignition methods in three sub-sections: 1) point-based ignition; 2) zone-based ignition; 3) volume-based ignition. Then in section III method of HCMI is reviewed in depth and it is pointed out that HCMI holds the key to perfect solution. Future directions are discussed in section I based on the reviews in the former sections and conclusions are drawn in section V.

II. EXISTING TECHNIQUES FOR IMPROVING ENERGY EFFICIENCY IN IGNITION AND COMBUSTION

It is a long way to fully understand the mechanisms of ignition and combustion. A general understanding is that some particles of the mixture become active radicals at ignition stage through heat or high electric field, and oxidation reaction happens between these active radicals. When the heat released by oxidation reaction is greater than heat radiation loss, the temperature of the combustible mixture continues to rise to create more active radicals and combustion begins. The key point of stable ignition is whether there exist enough active radicals during the initial ignition to maintain the chemical reaction chain. Hence the volume of ignited particles at the initial ignition defines three different ignition categories as: 1) point-based ignition; 2) zone-based ignition; 3) volume-based ignition. Besides the volume of ignition, the techniques of ignition in ICE also aim at lean combustion and pragmatic application. Also well pre-mixed gas is pointed out as a key factor for combustion performance.

A. Point-Based Ignition

Comparing with the whole space of cylinder, the flame kernel of spark plug and laser ignition is so small that can be considered as a point in the space of cylinder head. Only a few particles are broken into active radicals around igniter, hence the chemical reaction starts within small area. Though multi-pointed ignition with multiple spark igniters or by laser-induced ignition may have better lean combustion limits than single-point ignition, the separated flame kernels are still with millimetre level diameters.

1) Single-Point High-Energy SI

Single-point high-energy SI can expand lean limits for conventional SI, though such improvement is limited. Lean combustion with excess air or high EGR would require higher ignition energy to ensure reliable ignition, which is to modify the circuit of ignition system to produce long duration sparks [16] or increase ignition voltage. But that didn't noticeably affect burning cycles due to the slow flame speed of the highly diluted combustible mixtures [17].

2) Multi-Point SI

Instead of raising ignition energy of spark igniter, multi-point SI is another direction to push the lean limit of combustion. Patent [18] describes a multi-point spark

plug can create multiple sparks with a special electrode which has plural projections extending radially from the central. Another effort is made in [19] that at least 3 spark plugs are installed in a removable cylinder head to have multiple sparks when ignites. Although such multi-point ignition system enhances the combustion and prevent possible misfire by increasing ignition points, it does not raise the ignition and combustion efficiency or reduce emission significantly.

3) Laser Ignition

Laser ignition system has been under much attention recently for its many potential benefits, including the greater control over the timing and location of ignition, absence of electrodes and possibility to allow ignition in multiple locations [20]. Laser-induced spark ignition which depends on electrical breakdown at focal point of laser is described in [21]. It replace the spark plug with a laser plug and there is no need for plug replacement since it is without electrodes. However, the ignition efficiency improvement is negligible compare to traditional spark plug ones because it is only initiates the ignition at a single point [21] and only 30-70% of the incident laser energy can be utilised in laser-induced spark ignition [22]. Morsy et al. [22, 23] proposed a laser-induced cavity ignition which ignition occurs inside a conical cavity. A non-focused laser beam is directed into and confined in the conical cavity by multiple reflections at the surface of the cavity [23] where the energy is focused for ignition.

Though laser ignition has been proved and tested in lab for its reliability and feasibility, the practical implementation of this laser application has still to be fully realised in a commercial automotive application yet [23]. Some researchers have done the real engine experiments. Brake mean effective pressure (BMEP) and NO_x emission have been evaluated in a real engine with laser ignition. But costs for installation of laser ignition would be a problem for commercial use [24].

B. Zone-Based Ignition

If the active radicals can be more than those in limited points like point-based ignition, it is expected the initial flame kernel would be bigger. Further, such active radicals help the flame propagate faster. In this category, more active radicals as transient plasma is produced by increasing ignition energy and lengthening ignition duration so as to enlarge the ignition volume.

1) Through a High Voltage Pulse

The transient plasma is produced by discharge between electrodes. In order to maximise the transient plasma, an electrode is places in a small recessed chamber in the cylinder connected to ground line [25]. Discharge happens between electrode and chamber which creates radial pattern arcs when high DC voltage is applied between them. It is disclosed in [25] that this ignition system has potential for improving lean combustion operation and is potentially useful for gasoline engine emissions reduction.

2) Radio Frequency Corona Ignition

In [26], a radio frequency (RF) corona ignition system is presented which adopts RF plasma jet to ignition the

combustible mixture. Though the performance of radio frequency plasma for high pressure gasoline direct injection engines are tested by [26], the mismatch impedance of RF antenna would be the biggest challenge in this method. In operation around 90% of the radio frequency power was reflected from the actuator, and only 100 W were delivered to the plasma in the first prototype of ignition systems. With a matching network, the delivery power rise up to 380 W. Impedance match for a various load is important and matching network design would be difficult if aiming at well controlled, reliable ignition system.

3) *Corona Railplug*

Corona happens between two parallel rails when electrically break down between the rails because high voltage is added between them. The corona would move from its initial place to the other end of rails in a cavity to ignite combustible mixture. In [27], a plasma jet is produced during the process of corona moving resulting in a zone-based ignition with a high speed of flame propagation.

C. *Volume-Based Ignition*

In zone-based ignition methods, the ignition occurs in a pre-determined area, i.e., discharge chamber as in [25], between rails as in [27]. The mechanical design of cylinder must be changed, along with the igniter, to enlarge the volume of ignition. In volume-based ignition, low temperature combustion (LTC), HCCI and MI are introduced not only for its large volume of ignited gas when ignition happens, but also for the ignition location is no longer limited by the mechanical structure, but could be the whole space of cylinder head.

1) *Diesel Low Temperature Combustion*

In diesel engine, the temperature of combustion is much higher than in petrol engine. Hence the NO_x emission is higher though the efficiency is higher than in petrol ones. LTC is introduced by manufacturers for premixed combustible mixture with a better efficiency and lower emission. High level of EGR and direct-injection techniques are adopted to meet LTC [28].

The ignition happens which the fuel is injected into hot air, hence how and when the fuel are injected effects the ignition and combustion. For more contact area and longer time between injection and ignition, the mixture of air and fuel would be mixed better and the ignition volume would be bigger. Various fuel-injector types and techniques have been studied. Double injection technique was used to premix the air-fuel before ignition [29]. The first injection was used as an early injection for fuel diffusion and the second injection was used as an ignition trigger for all the fuel. However, liquid-spray impingement on the cylinder liner often occurs for very early injection [28]. Narrow angle direct injection is investigated by B. Walter which a narrow spray cone angle was selected [30]. Instead of premix the air-fuel mixture by early injection, another approach is to delay ignition to allow more time for premixing, which the ignition occurs at early stage of the expansion stroke. Injection timing retarded and increasing EGR rate can help prolong the ignition delay [31]. Performances of

combustion is studied under different EGR levels and injection timings in [32]. Although high EGR levels contribute to complete combustion, the diluted mixtures increase CO emissions. Injection the fuel at early timings helps alleviated this problem, but did not eliminate it [32].

2) *Homogeneous Charge Compression Ignition*

HCCI combines characteristics of ignition in conventional gasoline engine and diesel engines. The advantage of the HCCI is that the fuel and air are homogeneous which the auto-ignition would occur at multiple locations simultaneously, not a single point ignition like for the SI system.

With HCCI systems it is possible to save up to 30% on fuel consumption compared to traditional SI engines. The peak temperature is significantly lower than it would be during a typical spark ignition and the NO_x level is negligible [33]. Future, a HCCI engine can operate on most kind of fuels such as gasoline, diesel, and the majority of alternative fuels, like compressed natural gas (CNG) or liquefied petroleum gas (LPG) [34]. Unfortunately, the control of the temperature is difficult since pressure and temperature need to be monitored for a closed loop control and can only be adjusted through the inlet and outlet valves.

3) *Microwave Ignition*

A U.S. Patent issued in 1952 [15] discloses an ignition system with high frequency waves including radio frequency and microwave. Like laser, microwave can also be used as energy source for ignition. A microwave is electromagnetic radiation having a frequency within the range of 10^9 Hz to 10^{12} Hz, which has longer wavelength than laser.

In early research, magnetron was source of microwave for the cost reason, but the size and the fixed resonant frequency make magnetron infeasible for practical application. With the help of high power semiconductor, though power efficiency of which is about 20% lower than magnetron, the microwave source can be small and flexible enough to fit into a vehicle. Besides the potential of extreme lean combustion due to an actually whole space ignition by MI, ignition timing is controlled by timing of microwave emitting. Though MI is relatively new and with few accomplished, but it is the most promising and practical method and should have received more attention.

The name of MI varies in literature. This paper mainly focus on homogeneous charge microwave ignition (HCMI) where the combustible gas is homogeneous before ignition induced by microwave resonant. Up to this day, there are two major methods for microwave ignition, based on the fundamentals of breakdown of fuel molecules. The first method is to ignite by high temperature because of resonating of the pole-like molecules. These molecules show characteristic resonant frequencies. Thus a certain energy level microwave can be added to a specified molecule to induce the resonance. Such resonance makes the temperature of the specified molecule raise till breakdown. The second one is using cavity resonance to breakdown the molecules of fuel,

which the microwave would resonant in a cavity to break the air-fuel mixtures inside it. Both of MI types would be discussed in section III in depth.

III. HOMOGENEOUS CHARGE MICROWAVE IGNITION

The two types of mechanisms in MI are molecule resonant breakdown and strong electric field breakdown. Though there hasn't been an extensive theory on the quantum-mechanical process of molecule resonant breakdown. But in [35], paired electrons and its symmetry in chemical bonds are influenced by electric and magnetic moments which are generated with distributed parameters using a self-supporting oscillation system working under a superposition of mode conditions. When the chemical bonds in hydrocarbons are broken by a certain frequency microwave, radical chain reaction will be set off, which is beginning of combustion. Further, some technical procedures suggest that with surprisingly low excitation energy in the classic sense, chemical processes can be accelerated because of the interaction of electromagnetic radiation and hydrocarbons [35]. In strong electric field breakdown, a stable and strong electric field is created by microwave resonant. When the microwave power is big enough, the molecules will be breakdown under such electric field.

A. Molecule Resonant Breakdown

The breakdown happens due to the energy absorbed from microwave in the form of kinetic energy. Makita and Ikeda patented an apparatus for ignition with microwave as assistance [36, 37]. It adopts a similar ignition structure as spark plug. The difference lies on a miniature microwave antenna which emits microwave to the initial plasma generated by spark plug. It can enlarge plasma as much as 300 times of the spark discharge alone, with high working pressure to 2.0 MPa. This method can also be referred as plasma enhanced ignition because the main idea is to create as much plasma as possible at the initial ignition time so as to improve ignition and combustion performance. Through experiments in single-cylinder research engines, the combustion stability is improved and the lean limit of equivalence ratio is increased from 0.59 to 0.49 [36].

The Micro Wave Ignition AG (MWI) in Germany have filed several patents and papers for a microwave ignition system since 2005 [35, 38, 39]. The mechanism of ignition is stated briefly in [35], and the implement of this method is revealed in a patent [38]. If the frequency of microwave corresponds with the splitting of the atomic or molecular energy level in air-fuel mixtures, stimulated emission or absorption of electromagnetic radiation occurs which finally leads to the stimulation of the chemical bonding and thus setting off the radical chain reaction [35]. This space ignition method can be utilised to save up to 30% fuel consumption and prevents up to 80% of pollutant emissions. Further, it is point out there is no thorough theory to explain how microwave influence the chemical bonds in hydrocarbons, but it implied that traditional understanding about reaction between microwave and chemical bonds might need to be refined. They had an ultimate goal of supplying every new vehicle with their technology within ten years [39].

B. Strong Electric Field Breakdown

The microwave resonates inside a cavity which forms a steady and strong electric field to breakdown air-fuel mixtures. The advantages of resonant breakdown are that the air-fuel mixture has a lower breakdown voltage when using microwave and the resonant cavity itself is its own amplifier [40]. The problem in microwave resonant in a cavity is how to decide frequency of microwave to meet the resonant requirement of the cavity.

In [41] the shape of cavity is modified to ensure that resonance under fixed frequency. A similar solution, presented by West Virginia University, is to build a small cavity inside the igniter instead of an extra chamber which need to change the shape of engine cylinder [40, 42]. A high voltage is generated at the open end because of resonance. This high voltage, like generated from ignition coil in conventional spark ignition system, is used to breakdown air-fuel mixtures for ignition [40]. Meanwhile the resonant frequency can fixed in accordance with features of the cavity. The most notable one is called as a quarter wave coaxial cavity resonant (QWCCR) which where the length of the cavity is a quarter of the wavelength. In 2002, Schleupen patented an similar ignition device with QWCCR and high-voltage jointly mounting to a flexfilm as the substrate [43]. In 2003, Schmidt and Ruoss published a patent for a microwave igniter with resonant frequency at multiple times of the resonant frequency under QWCCR [44]. Unfortunately, with the QWCCR the ignition occurs just around the centre electrode of the resonator, which dissipates the advantage of using the microwave ignition system.

Another solution is given by the group in University of Glasgow. They use antenna to transmit the microwave into engine cylinder. When the frequency meets the resonant requirement of cylinder, the resonance would generate a steady and strong electric field inside the cylinder. The problems lie at this method is to find the resonant frequency for the irregular shape of cylinder head and to couple microwave into cylinder. The effort have made in computer simulations and digital prototyping of HCMI systems and in applying computational intelligence to their design and optimised zero prototyping in order to bring about this revolution [45, 46]. An HCMI system is divided into three parts: microwave source, transmission line, and engine cylinder (resonator). Once the control signal is sent the microwave is generated by the source and transferred to the resonator through the transmission line. The source frequency is equivalent to the natural frequency of the combustion chamber. The ignition timing of an HCMI system is easier to control than of an HCCI system simply by controlling the timing of the generation of the microwave in an open-loop manner similar to the SI system. To breakdown the air-fuel mixture inside the resonance an electric field intensity of 1×10^5 V/m is required [47]. Providing a resonance condition and hence an enhanced electric field strength for a viable design of the HCMI system is the main goal. Table I gives compares of different methods of HCMI.

TABLE I. COMPARES OF DIFFERENT METHODS OF HCMI

Institution	Method	Accomplishment	Further work
Imagineering, Inc., Japan [36, 37]	Microwave plasma assistant	Lean limit reach to 0.61; 300 times OH radicals of conventional spark; under 2.0 MPa	Feasibility for vehicle application.
MWI Micro Wave Ignition AG [35, 38, 39]	MWI	Reduce by 30% fuel consumption and 80% emission	Prototype for the large-size engine.
West Virginia University [40, 42]	QWCCR	Shorter ignition time; lean limit reach to 0.8	Impedance mismatch of microwave transmit.
University of Glasgow [45, 46]	Resonant in cylinder	CAutoD based design for microwave coupling problem	MI simulation and CAutoD based design of HCMI system

IV. FUTURE DIRECTIONS

For high-performance, lean-burn and low-emission petrol engines, increasing the ignition volume and reduce the combustion temperature should be aimed by future works in ignition and combustion of ICEs. HCMI, though, is a relatively new method for ignition, it has the potential for both space ignition and low temperature combustion. Also with the development of power semiconductors, the application in vehicles is foreseeable. For recent research in HCMI, most of the engine tests are in lab. There is a long way for real vehicle application and road tests. In this section, two directions of HCMI research are given, to shine a light towards practical application in and vehicles.

A. Enhancing Design of Experiments and Tests of Electromagnetic Properties

Experimental research on ICE ignition system is mainly on feasibility with a physical engine. This has not covered electromagnetic (EM) properties of ignition and combustion. It is necessary to investigate interactions between the EM field and the particles or particles themselves in an ICE environment. In particular, breakdown condition of fuel particles and EM properties of plasma should have drawn more attention. How the complex environment of pressure, temperature, shape and medium of gas would affect the breakdown condition is still unknown, though the breakdown condition is the premise of all the experiments and simulations on ignition. In most research, a rough assumption is made about the breakdown condition. Following the fuel-air mixture being broken down, particles form a transient plasma state. Properties of plasma have been studied to enhance ignition. However, in a high pressure and varying temperature environment as in an ICE cylinder, the properties of plasma in such an EM field are not explicit.

B. Virtual-Physical Design for Manufacture

The time and cost associated with physical prototyping should be taken into consideration as early as the conceptual design stage. Virtual prototyping is a sound solution because by replacing an actual prototype with a digital one, much time and cost would be saved in design through prototyping. With intelligent algorithms and Computer-Automated Design, optimizing the prototype would be much more efficient than the conventional manual trial-and-error method. Towards Industry 4.0 and smart factory, product-oriented design-prototype integration would deliver a higher product feasibility and performance in the future.

V. CONCLUSION

A comprehensive survey on the attempts and developments of greener ignition and combustion systems for ICEs has been presented in this paper, covering methods from a single point ignition to volumetric ignition. Plasma ignition is also discussed in this paper, but a low-power solution is a long way off. HCMI appears to hold the key to a perfect solution, while other practically feasible methods such as corona and laser ignitions have suffered from incomplete combustion.

Many tests on physical engines have proved that HCMI offers significant performance, but the problem lies in a production-oriented design. Virtual prototyping through Computer-Automated Design would help in this regard and could lead to novel processes for Industry 4.0.

REFERENCES

- [1] S. Taryma, J. A. Ejsmont, G. Ronowski, B. Swieczko-Zurek, P. Mioduszewski, M. Drywa, et al., "Road texture influence on tire rolling resistance," *Key Engineering Materials*, 2014. 597: p. 193-198.
- [2] O. A. Kutlar, H. Arslan and A. T. Calik, "Methods to improve efficiency of four stroke, spark ignition engines at part load," *Energy Conversion and Management*, 2005. 46(20): p. 3202-3220.
- [3] A. Thiruvengadam, S. Pradhan, M. Besch, D. Carder and O. Delgado, "Heavy-duty vehicle diesel engine efficiency evaluation and energy audit." 2014, Mechanical and Aerospace Department, West Virginia University, The International Council on Clean Transportation, Washington, DC.
- [4] C. Chan, "The state of the art of electric and hybrid vehicles," *Proceedings of the IEEE*, 2002. 90(2): p. 247-275.
- [5] K. Frenken, M. Hekkert and P. Godfroij, "R&d portfolios in environmentally friendly automotive propulsion: Variety, competition and policy implications," *Technological Forecasting and Social Change*, 2004. 71(5): p. 485-507.
- [6] O. Andersen, Implementation of hydrogen gas as a transport fuel, in *Unintended consequences of renewable energy*. 2013, Springer. p. 47-54.
- [7] S. A.-B. Maher AR, "Effect of compression ratio, equivalence ratio and engine speed on the performance and emission characteristics of a spark ignition engine using hydrogen as a fuel," *Renewable Energy*, 2004. 29(15): p. 2245-2260.
- [8] H. Helms, M. Pehnt, U. Lambrecht and A. Liebich. "Electric vehicle and plug-in hybrid energy efficiency and life cycle emissions." in *18th International Symposium Transport and Air Pollution*, Session. 113. 2010.

- [9] B. Simpkin, R. Marco, C. D. A. CRF, M. Abele, G. Heuer, A. Ferré, et al., "Improved energy efficiency for conventional vehicles through an enhanced dual voltage architecture and new components with an attractive cost-benefit ratio," 2011.
- [10] T. Briggs, T. Alger and B. Mangold, "Advanced ignition systems evaluations for high-dilution si engines." 2014, SAE Technical Paper.
- [11] O. Yaşar, Plasma modeling of ignition for combustion simulations, in Computational science—iccs 2001. 2001, Springer. p. 1147-1155.
- [12] J. Tagalian and J. B. Heywood, "Flame initiation in a spark-ignition engine," *Combustion and Flame*, 1986. 64(2): p. 243-246.
- [13] S. Pischinger and J. B. Heywood, "A model for flame kernel development in a spark-ignition engine," *Symposium (International) on Combustion*, 1991. 23(1): p. 1033-1040.
- [14] P. M. Najt and D. E. Foster, "Compression-ignited homogeneous charge combustion." 1983, SAE Technical paper.
- [15] E. G. Linder. "Internal-combustion engine ignition," United States 2617841, 1952. Patent
- [16] W. R. Aiman, "Extended spark duration improves engine operation at high exhaust gas recirculation rates," *Combustion Science and Technology*, 1977. 15(3-4): p. 129-136.
- [17] J. D. Dale, M. Checkel and P. Smy, "Application of high energy ignition systems to engines," *Progress in energy and combustion science*, 1997. 23(5): p. 379-398.
- [18] R. J. Schaus. "Spark plug with multi-point firing cap," United States 6608430, 2003. Patent
- [19] J. A. Davis. "Multipoint spark ignition system," United States 4805570, 1989. Patent
- [20] T. X. Phuoc, "Laser-induced spark ignition fundamental and applications," *Optics and Lasers in Engineering*, 2006. 44(5): p. 351-397.
- [21] D. Bradley, C. G. W. Sheppard, I. M. Suardjaja and R. Woolley, "Fundamentals of high-energy spark ignition with lasers," *Combustion and Flame*, 2004. 138(1-2): p. 55-77.
- [22] M. Morsy, Y. Ko and S. Chung, "Laser-induced ignition using a conical cavity in CH_4 -air mixtures," *Combustion and flame*, 1999. 119(4): p. 473-482.
- [23] M. H. Morsy, "Review and recent developments of laser ignition for internal combustion engines applications," *Renewable and Sustainable Energy Reviews*, 2012. 16(7): p. 4849-4875.
- [24] G. Herdin, J. Klausner, E. Wintner, M. Weinrotter, J. Graf and K. Iskra. "Laser ignition: A new concept to use and increase the potentials of gas engines." in ASME 2005 Internal Combustion Engine Division Fall Technical Conference. 673-681. 2005.
- [25] C. D. Cathey, T. Tang, T. Shiraishi, T. Urushihara, A. Kuthi and M. A. Gundersen, "Nanosecond plasma ignition for improved performance of an internal combustion engine," *Plasma Science, IEEE Transactions on*, 2007. 35(6): p. 1664-1668.
- [26] G. Bachmaier, R. Baumgartner, D. Evers, R. Freitag, T. Hammer and G. Lins, "Radio frequency ignition system for gasoline direct injection engines," *international Journal of Plasma Environmental Science & Technology*, 2012. 6(2): p. 140-148.
- [27] J. Ellzey, M. Hall, X. Zhao and H. Tajima, "Computational and experimental study of a railplug igniter," *Experiments in fluids*, 1993. 14(6): p. 416-422.
- [28] J. E. Dec, "Advanced compression-ignition engines—understanding the in-cylinder processes," *Proceedings of the Combustion Institute*, 2009. 32(2): p. 2727-2742.
- [29] R. Hasegawa and H. Yanagihara, "Hcci combustion in di diesel engine." 2003, SAE Technical Paper.
- [30] B. Walter and B. Gatellier, "Development of the high power nadi™ concept using dual mode diesel combustion to achieve zero nox and particulate emissions." 2002, SAE Technical Paper.
- [31] S. Kimura, O. Aoki, Y. Kitahara and E. Aiyoshizawa, "Ultra-clean combustion technology combining a low-temperature and premixed combustion concept for meeting future emission standards." 2001, SAE Technical Paper.
- [32] S. Kook, C. Bae, P. C. Miles, D. Choi and L. M. Pickett, "The influence of charge dilution and injection timing on low-temperature diesel combustion and emissions." 2005, SAE Technical Paper.
- [33] J. Warnatz, U. Maas and R. W. Dibble, *Combustion: Physical and chemical fundamentals, modeling and simulation, experiments, pollutant formation*. 2006: Springer.
- [34] K. Epping, S. Aceves, R. Bechtold and J. Dec, "The potential of hcci combustion for high efficiency and low emissions." 2002, SAE Technical Paper.
- [35] N. Hirsch and A. Gallatz, "Space ignition method using microwave radiation," *MTZ worldwide*, 2009. 70(3): p. 32-35.
- [36] Y. Ikeda, A. Nishiyama and M. Kaneko, "Microwave enhanced ignition process for fuel mixture at elevated pressure of 1mpa," *regulation*, 2009. 1: p. 2.
- [37] M. Makita and Y. Ikeda. "Ignition or plasma generation apparatus," United States 8226901, 2012. Patent
- [38] V. Gallatz, N. Hirsch and I. Tarasova. "Fuel ignition process for engine combustion chamber involves creating microwave radiation in combustion chamber from source outside it," DE 10356916 B3, 2005. Patent
- [39] V. Gallatz, N. Hirsch and I. Tarasova. "Method for igniting combustion of fuel in a combustion chamber of an engine, associated device and engine," United States US7770551 B2, 2010. Patent
- [40] R. Stiles, G. J. Thompson and J. E. Smith, "Investigation of a radio frequency plasma ignitor for possible internal combustion engine use." 1997, SAE Technical Paper.
- [41] K. Kimura, A. Endo and I. Takezaki. "Ignition system for internal combustion engine," United States 4446826, 1984. Patent
- [42] F. A. Pertl and J. E. Smith, "Electromagnetic design of a novel microwave internal combustion engine ignition source, the quarter wave coaxial cavity igniter," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 2009. 223(11): p. 1405-1417.
- [43] R. Schleupen. "Ignition device for a high-frequency ignition," United States 6357426, 2002. Patent
- [44] H.-O. Ruoss and E. Schmidt. "Device for igniting an air-fuel mixture in an internal combustion engine," United States 7204220, 2003. Patent
- [45] Y. Li, F. Sun and A. Capuano. "Em field enabled timing control of hcci engines." in Proc 7th Asia-Pacific Conference on Control and Measurement. 39-44. Nyingchi, Tibet, China 2006.
- [46] L. Schoning and Y. Li. "Multivariable simulation on a homogeneous charged microwave ignition system." in Automation and Computing (ICAC), 2012 18th International Conference on. 1-6. 2012.
- [47] F. Lambert, L. Guido and T. Manfred. "Ignition and combustion support device using microwave technology for a gasoline engine," WO 9937911A, 1999. Patent

Creative Computing for Personalised Meta-Search Engine Based on Semantic Web

Sicong Ma, Siyan Li and Hongji Yang

Central for Creative Computing
Bath Spa University
Bath, UK
sicong.ma13@bathspa.ac.uk
Siyan.li14@bathspa.ac.uk
h.yang@bathspa.ac.uk

Abstract—as the Web technology develops rapidly, search engine as an important part of web technology has been widely used by people. A huge number of results are produced when users input query to normal search engines. However, many of the results are irrelevant. This paper presents a system that has been used to adjust the web search programs satisfy individual client's needs. This work can be explained in three directions. Firstly, a novel technique presented is proposed combining semantic web technology and personalised technology. Secondly, a new meta-search engine is created to make the results more relevant. Finally this paper presents a running example to show the query will be analysed more comprehensively are generated fewer but more personalised results to individual users are generated compared with Google search engine.

Keywords-component: *Meta-Search Engine;*
Personalisation; Search Engine; Information Retrieval

I. INTRODUCTION

Nowadays, web search engines develop very fast and successfully. Advertisements earn enough economic benefits by providing the user useful information or satisfying the user's query. Meanwhile, the users are really difficult to find accurate answers in results pages as the algorithms are originally used to discover pages relevant to a query or a typical keyword, not to consider the query the user submitted. A special word could mean many things in different contexts and the predict context can only be defined by the user. A normal search engine offers relevant information without analysing the query. Therefore, how to find useful answers that users want become big challenges in Web Search Engines or Meta-Search Engines.

People pay more attention to web search engines in order to achieve their various information requirements. Although the Web Technology develops strongly, there are still some barriers for search engines, especially in how to find users' real needs. When different users input same keywords, different information could be needed. To avoid this problem, Personalised Meta-Search is created. In Personalised Meta-Search, how to sufficiently meet user's real-time information need is an important matter. The query that the user's input to search engine is the most important vector to evaluate information requirement. However the query may be short, ambiguous and incomplete, which cannot express user's information

needs clearly. Therefore, it is difficult to satisfy user's requirement only by the query.

Personalised Meta-Search is becoming more and more important. There are three main methods in Personalised Meta-Search: 1. Users need to register some personal information ; 2. Systems offer a relevant feedback after searching results; and 3. Re-ranking the results depends on existing favorite rating. Thus, user's information implicitly needs to be collected. In order to make search results more personalised, a client-system need to be built, recording users' input and analysing the relevant information.

The paper is organized as following: Section II presents related work of web information retrieval and current personalised Meta-Search engine. Section III describes our personalised Meta-Search engine approach. Section IV introduce Personalised Creative Meta-Search engine. Section V concludes this paper.

II. RELATED WORK

A. Web Information Retrieval

Information Retrieval is to search and collect the most relevant information within a special database, satisfying a user's query. Searches can depend on metadata or full-text indexing. Processing of an information retrieval starts when a query is input in the system. Queries should be formal statements to meet information requirement. As a query may have many meanings, several objects might meet the query in the processing of information retrieval. It perhaps reduces degrees of relevancy.

Two serious problems block Information Retrieval. The first one is collection of documents. This is quite different between IR problems and database problems. Databases pay more attention in building a precise structure in the process of searching and retrieving terms. In the IR part, index is an important process to develop a document representation through assigning content descriptors or terms to the document. These vectors are used in evaluating the relevance of information to a user query. Another one is that query cannot be specially understood. A word may have several meanings. Systems cannot distinguish which meaning users' real needs are. It perhaps presents lots of irrelevant information.

Two types of terms are included in IR systems, which is objective and nonobjective. Generally, there is no

disagreement on how to assign objective terms such as author name, document URL and date of publication, as they are extrinsic to semantic content. However, in terms of nonobjective terms, it aims to reflect information presented in the document. And there is no agreement on the choice or degree of applicability of these terms. Therefore, they are also regarded as content terms. What indexing generally concerns is assigning nonobjective terms to documents.

B. Current Personalised Meta-Search Engine

Personalised search task aims to use personal information to define the most relevant search results. Mainstream of personalisation techniques is to extract user-centric profiles or features, such as location, gender and click history, and combine such information with the original ranking function. Teevan et. al. encoded user profiles by using the extraction of relevant feedback to re-rank the retrieved documents [1]. Dou et. al. [2] evaluated several personalised search strategies on a large scale, such as user information depend on re-ranking [3], and indicated that personalisation has mixed effects on the ranking performance. These and other personalization models use large amounts of search history to learn interest profiles for each user need enough data available to operate personalisation effectively [4].

Memory-based personalisation techniques learn direct relevance between query-URL pairs [5]. For instance, the current user will choose a particular URL when this query is given. That can give high possibilities of visitation and well performance. However, query coverage is limited. Re-ranking the top-n results [2] as a common strategy is used to perform the application of personalization, once the model is learned. In other words, the personalized models have no opportunity to facilitate results outside of top-n, however, it can promote results of high interests to the current user, and turn these into the top-ranked results.

On the basis of a personalised web search model, Zhengyu Zhu et. al. [6], proposed innovative query expansion that depends on a representation of personalized web search organization. The novel system fixed on client machine is a middleware connecting a user and a Web search engine, which can research the user's favorite implicitly and then produce the user profile automatically. When the user enters query keywords, more personalized expansion words are produced by the proposed approach. The same as query keywords, these produced expansion words are forwarded to a famous search engine such as Ask or Google. These expansion words can facilitate search engine retrieval information for a user based on his/her implicit search objectives. The novel Web search representation can build an ordinary search engine personalised, especially through personalised query expansion making the search engine provide different search results to different users entering the equivalent keywords. The experimental observations indicate the results and use of the proposed work for personalised information service of a search engine.

Scout [7] is another Personalised Meta-Search engine, which is regarded as a machine to collect pre-session relevant feedbacks. A modified query based on the

feedback is supported for resubmission to the Scout after the user finishes the feedback.

Outride [8], as an intermediary between the user and a search engine is another personalised engine. When queries are entered, modified query in outride is produced based on the user profile and returned results are filtered or re-ranked before being presented to the user.

The user interface is a sidebar of the browser, which contains the user's bookmarks, surf history along with a web directory (based on the ODP) as well as search results. The user profile is initially drawn from the bookmarks that are imported by the user. Then it is updated with information derived from pages the user browsed. Vector space methods re-rank research results, which compare the titles and other page metadata with the user profile. Experiments with the system suggested that users could find information they are searching more quickly by using Outride compared to the use of the underlying search engine only. Google acquired outride in 2001, but to date there has been no release of the technology.

C. Biancalana et.al. [9], provided a novel method that is Personalised Web search with social tagging in query expansion. Search systems collect enough personal information according to categorization and shared data before users input keywords. Because of social networks and collaborative tagging systems, the system presents more personalised and precise results to the user. Social bookmarking methods make a great contribution to two core points. Firstly, they give a permit to each system, which remembers the browsed URLs. Secondly, system can use tags to obtain more users' information.

Kyung-Joong Kim et. al. [10], proposed a novel way for personalised Meta-Search engine based on fuzzy concept network with link structure. Most of the popular search engines collect precision results according to link structure. Compared with a text based search engine, a link-based search engine always presents a higher quality results. Furthermore, according to fuzzy conception, the system presents the result that meet user's preference. In many approaches, depending on a user profile, the fuzzy theory can draw user's interest roughly.

Other methods, which are query independent, have been proposed. Smyth [11] demonstrated that click through data from users in the same "search community" could improve search result lists. The users in the same "search community" mean that a group of people who use a special interest Web portal or who work at the same company. However, Smyth presented evidence, showing the existence of search communities. Compared to general Web users, a group of employees who work in a single company had a higher query similarity threshold. Mei and Church [12] found that geographic location might be regarded as a reasonable proxy in terms of community, because they observed that a group of users with similar IP addresses could enhance search results.

III. A PERSONALISED META-SEARCH ENGINE APPROACH

A. Client Systems

In order to let network manage individual users, Web developers have searched methods to separate each users with personal Web accesses. Web servers cannot record any successive connection information by same user. Netscape Communications create the “cookie” in 1995. It is regarded as a mechanism, adding in Web client-server systems [13]. Web servers can record short strings by cookies when client input query. For instance, a client chooses a search result after inputting the query into a search engine. The cookie will store this data. When the users search the same site, “cookie” will mark the previous results that the user chooses automatically. In the running systems that provide many user profiles, cookies could eliminate the issue of running network addresses to stand for users through separating Web problems with a replying user, rather than only an IP. Meanwhile, in order to defeat cookies, programming offer individual users a special ID. Because of privacy problems, many people switch off the cookie in their systems [14].

The inherent limitations of server-side inferences are identified by the “next generation” of monitoring. Thus, recent research has focused on software expansion on the client to give a more granular view of the session activities. This kind of data can effectively be combined with server data or third-party ancillary data to produce a more integrated user profile.

Personalised shopping as one of application is described in this paper [15]. TELLIM will produce custom multimedia presentations, matching display and customer -appeal when it is running. The system supervises client interactions such as opening an audio or video player in order to do this. Then the system conducts a deduction whether the end-user is of interest in a variety of presentation elements generated. Only preferred presentation elements are proceeding by the application of a learning algorithm whereas the suppressed elements are just presented as links. The technology combines dynamic HTML (DHTML) with a Java applet to monitoring event on the client.

Systems developed by lots of researchers in human-computer interaction (HCI) have been used to monitor end user activity on desktops. Hilbert and Redmiles [16] demonstrate the history of monitoring systems by a variety of surveys, principally concerning low-level user interface actions such as mouse clicks and keystrokes. Despite problems of a user interface can be exposed by such systems through measuring the total mouse travel with operation, it is difficult to deduce higher level behavior, such as information flows in application because of the primary nature of monitored actions.

As a result of the emergence of the web as a marketing medium, lots of companies start developing Web monitoring systems to track usage by individual users. However, the well-known system is probably DoubleClick, which uses a unique DoubleClick ID and is stored as a cookie in a computer. Even though most of cookies are just sent to server-stored cookies,

DoubleClick servers provide users a variety of ads to see when they browse the web. As the Web client have to connect with DoubleClick server to retrieve ads, DoubleClick ID cookie that seems not to be served by DoubleClick is even sent users view pages. Through connecting uses across many web pages, DoubleClick can create a user’s profile. Furthermore if a user’s demographic information is available, DoubleClick can learn his behavior segmented by demographics.

B. Semantic Web Technology

When the user is browsing a webpage with large amounts of hyperlinks, the system will recommend some more relevant links that are related to users. Each link is ranked and given scores. Then when the scores of links are above a threshold, the list of recommendation will be produced. The ranking is classified into two categories that are habit-based ranking and interest-based ranking. In terms of habit-based ranking, the system does the comparison between the activities and the rules based on the user habit during the period of webpage browsing. Links that match the rules based on the user habit get high scores. In terms of interest-base, Scores are given to links depending on their relevant weights. The system recommends main links by combining habit-based ranking and interest-based ranking

Semantic linking has recently seen a sharp increase in interest. It is focused on evaluation campaigns such as the Text Analysis Conference Knowledge Base Population (KBP) track.2. Thus, Semantic linking has been widely applied in diversifying micro blog posts [17], supporting forensic text analysis [18], and it is even used in providing second screen applications from subtitles [19]. The most advanced linking approaches typically have an effect on the structure of its underlying knowledge base by taking consideration of some different aspects such as hyperlinks between pages or ontology structure for tasks like reducing ambiguities [20] and measuring relevance between concepts [21].

Concept of semantics has been introduced with OWL-S to conquer this limitation. With respect to inputs, outputs, preconditions and effects, functionality of a service is stated in this approach. Input and output terms of the service are part of a set of ontologies, which are presented as concepts. A single concept from two or more syntactically different terms is allowed to be referred by the use of ontology. Therefore, limitations caused by syntactic difference between terms are removed, as matching can be used to describe input and output terms based on concepts of ontologies. For semantic technology and creative computing, if we presume that both, school information and query are defined in OWL-s format, after that a school information *SI* and query *QU* match if

```

10. void elementLevelSenseFiltering(Node node)
20. AtomicConceptOfLabel[] nodeACOLs=getACOLs(node);
30. for each nodeACOL in nodeACOLs
40. String[] nodeWNSenses=getWNSenses(nodeACOL);
50. for each ACOL in nodeACOLs
60. if (ACOL!=nodeACOL)
70. String[] wnSenses=getWNSenses(ACOL);
80. for each nodeWNSense in nodeWNSenses
90. for each wnSense in wnSenses
100. if (isConnectedbyWN(nodeWNSense, focusNodeWNSense))
110. addToRefinedSenses(nodeACOL, nodeWNSense);
120. addToRefinedSenses(focusNodeACOL, focusNodeWNSense);
130. saveRefinedSenses(context);

140. void saveRefinedSenses(context)
150. for each node in context
160. AtomicConceptOfLabel[] nodeACOLs=getACOLs(node);
170. for each nodeACOL in NodeACOLs
180. if (hasRefinedSenses(nodeACOL))
190. //replace original senses with refined

```

Figure 1: The pseudo code of sense filtering

- For each input element in SI , there is one input element in QU . Let QU_{in} and SI_{in} represent the input meanings of query and the school information. The system can correctly execute the task if all the input meanings defined in the school information are met by the clients' requirement.
- For each output element in QU , there is one output in SI . Let QU_{out} and SI_{out} represent the output meanings of query and the school information. The clients can perfume the system if all the output meanings defined in the query are met by the school information.
- For each personalised details in SI , there is one personalised details in QU . Let QU_{pd} and SI_{pd} represent the personalised details of query and the school information. The system can correctly execute the task if all the personalised details defined in the school information are satisfied by the clients' requirement.
- For each effect in QU , there is one effect in SI . Let QU_{effect} and SI_{effect} represent the effect of query and the school information. The client can perfume the system if all the effects defined in the query are met by the school information.
- In order to make the results more relevant, this algorithm will combine semantic web and pseudo code in a creativity way. The pseudo code in Figure 1 describes the sense filtering system. This system is to eliminate irrelevant results by concepts of labels. For instance, the meaning of label *education* and *accounting* are defined ([Education, {sensesWN#4}] & [Accounting, {sensesWN#3}]) before the first sense-filtering step. As to the senses of *education* contains the senses of *accounting*, Accounting#2 \square Education#1, the rest of senses are eliminated. Thus, the result is [Education, {sensesWN#1}] \cap [Accounting, {sensesWN#1}].

C. Ranking Methods

In order to solve information retrieval problems, many algorithms developed strongly. BM25 model as a query dependent algorithm is widely used [22], and PageRank model as a query independent algorithm is popular as well [23]. However, parameter tuning is one of the biggest challenges for most of algorithms to solve IR problems. Over fitting on the training set as another barrier block new algorithm developed. With the development of machine learning, two problems that are automated parameter tuning and over fitting can be solved effectively [24].

Personalised ranking is to present a ranked result to a user according to his personal information. For example, based on user's watching history, YouTube recommends a personalised ranked list of videos that the user may want to watch. H. Kim presents a novel method that orders the pages returned by a search engine according to a user's preference [25]. He uses Google Customer Search Engine without building his own search engine. The user profile will be created by personal bookmark per contents and Divisive Hierarchy Clustering (DHC) algorithm. A user profile draws a user's preference from general to personalise. The profile is regarded as a tree. Larger clusters of queries represent general preferences that like big branches towards to the leaves. More special preferences like the leaves are mean by smaller clusters of queries. Query as a phrase that has several meanings. Differences among all queries are presented in the bookmarked web page by branch node. The leaf node show more especially preferences queries. The relevance among queries is evaluated depended on the co-occurrence in a web page.

IV. PERSONLISED CREATIVE META-SEARCH ENGINE (PCMOOGLE)

For many years, personalised search have been explored, and many personalisation approaches have been examined. For various users in different search contexts, however, there still exists uncertainty on whether personalisation is always significant on different queries.

A Personalised Meta-Search Engine is developing and testing, named Creativity Personalised Meta-Search Engine (PCMoogole), which is shown in Figure 2. It provides search results personalised by modifying query, limiting search area and re-ranking third party results depending on user profile. A user profile will be built by explicit user feedback method and commercial system connections. Few techniques in PCMoogole are similar, such as Google Customer Search Engine, as it is a client-side system that re-ranks results and limits search area. The biggest differences among these technologies are on the profile content and structure as well as the semantic web. In order to make the results more precise and relevant, this meta-search engine is only suitable for special search area. The processing of PCMoogole is as follow:

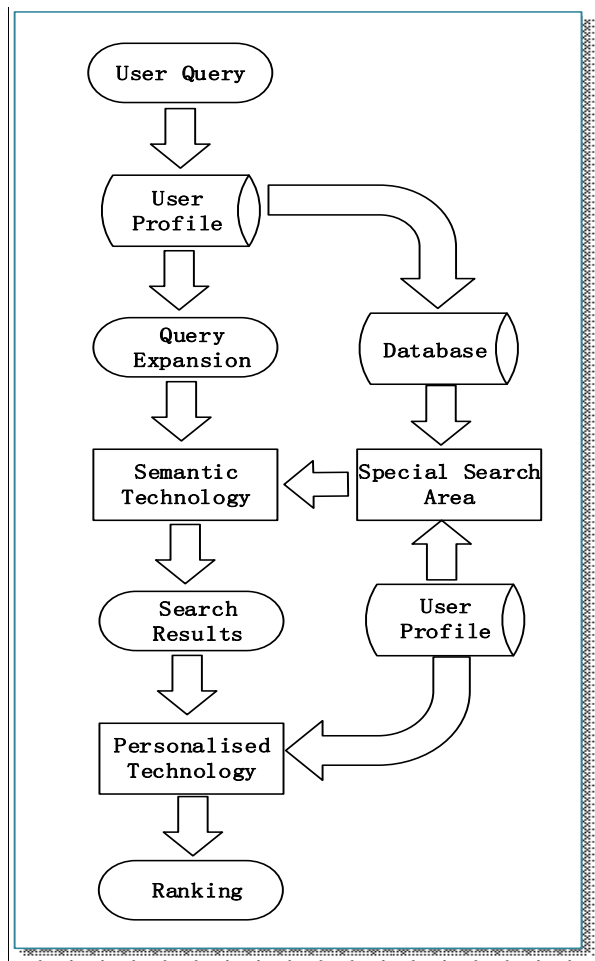


Figure 2: Architecture of PCMoogLe

- Step 1: After a user input a query to PCMoogLe, the system will recognise the user's interest by User Profile. According to User's personal information, the original query creates multiple

query expansion. Each of them is created by adding special personal information to the original query. Meanwhile, the system will record each query and views of URLs as a result of the query. It will store the number of client views to the page and a timestamp.

- Step 2: According to previous page views of web, PCMoogLe will record website views. Meanwhile, PCMoogLe will classify all websites. When the user inputs a query into search engine, it will classify websites according to user's information and query. Then the system will provide searches in the top 30 websites ranked with highest views in a category based on classification results.
- Step 3: In order to present more relevant results, the system matches extended queries and information from the database by semantic web technology and creative computing. In this step, pseudo will be used, eliminating the relevant results. Query expansion and special websites information will be defined (A and B represent query expansion and special websites information separately, $[A, \{\text{sensesWN}\#4\}]$ & $[Accounting, \{\text{sensesWN}\#3\}]$) after the system received the query expansion and special websites information. The results exact from $[A, \{\text{sensesWN}\#1\}] \cap [B, \{\text{sensesWN}\#1\}]$.
- Step 4: Depending on user's interest, the pages will be scored by DHC algorithm in bookmark system. The result from higher scoring website will present in the first page.
- Step 5: Results will present to the user. For example, as can be seen from the Figure 3 and Figure 4. The user input the same query to PCMoogLe and Google, separately. Google focuses on normal people, thus the results cover many aspects, such as advisement, maps, photo and some priority websites. However, PCMoogLe

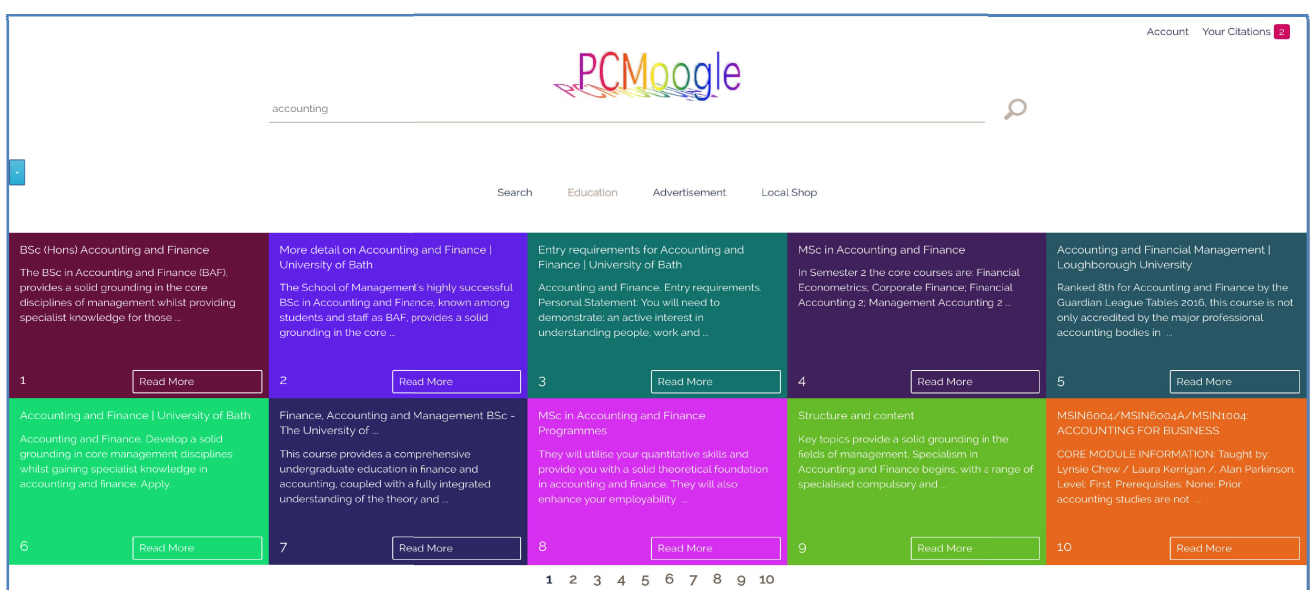


Figure 3: Ranked results generated by PCMoogLe

only presents 10 results in the first page. These results are based on each user-searched behavior, which are dwell time, IP address, and visitation rate and bounce rate, and ranks the results by semantic technology and creative computing. Thus, the results from PCMoogole do not have irrelevant information such as advertisement and photos. Meanwhile, only present 10 results in first page, user will find what they want easily.

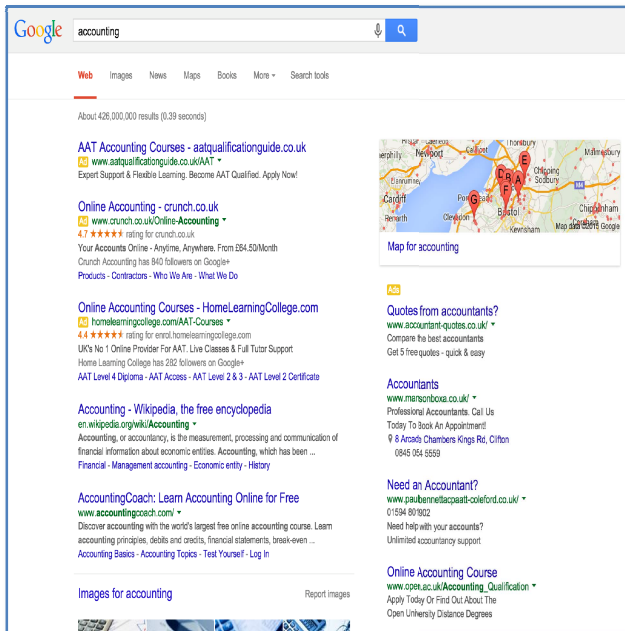


Figure 4: Ranked results generated by Google

CONCLUSIONS

Web search engines such as Google do not cooperate well with SWDS, as they are designed to work with natural languages and anticipate un-structured text made up of words included in the document. If current web search engines do not understand the structure and semantics of Semantic Web Documents (SWDs), they will not make full use of them. A novel meta-search engine PCMoogole is created. This search engine is on the basis of existing semantic meta-search engine within creative technology. The approach of PCMoogole applies the Semantic Web eliminated the number of web pages. This method will extend queries by special information. Thus, more relevant queries search in a smaller data set. It can decline the danger of eliminating new important terms. In the process of matching, pseudo code will be used to eliminate irrelevant results. PCMoogole is deeply dug by the combinations of user query and personal information with the use of semantic web technology and creative computing and collects the most relevant and personalised information to the users.

REFERENCES

- [1] J. Teevan, S. Dumais, and E. Horvitz, "Personalizing Search Via automated Analysis of Interests and Activities", *SIGIR'05 ACM*, pp. 449–456, 2005.
- [2] Z. Dou, R. Song, and J. Wen, "A Large-scale Evaluation and Analysis of Personalized Search Strategies", *WWW'07 ACM*, pp. 581–590, 2007.
- [3] P. Chirita, W. Nejdl, R. Paiu, and C. Kohlsch, "Using ODP Metadata to Personalize Search.", *SIGIR'05*, pp. 178–185, *ACM*, 2005.
- [4] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais, and B. Billerbeck, "Probabilistic Models for Personalizing Web Search", *WSDM'12*, pp. 433–442, 2012.
- [5] J. Teevan, D. J. Liebling, and G. Ravichandran Geetha, "Understanding and Predicting Personal Navigation", *WSDM'11*, pp. 85–94, 2011.
- [6] Z. Zhu, J. Xu, X. Ren, Y. Tian and L. Li, "Query Expansion based on a Personalized Web Search Model", *Third International Conference on Semantics, Knowledge and Grid*, pp. 128–133, 2007.
- [7] S. Kumar, L. ErtaAoz, S. Singhal, B. U. Oztekin, E. Han, and V. Kumar, "Personalized Proile Based Search Interface with Ranked and Clustered Display", *Technical report, University of Minnesota, Department of Computer Science*, 2001, <https://www.cs.umn.edu/tech-reports/listing/tr2001/01-023.pdf>.
- [8] J. E. Pitkow, H. Shtze, T. A. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar and T. M. Breuel, "Personalised Search", *Communications of the ACM*, Vol. 4, pp. 50–55, 2002.
- [9] C. Biancalana and A. Micarelli, "Social Tagging in Query Expansion: A New Way for Personalised Web Search", *International Conference on Computational Science and Engineering (CSE '09)*, Vol. 4, pp. 1060, 2009.
- [10] K. Kim and S. Cho, "A Personalised Web Search Engine using Fuzzy Concept Network with Link Structure", *Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, Vol. 1, pp. 81–86, 2001.
- [11] Smyth, B., "A Community-based Approach to Personalizing Web Search", *IEEE Computer*, Vol. 40 pp. 42–50, 2007.
- [12] Q. Mei and K. Church, "Entropy of search logs: how hard is search? with personalization? with backoff?", *WSDM*, pp. 45–54, 2008.
- [13] M. B. Cooley, M. Robert, and J. Srivastava, "Automatic Personalization based on Web Usage Mining", *Communications of the ACM*, Vol. 48, pp. 143–151, 2000.
- [14] J. Schwartz, "Giving Web a Memory Cost its Users Privacy", *New York Times*, C1 2001.
- [15] T. Jörding, Stefan Michel, "Personalized Shopping in the Web by Monitoring the Customer", *Active Web*, UK 2001.
- [16] B. Mobasher, H. Dai, T. Luo and M. Nakagawa, "Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data", *ITWP*, 2001.
- [17] E. Meij, W. Weerkamp, and M. de Rijke, "Adding Semantics to Microblog Posts", *WSDM '12*, pp. 563–572, 2012.
- [18] Z. Ren, D. van Dijk, D. Graus, N. van der Knaap, H. Henseler, and M. de Rijke, "Semantic Linking and Contextualization for Social Forensic Text Analysis. In European Intelligence and Security Informatics Conference (EISIC 2013)", pp. 96–99, 2013.
- [19] D. Odijk, E. Meij, and M. de Rijke, "Feeding the Second Screen: Semantic Linking based on Subtitles", *OAIR '13*, 2013.
- [20] D. Milne and I. H. Witten, "Learning to Link with Wikipedia", *CIKM '08*, pp. 509–518, 2008.
- [21] D. Milne and I. H. Witten, "An Effective, Low-cost Measure of Semantic Relatedness obtained from Wikipedia Links", *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pp. 25–30, July 2008.
- [22] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3", *TREC-3*, pp. 109–126, 1995.
- [23] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank Citation Ranking: Bringing order to the Web", *Technical Report 1999-66*, Stanford InfoLab, 1999.
- [24] T. Liu, "Learning to Rank for Information Retrieval", *Foundations and Trends in Information Retrieval*, pp. 225–331, 2009.
- [25] H. Kim and P. K. Chan, "Personalised Ranking of Search Results with Learned User Interest Hierarchies from Bookmarks", *Department of Computer Information Systems Livingstone College*, USA 2004.

Applying Semantic Web Techniques to Poem Analysis

Xuan Wang and Hongji Yang

Center for Creative Computing

Bath Spa University

United Kingdom

xuan.wang13@bathspa.ac.uk and h.yang@bathspa.ac.uk

Abstract—Computing technology has been eyed in many fields. Creative Computing addresses the challenges of reconciling the objective precision of computer science with the subjective ambiguities of the arts and humanities. Poetry is a creative language which is full of imagination and beauty. The ambiguity of poetry increases the difficulty in interpretation and appreciation. Based on Semantic Web techniques, the research intends to comprehensively analyse poetic elements such as syntax, style, metaphor and genre. The creativity and innovation of poetry will also be considered. In terms of the analysis results, there will be an assessment to the poem. The method aims to help promote the interdisciplinary study of language and contribute to the analysis and appreciation of verbal art.

Keywords—Poetry Analysis; Semantic Web; Ontology; Creative Computing; E-Assessment

I. INTRODUCTION

With the rapid development of computer industry, more and more people pay higher attention to combining the computer science with human arts. Since human mind creates beautiful literature, it is a great challenge for computer machine to understand and analyse the achievements. As one of the most significant elements in literature, poem is a research object of exclusive value and culture meaning.

The following sections review previous computing analysis work on poetry, and then introduce our research which applies Semantic Web techniques to analysing poems.

II. ACADEMIC BACKGROUND

A. Previous Research on Poem Analysis

The analysis of poem dates back to the 1940's when poet and literary critic Josephine Miles began her extensive work analysing the surface statistics of poetry across time [1, 2]. While Miles's work was influential in establishing a statistical framework for thinking about poetry, it was done largely by hand and thus limited in scope and size.

Most recently, Hayward's connectionist model of poetic meter incorporated more sophisticated and varied features in the analysis [3]. For every feature considered, including prosody, meter, and syntax, Hayward hand-assigned numeric scores to each syllable in ten samples of poetry. Analysing these scores allowed him to identify unique patterns for each poet and to note similarities within each period. However, this analysis required Hayward's personal assessment of the poems as well as assignment of feature scores. Since it is

unfeasible to apply this method to a large set of poems, Hayward's model also faces limitations in size.

One of the most thorough and sophisticated computing analysis of poems to date is the PoetryAnalyzer, where Kaplan and Blei's work on visualised comparison of style in American poetry [4]. Modern statistical and computational tools allowed the authors to integrate more features to analyse a large set of poems in an automated manner. The authors mapped poems from different poets and eras into a vector space based on three types of stylistic elements -- orthographic, syntactic, and phonemic -- in order to find stylistic similarities among poems.

Other research on the analysis of poetry focused on quantifying poetic devices such as rhyme and meter [5, 6], or classifying poems based on the poet and style [7, 8]. These studies show that computational methods can reveal interesting statistical properties in poetic language that allow us to better understand and categorise great works of literature. However, there has been very little work on assessment poetry. Moreover, it is necessary to establish a scientific, systematic evaluation poetry system.

In this paper, the research aims to use Semantic Web techniques to analyse poems and assess them. Moreover, it will focus on the creativity aspect.

B. Meaning of Semantic Web

The term "Semantic Web" was presented by the World Wide Web Consortium. It aims at converting the current web, dominated by unstructured and semi-structured documents into a "web of data" [9]. The emergence of the semantic web holds the promise of bringing the web to an all-new level with technologies that can import and channel data at the most granular level, providing a new frontier in data linking.

The Semantic Web is an evolution and extension of the existing Web that allows computers to manipulate data and information. There is the great appeal in the Web that has the potential ability to "know" and "understand" data with an even greater capacity to process better. This adds a more humanistic quality to standard data processing because the Semantic Web seeks to close the gap between merely providing documents to people and automatic data and information processing.

There are some advantages of the new technology. Firstly, the Semantic Web helps improve productivity and efficiency in terms of data and information dissemination to

people. The accuracy of web search can become more accurate and precise while removing more ambiguity that usually arises with current search engines. This is contributed by the idea that semantics allow for any existing knowledge-representation system to be exported onto the Web. Many organisations can use this in daily business functions to greatly speed up communication and information-sharing. So it can provide a better communication platform for us to analyse poems, which enables collaborative work more effectively.

Another advantage of the Semantic Web is the achievement of automation with minimal human intervention. It means that machines become capable to process and “understand” the data that they merely display at present. Due to this advantage, we can utilise Semantic Web techniques to make computers “understand” poems, which can facilitates poem analysis with more precision and comprehensiveness.

The Semantic Web consists primarily of three technical standards:

RDF (Resource Description Framework): The data modelling language for the Semantic Web. All Semantic Web information is stored and represented in the RDF.

SPARQL (SPARQL Protocol and RDF Query Language): The query language of the Semantic Web. It is specifically designed to query data across various systems.

OWL (Web Ontology Language): The schema language, or knowledge representation language, of the Semantic Web. OWL enables people to define concepts so that these concepts can be reused as much and as often as possible.

In Semantic Web, an ontology formally represents knowledge as a set of concepts within a domain and the relationships between pairs of concepts. Some researchers have constructed ontologies about poetry such as Tang ontology and Su Shi ontology, which are based on 300 Tang poems and poems by Su Shi [10]. Yao designed a Chinese Ancient Poetry Ontology [11] and Feng presented a Poetry Learning Environment using the Learning Context Ontology, Poetry Noun Ontology and Music Emotion Ontology [12].

It can be seen that current poetry ontologies are limited to Chinese poetry in literature, and they are not comprehensive enough. The research intends to design ontologies of poetic elements such as syntax, style, metaphor and genre. Moreover, it will focus on analysing the creativity aspect.

III. SYSTEM DESIGN

In order to systematically evaluate poems, a poetry analysis system is developed using Semantic Web techniques and supporting tools. Based on existing poem resources, the system increases ontologies and processing module, which could analyse poems from different aspects.

A. System Architecture

The system has three layers: User Interface Layer, Data Processing Layer and Data Storage Layer as shown in Fig.1.

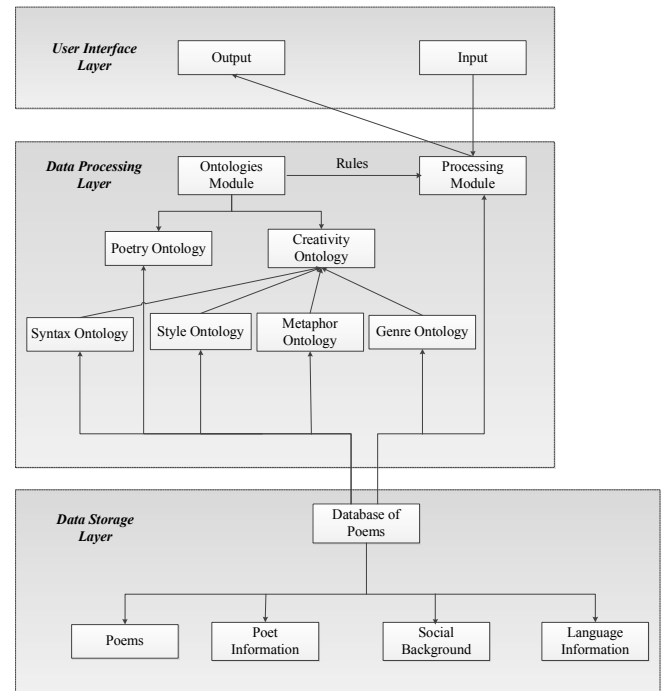


Figure 1: System Architecture

1) User Interface Layer

User Interface Layer is the entrance of the system, which communicates with the users. It displays the information and provides the interface to access the system for users. Poems could be input to the Data Processing Layer, and it will return analysis results to the output interface after processing.

2) Data Processing Layer

Data Processing Layer receives the requests from User Interface Layer, and handles data processing and returning the analysis results to the client. It includes two modules: Ontologies module and Processing module.

a) Ontologies Module:

There are six ontologies module in this part: Poetry Ontology, Creativity Ontology, Syntax Ontology, Style Ontology, Metaphor Ontology and Genre Ontology. The assessment elements are shown in Fig 2.

Poetry Ontology: an ontology of poetry aspects, like categories, diction, sound patterns, rhyme, meter and stanza, etc, as shown in Fig 3.

Syntax Ontology: an ontology of poem syntactic, such as line break, rhythm, repetition and rhyme.

Style Ontology: an ontology of poem diction, such as formal, stately, noble or informal, etc.

Metaphor Ontology: an ontology of metaphor. Many abstract and common concepts can be embodied or evoked by surprising metaphor. The analysis will focus on negative and positive emotion.

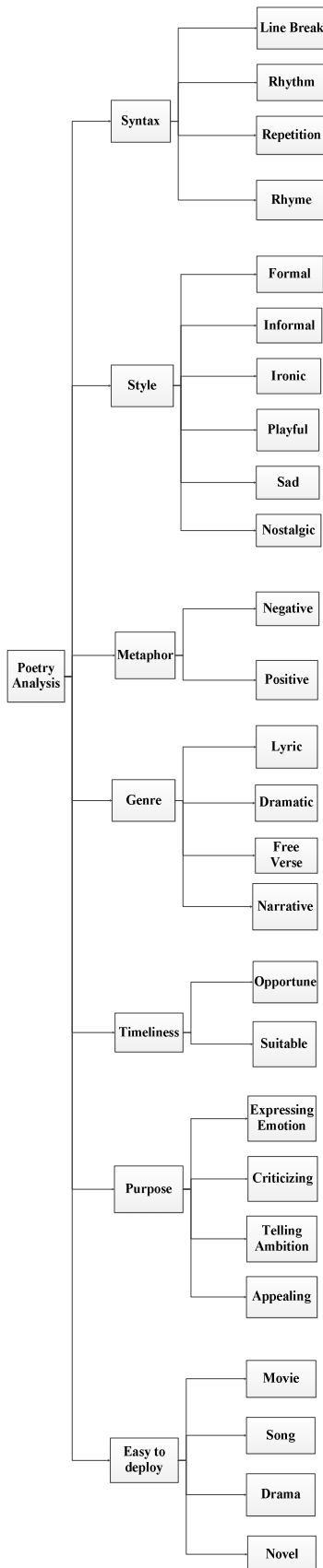


Figure 2: Assessment Elements

Genre Ontology: an ontology of poem styles, such as lyric, dramatic, free verse and narrative.

Creativity Ontology: an ontology of creativity character, such as new and useful. Creativity has been much used and becomes a hot topic. However, what is creativity? How to define creativity? In the Eighth Oxford English Dictionary, “creativity” is defined as the use of imagination or original ideas to create something. In a 2003 summary of scientific research into creativity, Michael Mumford suggested, “Over the course of the last decade, however, we seem to have reached a general agreement that creativity involves the production of novel, useful products. [13]” According to the analysis, “New” and “Useful” will be considered as the creative factor of poems. Based on the following ontologies, the system will also analyse other three aspects of poetry:

Timeliness: According to the social background and language information, if the theme or diction is opportune at that time, it could be considered as timely.

Purpose: There must be a purpose for a poem. By analysing a certain number of poems, four types of poetic purposes are elicited out by the research: Expressing Emotion, Criticising, Telling Ambition and Appealing. Expressing Emotion is about evaluating the sentiment aspect of a poem, either positive or negative. Criticising is about estimating what is criticised by a poem and the reasons underneath the criticism. Telling Ambition is about apprising what kind of ambition is expressed by a poem and the circumstance that triggers it. Appealing is about evaluating what is appealed by a poem and the historical events associated with it.

Easy to deploy: It is known to us all that various forms of art including poetry, essay, novel, music, dance, drama and movie are not independent but interrelated with each other. Through editing materials, they could be transformed into each other, which is one of fascinating features of art. For example, the famous tragedy *Romeo and Juliet* written by Shakespeare early in his career is always adopted into opera. Therefore, an assessment on whether it is easy for a poem to deploy into other forms of art should be considered.

b) Processing Module:

Processing module is responsible for data processing and returning results to the User Interface Layer. Based on the original database of poems, the ontologies data will be processed.

3) Data Storage Layer

Data storage layer stores the database of poems, which has four modules: Poems, Poet Information, Social background and language Information.

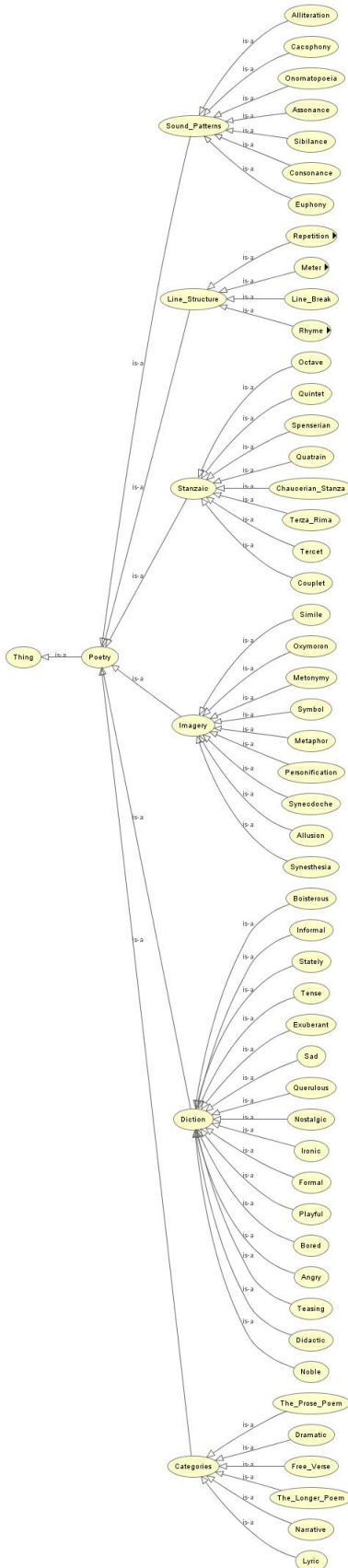


Figure 3: Ontology of Poetry

B. System Realisation

The whole system realisation is complicated and still in process of developing. However, the key points of the system are building Ontologies, processing Ontologies by Jena and reasoning Ontologies by Racer. Some algorithms are also presented.

Taking the creativity analysis of metaphor for instance, the algorithm is comprised of three phases. Firstly, plain text of poem is read from user input. It will be directed to a word processor which distinguishes and marks each word by various colours, e.g. red for adjectives, blue for nouns, green for verbs, *etc.* Since only the 'adj+n' draws our interest in terms of metaphor, only red concatenated with blue phrases are extracted and stored in a variable. Note that the extraction takes advantage of a localisation function which selects phrases from the text. Then the adjective part and noun part of these phrases are grouped in arrays, respectively. The number of nouns concerned is denoted using size function. Lastly, iterations are launched for n_number times, each of which examines if the adjective used to describe the noun is incorporated in its common metaphor database. Otherwise, accumulate credits for exceptional uses into k . The creativity rate of metaphor is derived from k divided by n_number .

The pseudo code of this part is shown as below:

Program metaphor_analysis

```

input_text = pscanf('poem')
WORD_PROCESSOR(input_text);
adj_n_array = LOCATE('RED+BLUE' or
                    'RED+RED+BLUE' or
                    'RED+RED+RED+BLUE');

adj_array = LOCATE('RED' or
                  'RED+RED' or
                  'RED+RED+RED', adj_n_array);
n_array = LOCATE('BLUE', adj_n_array);
n_number = SIZE(n_array);
k = 0;
For i = 1 to n_number do
{database = ADJ_COMMON(n_array( i ))
  if adj_array( i ) □ database
  else if
    k = k + 1
  end if };
creativity_rate = k / n_number;
```

Another algorithm is applied for style analysis with five steps. Firstly, plain text of poem is read from user input. It will be directed to a word root program which extracts roots of words, e.g. amen from amenable, break from broken, compete from competition, *etc.* Then the roots with sentimental colour are selected and stored in `style_root_array`. The number of roots, i.e. `n_length`, is denoted using `size` function. Create an `accumulate_style` vector of which the values are zero with `n_length` dimension. Then evaluate each root and derive a `style_vector` of which values represent mark of its stylish category. For example, 'die' is marked as (0,0,5,0,1) in integer range 0 to 5. The dimension of vector is the stylish categories of interest while each column stands for a certain aspect such as formal, informal, sadness, delightful or ironic. Thus, a root scores up to 5 if it presents stronger style, whereas, remains 0 if it is neutral in terms of that stylish category. Sum up theses style vectors to `accumulate_style` with the help of loop structure. At last, an average stylish evaluation is derived via dividing the accumulation by the number of roots to undermine the influence of poem length diversity. The `VALUE` function is designed to estimate finally the contributions of all significant roots to the stylish characteristics of the poem.

The pseudo code of this part is shown below:

Program style_analysis

```
input_text = pscanf('poem');
root_text = WORD_ROOT(input_text);
style_root_array = SELECT_STYLE_ROOT(root_text);
n_length = SIZE(style_root_array);
accumulate_style = ZEROS(n_length);
For i = 1 to n_length do
    {style_vector = ROOT_EVALUATE(style_root_array( i ) )
    accumulate_style = accumulate_style + style_vector
    };
STYLE = accumulate_style / n_length;
VALUE(STYLE);
```

C. Example

As an example, the poem below will be analysed.

Dreams

*Hold fast to dreams
For if dreams die
Life is a broken-winged bird
That cannot fly.*

*Hold fast to dreams
For when dreams go
Life is a barren field
Frozen with snow.*

—Langston Hughes

Metaphor analyser would read through the text marking adjectives and nouns in red and blue, respectively. The `LOCATE` function would focus on those adjective-concatenate-noun phrases. In the given poem, 'broken-winged bird' and 'barren field' are the only two valid phrases such that the total number of core nouns equals 2. After the common adjective matching process, 'barren' is believed a usual metaphor for 'field'. However, 'broken-winged' seems an uncommon vivid sketch of 'bird' which adds 1 to the creative stack. Hence, the creativity rate yields 0.5.

Meanwhile, the style analyser initially rewrites the poem in root words which reads

Dream

*Hold fast to dream
For if dream die
Life be a break wing bird
That cannot fly.*

*Hold fast to dream
For when dream go
Life be a barren field
Freeze with snow.*

'Dream', 'dream', 'die', 'break', 'dream', 'dream', 'barren', 'freeze' and 'snow' draw our attention in sequence, which will be extracted to form an array. The stylish value vector should then be determined and accumulated in terms of formality, informality, sadness, delightfulness and irony.

The values of the vectors are shown in the table on next page.

Consequently, the style-characteristic vector is derived from accumulative style vector divided by the length of stylish root array. It is the `VALUE` function that eventually interprets the determination on the style of the poem according to the style-characteristic vector. And the style value of this poem is (0, 0, 1.33, 0.44, 0.33). Since there shows null for formality or informality, it could be drawn that the poem is not interpretable in terms of formality. In contrast to delightfulness, sadness scores significantly higher, which suggests the poem is of remarkable sadness. As for irony, 0.33 is relatively smaller than threshold value such that it could not be determined as an ironic poem for lack of evidence.

TABLE 1: VALUES OF ROOT WORDS

Style	dream	die	break	barren	freeze	snow	Total
Formality	0	0	0	0	0	0	0
Informality	0	0	0	0	0	0	0
Sadness	0	5	2	3	1	1	12
Delightfulness	1	0	0	0	0	0	4
Irony	0	1	0	2	0	0	3
Times	4	1	1	1	1	1	

IV. CONCLUSIONS

The research presented in this paper aims to analyse the creativity of poetry based on the elements of syntax, style, metaphor as well as genre by using Semantic Web Techniques. A three-layer system is proposed based on Ontologies of analysis elements and relevant algorithms. The algorithms for the creativity analysis of metaphor and style analysis have been presented in details. Moreover, an example poem is analysed and its evaluation result is obtained. According to the result, the creativity rate of metaphor in the poem has been derived and the style has also been concluded. Besides metaphor and style, the system also investigates other aspects including syntax, genre, purpose of which trained databases are developed based on statistical analysis of all kinds of poems.

Since the great potential of the pervasive utilization of computing in every field, more and more fascinating innovations done by computing are demanded to facilitate the evermore sophisticated society. The research is aiming to use the great power of computing to analyse poems. However, there are far more work could be conducted.

By analysing poems, insights about the decisive elements and impacting factors of building creative poems will be excavated. According to the results above, the system makes the further innovation about generating creative poems through computing possible. Hence, one of the promising future considerations is using computing to generate poems, both traditional and inventive. The former is about constructing poems based on traditional formats with new content, such as using creative words. The latter is about composing poems with totally new formats and content, which requires more creativity for computing.

REFERENCES

- [1] J. Miles, "Major Adjectives in English Poetry: From Wyatt to Auden", *University of California Publications in English*, vol. 12, No. 3, 1946, pp. 305-426.
- [2] J. Miles, *Style and Proportion: The Language of Prose and Poetry*, Brown and Co, Boston, 1967.
- [3] M. Hayward, "Analysis of a Corpus of Poetry by a Connectionist Model of Poetic Meter", *Poetics*, vol. 24, 1996, pp. 1-11.
- [4] D. Kaplan and D. Blei, "A Computational Approach to Style in American Poetry", *7th IEEE International Conference on Data Mining*, Omaha, USA, 2007, pp. 553-558.
- [5] D. Genzel, et al., "Poetic Statistical Machine Translation: Rhyme and Meter", *EMNLP Conference on Empirical Methods in Natural Language Processing*, SIGDAT, Massachusetts, USA, SIGDAT, 2010, pp. 158-166.
- [6] E. Greene, et al., "Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation", *EMNLP Conference on Empirical Methods in Natural Language Processing*, SIGDAT, Massachusetts, USA, SIGDAT, 2010, pp. 524-533.
- [7] Z. He, et al., "SVM-based Classification Method for Poetry Style", *IEEE International Conference on Machine Learning and Cybernetics*, Hong Kong, China, 2007, pp. 2936-2940.
- [8] A. C. Fang, et al., "Adapting NLP and Corpus Analysis Techniques to Structured Imagery Analysis in Classical Chinese Poetry", *AdaptLRTtoND Workshop on Adaptation of Language Resources and Technology to New Domains*, FLAReNet Project, Stroudsburg, PA, USA, 2009, pp. 27-34.
- [9] T. B. Lee, et al., "The Semantic Web", *Scientific American Magazine*, Macmillan Publisher Ltd, May, 2001, pp. 29-37.
- [10] C. Huang, et al., "Reconstructing the Ontology of the Tang Dynasty: A Pilot Study of the Shakespearean-garden Approach", *18th Pacific Asia Conference on Language, Informaion and Computation*, PACLIC Steering Committee, Waseda University, Tokyo, 2004.
- [11] R. Yao and J. Zhang, "Design and Implementation of Chinese Ancient Poetry Learning System Based on Domain Ontology", *IEEE International Conference on e-Education, e-Business, e-Management and e-Learning*, Sanya, China, 2010, pp. 460-463.
- [12] J. Weng, et al., "Constructing an Immersive Poetry Learning Multimedia Environment using Ontology-based Approach", *IEEE International Conference on Ubi-Media Computing*, Lanzhou, China, 2008, pp. 308-313.
- [13] M. D. Mumford, "Where Have We Been, Where are We Going? Taking Stock in Creativity Research", *Creativity Research Journal*, Routledge Company, vol.15, 2003, pp. 107-120.
- [14] D. Rubin, *Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes*, New York: Oxford University Press, 1995.
- [15] R. Lea, et al., "Sweet Silent Thought: Alliteration and Resonance in Poetry Comprehension", *Psychological Science*, SAGE Publications, vol.19, 2008.

Creative Computing for Decision Making: Combining Game Theory and Lateral Thinking

Lin Zou and Hongji Yang
Centre for Creative Computing
Bath Spa University
Bath
England

Abstract—Our research aims to improve the Traditional Decision Making Steps with efficiency and effectiveness. Creative Decision Making Steps are created through combining Game Theory and Lateral Thinking approach in an Enhanced Computer System. Computer system is able to help people make decisions as a result. In the current situation, computer system is a tool for human to output result after inputting data. Users could use the result to make decisions. In the past, data collection is limited to hardware. But big data and cloud computing change the situation, besides the infinite storage, the relationships among all the information are presented at the same time. Which means, if combine the creative methodology and logical process such as Game Theory and Lateral Thinking in computer system, computers will become more reliable in Decision Making.

Keywords—component; Creative Computing; Decision Making; Game Theory; Lateral Thinking

I. INTRODUCTION

Making a decision requires a logical and strict process. According to Pam Brown's study [1], there are seven steps in decision making, which are "Outline your goal and outcome, Gather data, Develop alternatives, List pros and cons of each alternative, Make the decision, Immediately take action to implement it, Learn from and reflect on the decision"

There are many decision making techniques for individual and group decision making, but there are various ways to make decisions. The first one is based on subjective experience, the other one is based on objective data analysis. A suitable decision usually combines subjective and objective factors, which makes a rigid computer system hard to achieve. Game Theory has developed into an umbrella term for the logical side of decision science, including both humans and non-humans. Therefore, it is possible to apply Game Theory and Lateral Thinking into a computer system to let computer system makes suitable decisions. The decision making is still based on a strict logical process, but the decision made by computer system will be influenced by Game Theory, however, the result will become more creative due to application of Lateral Thinking.

Our research aims to improve the efficiency and effectiveness for Decision Making in a way of Creative Computing. It also seeks new approaches which makes computer system become more creative. The development of Game Theory and Lateral Thinking has attracted a lot

attention. These two ideas are able to help Decision Making and Creative Computing to some extent. Therefore, how Game Theory and Lateral Thinking works should be evaluated firstly, and how Game Theory and Lateral Thinking could help in computing modelling would be considered as the crucial part of our research. According to the fundamental knowledge, Nash equilibrium and situation puzzle will be discussed as the scope. Also, Semantic web knowledge will be used for our research to examine what is creative and what is the creative in computing especially in technocratic paradigm rather than the rationalist paradigm and the scientific paradigm [2]

II. TRADITIONAL DECISION MAKING AND COMPUTER SYSTEM

Making decisions are essential components of life. For example, should a driver turn left or right on the road in order to take the shortest distance to destination, which stock should be bought for more profit to earn, does a patient need to see a doctor if he feels uncomfortable. These possible options are selected by human thought through a process which can be called as Decision Making Steps. A good decision will weight the positives and negatives to consider all the possibilities. Based on different situations, a good decision also requires particular decision techniques [3]. According to Dr. Pam Brown [1] Decision Making Theory, there are 7 steps in Decision Making. Recently, with the rapid development of computer science, whether computers can help human making decisions have been discussed more frequently [4]. Some steps have already been realised at computer platform, such as data gathering. Other steps have possibilities to be realised at computer platform in a conditional situation.

A. Traditional Decision Making Steps

Dr. Pam Brown has stated that there are 7 steps in Decision Making process, which are Outline the goal and outcome, Gather data, Develop alternatives, List pros and cons of each alternative, Make the decision, Immediately take action, and Review decision, every step in the decision making contains many factors, such as social, cognitive and cultural obstacles [1]. Traditional Decision Making Steps shows in Figure 1.

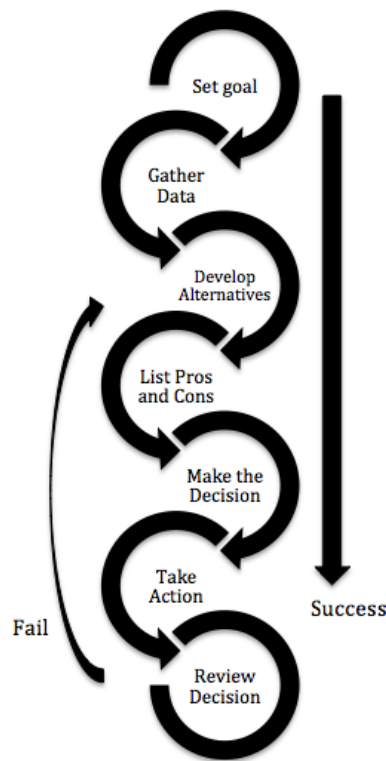


Figure 1. Traditional Decision Making Steps

- Step 1: Outline the goal and outcome. At first step, a person realises that he needs to make a decision. It is very crucial to clarify the nature and purpose of first thought through an internal process before a decision has been made.
- Step 2: Gather data. A precise decision requires relevant and detailed information to be collected before it is made. It is crucial to determine what information is useful, the way of gathering information and the source of acquiring information. This step requires think from internal to external, involves both qualitative and quantitative data.
- Step 3: Develop alternatives. This step could also be named as a brainstorm. A set of alternatives is discovered through analysis data gathered. Besides, people could also use experience to imagine new alternatives. By listing all the possibilities, a comprehensive view of action will be presented.
- Step 4: List pros and cons of each alternative. It is time to forecast each alternatives outcome and evaluate the positives and negatives for each decision after a list of alternatives drawn. Besides, it is necessary to determine whether the alternatives will meet the goal they set in Step 1 at the end. Ranking priority order of each alternative shows which decision has the higher potential for reaching best outcome. In real life,

individuals have different value system, the priority order will not be fixed.

- Step 5: Make the decision. After considerations in Step 4, a final decision will be made in Step 5. This decision could even be a combination of alternatives. Usually, decision will very likely to be similar to the top choice in the list which placed at Step 4.
- Step 6: Immediately take action. After decisions are made, people will start to take action.
- Step 7: Review decision. Results will be examined after taking actions. However, the decision might not effective immediately. The result could reach the goal, and also it has a possibility to fail. If a decision does not work, a new decision should be made by modifying the certain steps for previous decision.

B. Current Computer System Support on Traditional Decision Making Steps

Nowadays, computer is able to help in a wide range. Such as Economics, Physics, etc. Since 1987, a study of computer system to support clinical decision making has been proposed by Edward H. Shortliffe [5], with the rapid development of computer science, this approach is applied more and more often and not only in clinical situation. As a tool, computer system could be used to collect data, make statistics and calculation with no problem. According to Sprague study, Decision Support System has been defined [6], Decision Support system aims to gather and store the knowledge-based data, in the meantime, suitable for different environment users. This could be seen as using the computer system to do the Step 2 in Decision Making process.

Recently, Decision Making Software has been developed due to the demand of analysis large amount data. Such as 1000Minds, Expert Choice, and D-Sight. This software is not only using for data collection and analysis, but also has the ability to do the pairwise comparison, sensitivity analysis, and group evaluation. They open for users who interested in a particular area, for example, 1000Minds is finance based software, any users who would like to gain the advice for investment could be attracted by 1000Minds. This software is mainly used for supporting the Decision Making Process with Sensitivity Analysis and Fuzzy Logic calculations. The functions are similar to Step 3 in Decision Making Process.

Besides Decision Making Software, a search engine can also help in modern world due to the big data and a large amount of information stored in the cloud storage. Users share the relevant information each other make the search engine provide alternatives through keyword search. Still, this is also a similar function with Step 3.

C. Limitation of Traditional Decision Making Steps

Computer System has the ability to help Decision Making. However, Belton stated that Decision Making Software could be able to support decision making, but

software cannot replace decision making since it does not have the driving force [7]. This is also the limitation of computer system due to the nature of a binary system. A software engineer is required to apply strict logic when they are programming.

Fuzzy Logic becomes prevalent in programming to help computer system to solve more complexity problem for the accurate result. Fuzzy Logic has been applied on Decision Making Software already, but it has limitations such as decision maker can only gathering a small amount of information at one time [8].

Decision Making Software provides alternatives after users input the relevant information, but it assumes that people already have the clear aim. However, people also face problems with single goals and multiple goals [9], complicate think troubles users from the start of the traditional steps. In the meantime, the factors consideration is limited due to the lack of background knowledge provide. Therefore Traditional Decision Making Steps will provide an inaccurate result. Furthermore, people are hard to solve all problems through direct actions.

III. GAME THEORY AND LATERAL THINKING

Due to the development of Internet of Things and Semantic Web, it is possible to make computer system find out the connection between large amounts of information. Furthermore, computer system is able to gain the ability of forecasting after input relevant knowledge with suitable algorithm applied. For example, Game Theory and Lateral Thinking

Game Theory is a mathematical model that has been widely used in Cooperation Decision Making. It is generally accepted that the modern Game Theory in contemporary economics is introduced by mathematician Von Neumann and economist Oscar Morgenstern in 1950s. It is a principle that can be based on when analysing trade-offs in a conflict, and tries to determine the best outcome possible with the known conditions, both objectively and subjectively[10].

In modern world, Game Theory has been applied to economics, politics, and international strategy. Game Theory considers the separate individuals makes dependent decisions, different interaction could provide similar incentive structure [11]. Many types of Game Theory have been discussed based on different set, such as cooperative/non-cooperative, symmetric/asymmetric etc. Non-cooperative is developed widely by economists in recent years. Studies apply exclusively in the situation where interests of individuals not only conflict but also depend on the other's acts [12].

People work following the rules and a computer is following the rules made by human. However, not all the procedures have rules could be followed. People draw on experience to solve uncertain problems, and creativity is a result of uncertainty. It is hard for computers to gain experience and solve the problems in an indirect way.

On the other hand, Lateral Thinking aims to solve problems through a creative and indirect way. Bono has stated that it is important to think in a creative way, which

keeps distances from vertical logic or horizontal imagination[13]. Vertical logic represents a traditional way to solve problems. Horizontal imagination shows that people confused with many ideas. It could also be used for the purpose of making suitable decisions in different situations, also it provides a kind of acting strategies to computers. Furthermore, situation puzzle would also provide ability of improving the accurate forecasting of results. Situation puzzles contains a list of possible fitting answers, hosting puzzle player will answer the questioners with only "yes" or "no", to find out full aspect even with unexpected angle of one thing [14]. If applying Lateral Thinking in the Decision Making Steps, data collection followed way like situation puzzles, which can react itself with suitable changes. Furthermore, if Lateral Thinking could also help the computers could gain the ability to solve uncertain problems, it could say that computer get the creative reaction to do the job. This is a subjective ambiguity of human creativity which reconciled with the objective precision of computer systems [2].

A. Game Theory Approach in Decision Making Steps

Game Theory is able to change the Traditional Making Steps. Due to the significant logic method in Game Theory, Traditional Decision Making Steps could be reformed in a new way.

Combine the Game Theory algorithm and Traditional computer system through semantic web knowledge will result in an Enhanced Computer System. By establishing useful ontologies, users is able to suggest decisions through inputting data. New input information will be understood by enhanced computer system, a comprehensive analysis will be presented to help Decision Making Step 3 and Step 4. Furthermore, Enhanced Computer System will present a list of possible outcomes with forecasting such as Games results.

B. Lateral Thinking in Decision Making Steps

Besides Game Theory, Lateral Thinking also provides ideas for improving Steps 1 and 2. There will be more interactions between users and computers. At the start, instead of a clear goal, users could generate an idea. After users input conditions, Enhanced Computer System needs to provide a goal set and brief suggestion to help users know a clear goal. Furthermore, in Step 5 and 6, situation puzzles can help people to find out the exact aim and reason, which improve the possibility of success in Decision Making

The Enhanced System will add Lateral Thinking and Game Theory on Software Process and Software Engineering Tools and Method. For instance, people need credit rating for financing, the system will collect actions from people without showing the information to anyone else by using the zero-knowledge proof. The system will generate the score of credit rating to debtors. In the case, the human behaviour will influence the score, if a smoker lived in 40 years ago, the system would not take notice of smoking, but after 40 years, when a certain amount of feedback of tobacco use kills from data collection, the system would reconsider the calculation of score in order to meet the new environment. In the meantime, debtors and creditors could always be matched together who

demand profitable interests due to the nature of non-corporate Game Theory. In city transportation, the driver users could interact together and the computing result would generate the best choice of a road for each driver to avoid congestion. The users will be provided a reasonable plan when process creative computing with Lateral Thinking and Game Theory in a computing system, at least in competitive advantages.

IV. EXPERIMENT FOR NEW DECISION MAKING STEPS

It is possible to examine the applicability of Game Theory for creative computing through defining and making an explanation of Game Theory in different areas by using the semantic web and computer modeling techniques. Decision making depends on considering all interact factors rather than only objective precision but also subjective ambiguity. It is a problem that current computing systems do not play well. Game Theory could provide an adaptations idea in existing computing systems to meet the competitive environment.

A. Game Theory effectiveness in Decision Making Steps

There are varies games and models in Game Theory, and it is easy to understand the relationships between Game Theory and Decision Making Steps through Nash-Equilibrium. Nash-equilibrium is a concept that in a game consisting of players, an action profile of each player and a payoff function for each one, no individual player can gain a higher payoff by diverging singly from his or her profile [15]. This concept is widely used in predicting the outcome of a strategic interaction in the social science. For example, following the traffic light is a Nash Equilibrium game in real life. The conditions includes:

- In a perpendicular directions, there are two cars moves towards cross
- The traffic light will go red if the other one is green

If police will not fine the drivers, would they break the law to pass the crossing road while in a red light?

		DRIVER A	
		GO	STOP
DRIVER B	GO	-3, -3	1, 0
	STOP	0, 1	-1, -1

Figure 2. Traffic Light Games

In Figure 2, there are two drivers A and B, they have two options to be chosen when a red light display or green light display, which are "Go" and "Stop". As Game Preference, "-3" represents very bad outcome, driver crashes in accident; "-1" represents bad outcome, driver will stop and wait for infinite time; "0" represents acceptable, driver will stop when red light occur and move when green red display; then "1" means advantages,

driver moves without stop. Therefore there are 4 combination outcomes in this example:

- Driver A and Driver B will both choose go at the same time The result for two drivers will be both crashes in an accident.
- Driver A and Driver B will both choose stop at the same time. The result for two drivers will be waiting for an infinite time on the road, but fortunately they will not get hurt.
- Driver A will stop and Driver B will go. Driver A needs to wait until Driver B passes. Driver B gains the significant advantage in this situation which let him pass first. Driver A in a competitive disadvantage situation, however, Driver A will not wait infinite time and pass the crossing road without a crash. This is a Nash Equilibrium Outcome.
- Driver B will stop and Driver A will go. Driver B needs to wait until Driver A passes. Driver A gains the significant advantage in this situation which let him pass first. Driver B in a competitive disadvantage situation, however, Driver A will not wait infinite time and pass the crossing road without crash. This is a Nash Equilibrium Outcome.

Therefore, after outcomes analysis, it is easy to show the Nash Equilibrium Outcomes. Since both Drivers do not want to get crash accident in the situation, therefore, they will not choose to go when the other one is going. And Driver A choose is going, deciding to stop is best outcome for Driver B, and if B stop, Driver A's best outcome will be chosen to go, since both Driver A and B does not want to wait for infinite time. Driver A and B will not change the strategy, this is Nash Equilibrium.

Furthermore, in Decision Making Step view, this example presents a full analysis as step 3 and step 4, all possible alternatives are listed with the advantages and disadvantages compare. Then the case presents Nash Equilibrium results as the best outcome. In the end, example gives suggestions to take action as step 6. This shows that Game Theory can support Decision Making Steps effectively.

B. Combine Game Theory and Lateral Thinking with Computer System

Game Theory could provide an understanding of the nature of computation in both mathematics and modeling techniques for computing systems [16]. Similar criteria are studied in both distributed computing and Game Theory: dealing with systems where there are multi agents, facing uncertainty, and having possibly different goals [17]. However the difference is obvious in practice.

Based the understanding of Game Theory, Lateral Thinking and Computer System, we will begin with a research to study the applicability of enhancing existing computer system. Our research will start with a Quantitative Positivistic Method. Experimental Simulation will be chosen as a primary approach. This methodology employs a closed simulation model to mirror a segment of

the "real world". And it is better to follow the steps to gain scientific research result in creative software engineering.

- Collect and process real system data. Data such as a performance of an existing system, variables input should be collected for future use.
- Formulate and develop a model. Game Theory and Lateral Thinking are human subjects that exposed to a conceptual model, which needs to be simulated in an acceptable software form. Structured Equation Modeling will be suggested to analyse data. According to its advantages, Multivariate technique can be helpful to estimate a series of interrelated dependence relationships simultaneously by combining factor analysis and multiple regression. This can help to examine the dependence relationships and representing unmeasured concepts with multiple variables.
- Document model for future use. Experimental design such as assumptions, input variables, and objectives should be documented in detail.
- Select an appropriate experimental design. The design relies on a Two-level factorial design. It presents the design cannot interpret main effects of involved factors in isolation. The length and number of the independence runs should be considered in a different random number stream and the same starting conditions.
- Establish experimental conditions for runs. To obtain accurate information from each run in order to determine the performance measure of a system is changed or not over time.
- Interpret and present results. Computer numerical estimates such as mean and confidence intervals of the desired performance measure. The hypotheses of performance measure need to be tested.

C. Create Creative Decision Making Steps with Creative Computing

If Enhanced System works, it will change the Traditional Decision Making Steps into a Creative Decision Making Steps which shows in Figure 3. Though the Decision Making is a strict process, people still need creativity to solve uncertain problems. Therefore, Creative Decision Making Steps will combine the creativity process and strict process. Lateral Thinking and Situation Puzzle will provide the uncertainty and creativity of the result. Game Theory will provide the strict and logical process for Decision Making.

Figure 3 shows that in Creative Decision Making, first 3 steps replace the Traditional Decision Making Step 1. which brings the creativity for Enhanced Computer System to help users achieve creative thinking at the beginning of their Decision Making.

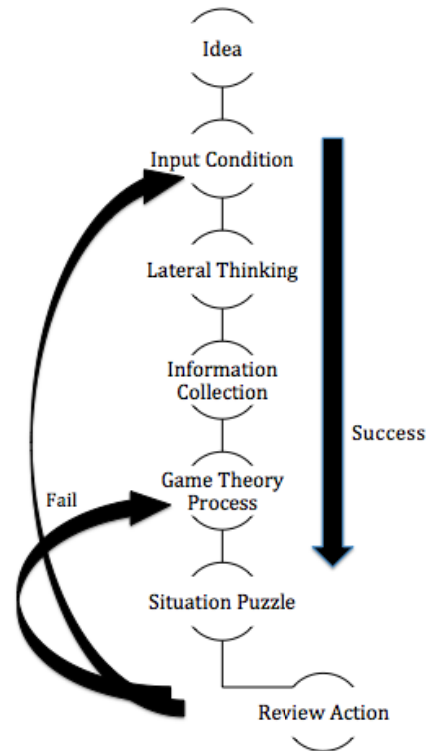


Figure 3. Creative Decision Making Steps

- Step 1: Start an Idea. The difference from Traditional Decision Making Steps is replace set the goal to raise an idea. This is because users might confuse what they want at the start. This is a very flexible and creativity approach. If users have a very clear goal set, they can skip this step.
- Step 2: Input Conditions. After users have an idea, there is a background research for users, system will consider more factors than Traditional Decision Making. In the meantime, the ideas from users raise in Step 1 will be used in this step to make Enhanced System prepare "Lateral Thinking".
- Step 3: Lateral Thinking. This step provides a set of goals which is able to help users clarify the goal and outcome. Users will have a brief perception from their future actions.
- Step 4: Collect Information. This step is same with Traditional Decision Making Steps. But there is a tiny difference since in the Enhanced System, the information requires is complex and more than Traditional Decision Making Steps. Probably, this steps demands Semantic Web technology to make Enhanced Computer System understands the questions and gathers more suitable result.

Creative Decision Making still needs a strict logic and process to make serious decisions for users. Therefore Game Theory Process replaced the Traditional Decision Making Steps 3 to 4. Nevertheless, this step includes 3 elements, Interaction Analysis, Define Preference and Outcome Set in the same time due to the nature of Game Theory.

- Step 5: Interaction Analysis. Not only users are considered in the Decision Making Process, Enhanced Computer System will use the background information to establish a game contains more players rather than just users. And then start an Interaction Analysis in this Step.
- Step 5: Define Preference. There will be four preferences defined in the meantime, which are Worst, Worse, Normal and Good. Preferences are used for comparing each other in order to see the competitive advantages.
- Step 5: Outcome Set. A list of outcomes will be presented. Users will have a comprehensive view on all the possibilities. There is no best outcome, but a more suitable outcome for different individuals.
- Step 6: Situation Puzzle. This step replaces the Taken Action in order to provide a creative result for Decision Making, through answering question, users are able to get predictions and forecast for each outcome show in Step 5.
- Step 7. Review Action. This step is same with Traditional Decision Making, however the return is return to Game Theory Process or Input Conditions, if users do not satisfy the result, he will first to change an outcome at Step 5, or they have to change the conditions to get more accurate decisions.

D. Potential Problems and Limitation of New Decision Making Steps

Though the perspective view of Enhanced Computer system is expectable. There are still some issues and limitations need to be considered. First of all, Nash Equilibrium has its own limitations in a real world. It requires further mixed strategy to make a suitable for unique situations such as Matching Pennies problem. Nevertheless, the database and feedback usually requires a large amount of computation which is costly and limited. And the data collections require an extended period to be analysed. Confidentiality agreement might occur in business information in further test. Apply Lateral Thinking interact system and self-improve system demands a complicated logic. The interactive proof system is complex. For Decisions made through Computer Support, it probably causes potential problems due to fuzzy responsibilities, which might cause legal issues in the future.

V. CONCLUSIONS

Our research will test the applicability of applying Game Theory and Lateral Thinking in creative computing to improve the Traditional Decision Making Steps through digital technologies especially for software engineering. Game Theory is a strict logic process, Lateral thinking is a creative thinking method. When they combine with each other, the result makes contribution especially in decision

making and computing systems. Though Lateral Thinking provides only 2 steps in Creative Decision Making Steps, it still makes strict decisions more creative, computer system will also gain benefit from a creative algorithm. These small steps change everything from Traditional Decision Making Steps, and make the result much more creative. Meanwhile, discussion on related limitation and new system provide a balanced perspective. It is also indicated that further process needed to be discovered.

REFERENCES

- [1] P. Brown, *Career Coach - Decision-Making*, Available: <http://www.pulsetoday.co.uk/career-coach-decision-making/10967084.article>, March 2015.
- [2] A. Hugill and H. Yang, "The Creative Turn: New Challenges for Computing", *International Journal of Creative Computing*, Inderscience Enterprises Ltd. vol. 1, no. 1, pp. 4-19, 2013.
- [3] V. Leicherova, and M. Januska, "Recommendations for the Selection of the Appropriate Decision-Making Style for the Selected Problem Situations Using the Vroom-Yetton-Jago Model", *Vision 2020: Innovation Development Sustainability Economic Growth*, Vienna, Austria, pp.908-920, 2013.
- [4] A. J. Maule, "Can Computers Help Overcome Limitations In Human Decision Making?", *Journal of Human-Computer Interaction*, Thomson Reuters, vol. 26, no. 2, pp. 108-119, 2010.
- [5] E. H. Shortliffe, "Computer Programs to Support Clinical Decision Making", *Jama*, AMA, vol. 258, no.1, pp. 61-66, 1987.
- [6] R. H. Sprague Jr, "A Framework for the Development of Decision Support Systems", *MIS Quarterly*, Management Information Systems Research Center, pp. 1-26, 1980.
- [7] V. Belton and T. Stewart, *Multiple Criteria Decision Analysis: An Integrated Approach: Science & Business Media*, Norwell: Kluwer Academic, 2002.
- [8] M. O'Hagan, "A Fuzzy Decision Maker", *Fuzzy Logic '93 Computer*, Fuzzy Logic Inc., 2000.
- [9] R. Grünig and R. Kühn, "Decision Problems", *Successful Decision-Making*, Berlin: Springer, pp. 7-14, 2013.
- [10] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, NJ: Princeton University Press, 2007.
- [11] R. J. Aumann and R. Serrano, "An Economic Index of Riskiness", *Journal of Political Economy*, University of Chicago Press, vol. 116, no. 5, pp. 810-836, 2008.
- [12] K. Sydsæter, A. Strøm, P. Berck, A. Strøm and P. Berck, *Economists' Mathematical Manual*, 3rd ed., New York: Springer, 2005.
- [13] E. De Bono and E. Zimbalist, *Lateral Thinking*, New York: Penguin, 2010.
- [14] P. Sloane, *Lateral Thinking Puzzlers*, New York: Sterling Publishing Company, Inc., 1992.
- [15] D. Fudenberg and J. Tirole, "Perfect Bayesian Equilibrium And Sequential Equilibrium", *Journal of Economic Theory*, Elsevier, vol. 53, no.2, pp. 236-260, 1991.
- [16] F. Moller and G. Struth, "Modelling Computing Systems: Mathematics for Computer Science", London: Springer, 2013.
- [17] J. Y. Halpern, "Computer Science and Game Theory: A Brief Survey", *Communications of the Acm*, ACM, vol.51, no.8, pp.75-79, 2007.

A Decision Tree based Recommendation System for Tourists

Preethiengburanathum (PhD student)^a, Shuang Cang^b, Hongnian Yu^c,
Faculty of Science and Technology, Bournemouth University, UK.
Email: ^aptheiengburanathum@bournemouth.ac.uk, ^cyuh@bournemouth.ac.uk
Faculty of Management, Bournemouth University, UK.
Email: ^bscang@bournemouth.ac.uk

Abstract—Choosing a tourist destination from the information that is available on the Internet and through other sources is one of the most complex tasks for tourists when planning travel, both before and during travel. Previous Travel Recommendation Systems (TRSs) have attempted to solve this problem. However, some of the technical aspects such as system accuracy and the practical aspects such as usability and satisfaction have been neglected. To address this issue, it requires a full understanding of the tourists' decision-making and novel models for their information search process. This paper proposes a novel human-centric TRS that recommends destinations to tourists in an unfamiliar city. It considers both technical and practical aspects using a real world data set we collected. The system is developed using a two-steps feature selection method to reduce number of inputs to the system and recommendations are provided by decision tree C4.5. The experimental results show that the proposed TRS can provide personalized recommendation on tourist destinations that satisfy the tourists.

Keywords: Recommendation System; Tourist Destination, Feature Selection; Filtering methods; Mutual information; Classification; Decision Tree

I. INTRODUCTION

The tourism industry is an extremely important sector on a global scale and contributed 9.5% to the total world's economy in 2013. It is expected that tourism will contribute around 10.3% GDP in 2023. South East Asia is expected to be the fastest grown regions in terms of its Travel and Tourism contribution to the GDP. In particular, Thailand, Indonesia, Singapore and Myanmar were identified as the countries possessing the most attractive tourism features in 2013 [1].

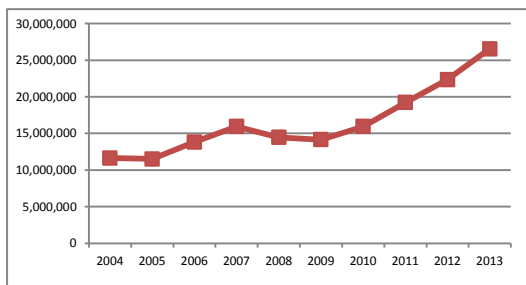


Figure 1. Number of international tourist arriving in Thailand from 2004- 2013 [1]

International tourist arrivals in Thailand have doubled over the past nine years (See Fig 1). In 2013, Thailand is the 10th most visited destination worldwide[1]. The country attracts 26.5 million international tourists grew by 18.76% over 2012 [2]. Increasing both tourist numbers (international and domestic) and the benefits from tourism are the primary objective of the Thai government. In 2013, tourism generated 1.79 trillion BHT (\$55.49 billion) in revenue for Thailand[2].

The Internet is now considered to be the main information source of tourists for information on products and services [3]. Due to the huge volume of heterogeneous information available on the Internet, the search for destinations, as known as travel planning can overwhelm tourists. The travel-planning task is complex and dynamic such that there are many factors involved when making a decision, for examples, the quality of the attractions, travel routes, hotels, numbers of traveler, leisure activities, weather, etc.[4]. Recently, tourism has substantially benefited from ICT, and especially from Internet technology [5]. With the development of decision support tools, also known as Recommendation Systems (RS), tourists and tourism providers can search, select, compare, and make decisions more efficient than ever.

Most of the previous TRSs have focused on estimates of choosing the destination, activities, attractions, tourism services (e.g. restaurants, hotels, and transportation) based on the user's preferences and interests. With regard to technical aspects, these TRSs only provide filtering, sorting and basic matching mechanisms between the items and the user's hard constraints. However, they are lacking in technical aspects (e.g. sparsity, scalability, transparency, system accuracy, theories to improve personalization, etc.) and practical aspects (e.g. user satisfaction, usability, etc.).

One of the greatest challenges in developing a TRS that provide personalized recommendations of tourist destinations is to enhance the tourist decision-making process. In order to achieve this, it requires a deep understanding of the tourists' decision-making and develops novel models for their information search process. Also, uncertainties involved in the information search stage of a tourist decision process need to be

eliminated. By reducing more parameters in the system, the model complexity could be decreased. In return, the recommendation performance and the level of user satisfaction of the system can both be increased.

This paper proposes a novel human-centric TRS that recommends destinations to tourist to solve the mentioned challenges. The proposed TRS is processed offline using the Data Mining (DM) process. This includes data acquisition, variables selection by using feature selection methods, decision making by using decision tree C4.5, and interpretation of the decision tree. The proposed TRS has three main innovations. Firstly, two feature selection methods are used to remove the unnecessary (both irrelevant and redundant) inputs into the system and to decrease the model complexity. Secondly, a decision tree C4.5 is used as a classifier to identify the tourist destination selection process. Lastly, the proposed system uses real world data that have been collected by us from Chiang Mai, Thailand.

The paper is organized into the following sections. Section 2 provides background on recommendation systems in the tourism domain. Section 3 describes the data collection process used in this paper. Section 4 presents the proposed TRS framework using the DM approach. The experiment setup for this study is demonstrated in Section 5. Section 6 shows the results and the evaluation analysis of the proposed TRS. Finally, we present some tentative conclusion and our future work in the last section.

II. BACKGROUND

A. Recommendation System

A recommendation system (RS), a subset of Decision Support Systems (DSS), is a tool that can recommend an item based on the aggregated information of the user's preferences [6]. It supports users by providing valuable information to assist them in their decision-making processes based on their priorities and concerns [7]. RS plays an important role and is common in many popular e-commerce websites, such as Amazon, Netflix, Pandora, etc. The e-commerce RSs suggest items to the user which involve news, articles, people, URLs, and so on [8].

B. Travel Recommendation Systems

Tourism is a leisure activity that involves complex decision processes, for example, selecting destinations, attractions, activities, and services. Thus, TRS attract the attention of many researchers from the fields of both academics and industry. Various TRS have been developed/deployed in and on many kinds of platforms (e.g. desktop, browser, mobile). TRSs recommend results to a user for the purposes of estimating user interest, choosing Points of Interests (POIs), identifying services or routes, ranking them in sequence, or as a holistic trip plan.

Most of the current TRSs aim to support an individual tourist, although there are some systems that support travel agencies as well [9]. They share similar

frameworks but differ in the selection of technology, theories to improve personalization, data inputs, interaction style, and recommendation techniques. Fig 2 shows the general framework of the recent TRSs. Information from various sources (e.g. sensors, GPS Coordinates, surveys, reviews, etc.) are integrated and kept in the repository (e.g. database schema, ontology).

The recommendation engine can be composed of several subsystems such as an optimization subsystem, a statistical subsystem and an intelligent subsystem and so on. This is to suggest, rank, or predict the items (i.e. destination, attractions, activities, and services) based on user requirements, preferences, or some hard and soft constraints (e.g. user demographic information, number of travel days, travel budgets, travel type, etc.).

Generally, before or during the trip, the TRS would take some inputs from the tourist (implicit, explicit, or both) to create a user profile and calculate the recommended result which is then sent back to the tourist. Tourists can visualize the results from the system in many ways, such as by destination icons on the map interface with a route between point-to-point, agenda, and itinerary. Most TRSs present the result with the use of spatial web services such as the Google Maps API service.

Lately, some TRSs are able to adapt the results to the user by taking the user context information (e.g. location, weather) into account. Some TRSs provide a functionality to let the user modify the generated result and adapt the results based on user feedback or user rating [10], [11].

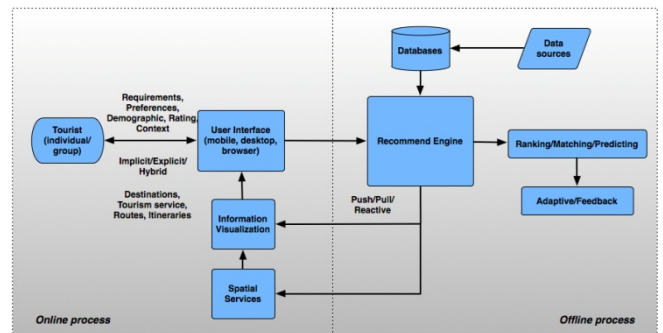


Figure 2. General framework of the travel recommendation systems

C. Recommendation techniques

According to [12], RS can be classified by the degree of personalization, including the usefulness and accuracy of the recommendations. The degree of personalization can be defined from low to high, including non-personalization, ephemeral personalization (short term), and persistent personalization (long term). The non-personalized RS is a fairly simple system that does not take the user preferences into account when making recommendations. For instance, the RS only generates a list of the most popular items based on the number of reviews or number of purchases (i.e., editor's choices or top-sellers). As a result, the recommended results would likely be of value to other generic users of the system. Due to their limited decision making power, non-

personalized systems have not been a focus of RS research [7].

Concerning the information incorporation related to the system users (i.e. user preferences, socio-demographic information, etc.), an ephemeral and personalized RS is more advanced than a non-personalized RS. In other words, every user would be able to see a different list of recommendations depending on his/her preferences. For example, Trip-advisor¹ recommends a destination based on the user's socio-demographic information. In fact, there are many types of personalized RSs that have been analyzed in previous studies, and the researchers have categorized them according to the method of the information-filtering techniques [7], [13]–[15]. In the next section, we will briefly investigate the recommendation engine (Fig. 3) which is composed of several recommendation techniques based on findings in [14]. The advantages and disadvantages of each type, and the hybrid filtering approach applied (i.e. the networking of several RSs) will be discussed.

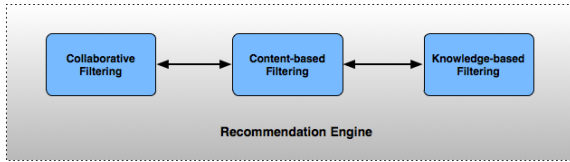


Figure 3. Recommendation Engine

a) *Collaborative filtering*: This approach is widely adopted by the most implemented recommendation systems. It recommends item(s) to the user based on the feedback of other users who share the same attributes, and suggest popular items to the user. This approach still suffers from a cold-start problem, where the new item or user would need to be rated before a recommendation can be made.

b) *Content-based filtering*: This recommendation technique suggests items to the user based on his/her previous searches or queries for items. The main drawback is the cold-start problem for the user, in which the user needs to provide a significant amount of information before the system can generate a recommendation. Otherwise, the system needs to have archived large historical data set in order to generate quality results [13]. Another common problem is over-specialization, since the system is most likely to suggest the item that the user liked the most, with less diversity among the recommendations [7].

c) *Knowledge-based filtering*: This technique recommends items to the user based on the knowledge of the domain. In other words, the system has some knowledge of how the particular item relates to a particular user. Predominantly, this technique can be achieved by using case-based reasoning or ontological methods. This recommendation technique can be found in [9] and [16], where the system exploits the travel agencies' and group expertise's past experiences.

d) *Hybrid filtering*: The above mentioned recommendation techniques have some strengths and weaknesses. The purpose of the hybrid recommendation technique is to achieve the best performance and to remove the weaknesses/disadvantages of one technique by complementing it with the advantages of another technique. Also, there are many hybridization methods, such as combining recommendation techniques together including weight, switching, mixed, feature combinations, cascades, feature augmentations, and meta-levels [13].

The latest Information and Communications Technology (ICT) e.g. Artificial Intelligent (AI), Semantic web, Communication network, etc. provides new opportunities for researchers to design and implement a TRS that is more intelligent, interactive, and adaptive, while being automatable, and supporting a higher degree of user satisfaction than ever before. We aim to develop a system to achieve those characteristics.

III. METHODOLOGY

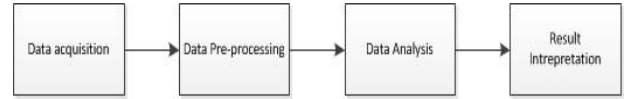


Figure 4. Data Mining Framework

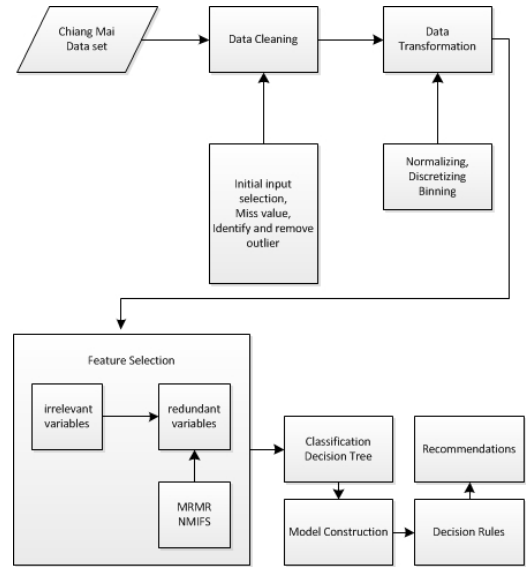


Figure 5. Methodology of the proposed destination TRS

The proposed DM framework shown in Fig. 4 consists of four phases including data acquisition, data pre-processing, data analysis, and result interpretation. (1) For data acquisition, the designed questionnaire, which has four parts, is distributed and collected from Chiang Mai, Thailand. (2) The collected data is pre-processed using several data pre-processing techniques involving data cleaning, data transformation, and feature selection methods. (3) The third phase involves the data analysis processes using a decision tree C4.5 as classifier. The aim of the third phase is to identify suitable features and find

¹ www.tripadvisor.com

the optimal models. (4) The final phase involves the interpretation of the obtained optimal decision trees and the extracted decision rules. The flow of the processes is described in Fig 5.

A. Data acquisition

To understand tourist's search behaviour in assessing travel information and decision-making processing for destination choice, we use a questionnaire as a data collection method due to its effective mechanism for collecting information from tourists. Pre-study on variety of factors that influence tourist's preferred destinations were identified for questionnaire design. The questionnaire design contains four parts containing a set of factors related to tourist's preferred destinations as following:

1) *Trip characteristics*: These variables are the most important variables when tourists select their destinations [17]. This includes trip length, travel purpose, trip composition, and etc.

2) *Tourist characteristics*: These variables include psychological, cognitive and socioeconomic status variables that influence on the tourist destination choice process [17].

3) *Travel motivations*: Travel or tour motivation is one of the important factors we have found from literature reviews when tourists are selecting their destinations. This variable describes the reason that a tourist chooses to visit a destination [18].

4) *Tourist sociodemographic information*: The individual demographics may influence the information seeking behaviour [19].

4,000 Questionnaires were distributed and collected at the five preferred tourist destinations in Chiang Mai, Thailand. The list of the preferred destinations was retrieved from the Trip-advisor website¹. The survey was distributed to both international (60%) and domestic tourists (40%). The destinations included Art in Paradise (27.7%), Mae Sa Waterfall (22.06%), Huay Tung Tao Lake (19.18%), Museum of World Insect and Natural Wonders (16.97%), and Bua Thong Waterfall (14.09%). The participants took 15-30 minutes on average to complete the questionnaire. 3,695 valid questionnaires with 145 variables were imported to data pre-processing stage, while 35 samples were rejected as being incompletely filled in.

The proposed framework uses variables extracted from questionnaire as inputs for classification of the tourist's preferred destination, including travel characteristics, tourist behavior, tourist expenditure behaviour, travel motivations, and tourist demographic information as described above.

B. Data Pre-processing

Real world data are generally incomplete, noisy, and inconsistent. For example, with surveys like ours, respondents may intentionally submit incorrect data because they do not want to submit personal information, or there may be data entry errors. The best classification

results require good quality data. To achieve this, we pre-processed the survey data through data integration, data cleaning, data transformation, and variable selection using feature selection methods.

Feature selection or variable selection is a process of selecting subsets of relevant features that describes the output classes. It is very important process for not only the utilization and usability, but also for accuracy improving. In this paper, we try to use small number of variables, which should contain the maximum information at the same time. In other words, it is to reduce the number of necessary user inputs as well as to increase performance of the classification model. In this paper, we propose a two-step filtering method based on Mutual Information (MI) to rank the features and remove irrelevant and redundant features from the dataset.

MI is used as a measurement in the feature selection process to characterize both the relevance and redundancy of the variables. If the variables were independent of each other, the MI value is zero. The greater the MI value, the more significant the dependent variable was. Given a set of X and Y , $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y , and $p(x, y)$ is the joint probability distribution function of X and Y . The MI for discrete variables is presented as:

$$MI(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

1) First filtering method

The purpose of the first filtering step is to rank the variables and remove any independent variables that are unrelated to the dependent variable. We applied a Max-Relevance feature selection algorithm [20], in which we chose MI as the measurement to remove the irrelevant features. We computed the MI score between each independent and dependent variable. Then, we ranked them in descending order and used a threshold value (the threshold value is chosen manually) to remove features that contributed less or were not related to the predictive power.

2) Second filtering method

In the second filtering step we used two mutual information-based, feature-selection algorithms: Minimum Redundancy Maximum Relevance (MRMR) [20] and Normalized Mutual Information Feature Selection (NMIFS) [21] to remove the redundant variables. The optimal feature space was chosen using the maximum MI G value. Feature selection stops when $G < 0$ is reached.

a) MRMR algorithm

The idea of the MRMR algorithm [20] is the algorithm using the MI value to rank the features based on the minimal redundancy and maximal relevant criterion. MRMR calculates redundancy for every pair of features

and calculates the relevance between the feature and the class. It is formulated as (1) below.

$$MRMR = \max_{i \in \Omega_s} [I(i, h) - \frac{1}{|S|} \sum_{j \in S} MI(i, j)] \quad (1)$$

b) NMIFS algorithm

NMIFS [21] is a modification of the MRMR algorithm (See (2) and (3)), it normalized the original MI value by the minimum entropy ($H(i)$ and $H(j)$) of both features as shown in the equation below.

$$MI2(i; j) = \frac{MI(i; j)}{\min\{H(i), H(j)\}} \quad (2)$$

$$NMIFS = \max_{i \in \Omega_s} [I(i, h) - \frac{1}{|S|} \sum_{j \in S} MI2(i, j)] \quad (3)$$

C. Data Analysis

Decision tree is chosen as a classifier/model for the proposed TRS because it provides several benefits for decision maker such as simplicity, interpretability. Decision-making is easily understood due to its flowchart-like model. For technical aspects, it handles the TRS's technical issues in terms of sparsity and scalability. The decision tree consists of nodes and leaves. The first node is called the root node, where the instances from the test set start to navigate down to a leaf. Other nodes, referred to as internal nodes, involve testing a particular attribute; this is where the split – either binary or multi – occurs. The leaf nodes represent a class label (i.e., the output of the classification) or the final decision of the instance from the test data. [22]. To recommend a destination to tourist, we must traverse the decision tree from the root to the leaf. Many decision trees exist, such as Hunt's algorithm, Top-down Induction of Decision Tree (TDIDT), ID3, CHAID, CART and C4.5. They differ in terms of splitting criteria, pruning, type of attributes, etc.

C4.5, an extension of ID3, was devised in [23]. It was chosen for this study because C4.5 tried to solve ID3 main drawbacks. ID3 [24] is the most simple decision tree algorithm but has many drawbacks such as that the optimal solution is not guaranteed, over-fitting problem with the training data set, and it supports only nominal variables. On the other hand, C4.5 supports two types of splitting criteria, including the information gain and the entropy-based criterion. It also supports both nominal and scale variables. In order to avoid the over-fitting problem, C4.5 supports tree pruning (e.g., confidence-based and error-based pruning). Moreover, C4.5 allows attributes to be missing.

IV. EXPERIMENT DESIGN

A. Representation of data set

Table 1 describes the characteristics of the data set used in this study. The data set contains five tourist's preferred destinations. However, the decision tree model that was constructed using all five destinations archived a very low of rate classification accuracy of 36.1%. In addition, the decision tree model was too complex such that it has a large tree size and number of leafs, which

makes it difficult to interpret for the decision-maker. To solve this problem, this multi-classes classification problem is divided into several sub problems by investigating the type of tourist's preferred destinations, combining the knowledge from Chiang Mai tourism domain experts and destination information from the trip advisor website.

Hence, the two categories were constructed and are presented in Table 2. The decision tree models were constructed based on these categories. The Museum data set presents a binary classification problem and the Nature data set presents a multi-classification problem. The Museum data set consists of two classes, as there both are specialized museum. The Nature data set consists of three classes, two of them represent the waterfall and one of them represents the lake.

TABLE 1. CHARACTERISTICS OF THE DATA SET USED IN THIS STUDY

Data set	# Features	# Classes	# Sample
Tourist destination choice	145	5	1,632

B. Data pre-processing

Initial selection is the first step for the process of cleaning the data. In this phase, knowledge acquired from tourism domains is used to select the features that are not related to output classes. Next, missing value analysis is performed for both data sets. Continuous variables were discretized using the binning method. The bin size is chosen as 10. Some of the discrete variables were normalized using tourism domain expert knowledge. After the data set had been cleaned and transformed, the proposed two-step filtering method was applied. This was done to remove the irrelevant and redundant features from the data set. For the first filtering step, different numbers of thresholds were used based on each data set to select between 17-18 relevant features. Then, MRMR and NMIFS feature selection algorithms were applied to the sub-set feature in order to remove the irrelevant features.

C. Classification and model construction

After the irrelevant and redundant features were filtered out, and the designated features were selected, we then constructed a classifier using a decision tree. An investigation of C4.5 performance from the two feature selection algorithms is carried out.

The K repeat holdout method was applied in this experiment. In each iteration, a 60% sample from each data set was randomly selected for training, 20% was used for validating, and the rest was used for testing, with stratification (i.e. each class has the same proportion in training, validation, and testing sets). The predictive accuracy of training, validating sets on the different iterations was averaged. Different configurations on confidence levels for decision tree pruning are used to find the optimal models for the two data sets. The confidence levels ranged from 0.1 to 0.5, with a step size

of 0.1. The optimal model is found when the following two conditions are met.

1. Best of mean of accuracy of validation sets.
2. Mean of accuracy of validation set must be equal or less than the mean of accuracy of training set.

V. RESULTS AND SYSTEM EVALUATION

Table 2 presents result of classification rate using C4.5. For single best learner, it can be seen that the Museum data set achieved a classification rate of 80%. The Nature data set revealed a classification rate of 49.72%. Regarding the performance of the two feature selection algorithms, the NMIFS algorithm is considered superior to the MRMR algorithm for both of the data sets.

TABLE 2. ACCURACY RATE FOR EACH DATA SET

Data set	# of classes	#Sample	Confidence level	Single best learner accuracy rate
Museum	2	729	0.39	80%
Nature	3	903	0.24	49.72%

Fig 6 shows the data pre-processing result from the Museum data set. Fig 6(a) presents the MI value from the first filter method, the threshold was set as 0.022, 128 variables were removed from the data set. Fig 6(b) shows the MI G values from both of the feature selection algorithms. Feature selection stopped when negative values were reached.

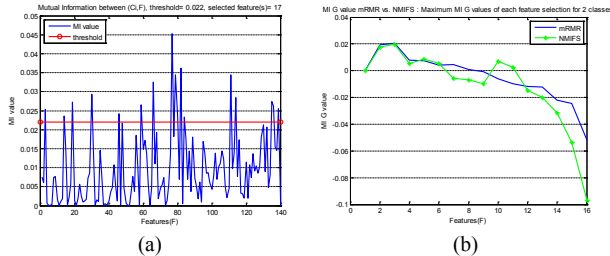


Fig 6. MI value (a) and MI G value (b) from the two-step feature selection method of the Museum data set

Table 3 presents the selected features from both of the feature selection algorithms of the Museum data set. The bold variables indicate that the corresponding feature belongs to the optimal subset. For the second filtering method, MRMR algorithm selected eight optimal features and NMIFS selected six optimal features for the Museum data set. It can be seen that feature *a* is the most important. This can be explained by the notion that one of the museums is specialized in insects. The feature *c*, *d*, and *b* were combined to help classify the data set. The optimal decision tree for the Museum data set is obtained and the decision rules are generated, combining four selected features from the NMFIS (See Fig 7 and 8). The obtained decision tree is viewed as being simple with the size of 17 and it has a number of leafs equal to 10. For the Nature data set, b2 (trip purpose) was selected as the most important feature.

TABLE 3. FEATURE RANKING BASED ON THE MRMR AND NMIFS ALGORITHMS (MUSEUM DATA SET)

Algorithm	Selected feature
MRMR	a c d b e f g h i j n k l m o p
NMIFS	a c d b g h f j i o k e n l m p

a:deepest impression is wildlife b:to visit place I have never been before c: The people who are companying are friend d: books and guides influences your decision to visit Chiang Mai

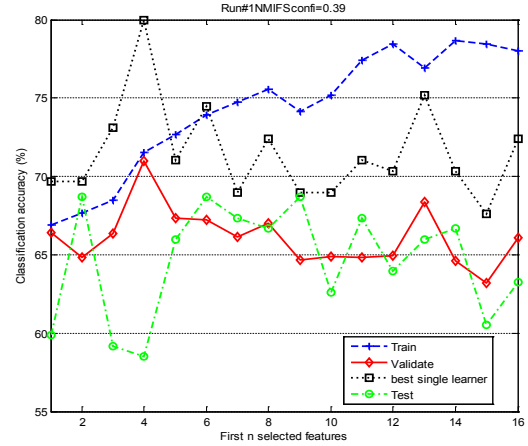


Figure 7. Accuracy rate for the Museum data set

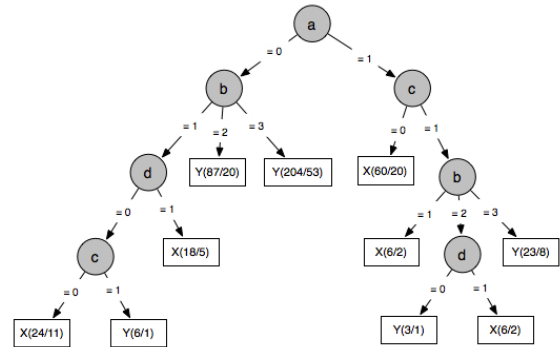


Figure 8. The optimal decision tree of the Museum data set using validation data. (X: Museum of world insects and Natural Wonders and Y: Art in Paradise, Chiang Mai 3D Art Museum)

Beside the accuracy rate, the confusion matrix is also used to evaluate the model's performance; it contains information regarding the actual and predicted classification done by the obtained optimal decision tree. According to Table 4, we can see that Museum of World Insects had a higher value of false positive (i.e. Museum of World Insects samples that were incorrectly classified as 3D Arts Museum samples).

TABLE 4. CONFUSION MATRIX OF THE MUSEUM DATA SET

		Predict	
		Museum of world insect	3D arts Museum
Actual	Museum of world insect	26	21
	3D arts Museum	8	90

To make it easier for a decision-maker to interpret the results, decision rules of the Museum data set are generated from the obtained optimal decision tree as shown in Table 5. There are eight rules generated for the Museum data set.

TABLE 5. THE DECISION RULES OF THE MUSEUM DATA SET

if $a == 0$, then
if $b == 1$ then
if $d == 0$
if $c == 0$ then, class = X;
elseif $c == 1$ then, class = Y;
end
elseif $d == 1$ then, class = X
end
elseif $b == 2$, then class = Y;
elseif $b == 3$, then class = Y;
end
elseif $a == 1$
if $c == 0$ then, class = X;
elseif $c == 1$
if $b == 1$, then class = X;
elseif $b == 2$
if $d == 0$ then, class = Y;
elseif $d == 1$, then class = X;
end
elseif $b == 3$, then class = Y;
end
end

VI. CONCLUSION

In this paper, a decision tree based tourist recommendation system has been presented in attempt of solving the current challenge of the destination TRS. The data set has been decomposed into two sub data sets using relevant tourism domain knowledge. This was done to increase classification accuracy rate and to reduce the complexity of the decision tree. The optimal decision trees from NMIFS with the highest accuracy rate and simplicity (i.e. less number of leaf and tree size) have been constructed for destination choice. The decision rules from decision trees were extracted. It can be seen that NMIFS is the optimum method because it uses fewer number of feature than MRMR for both of the data sets. Finally, the experimental results confirm applicable of the proposed a TRS. The proposed TRS satisfies the tourists' requirements who plan to visit or during their visit the city of Chiang Mai.

For future work, different types of classifiers can be considered to increase the classification accuracy rate for the data sets. Moreover, front-end web application and an interactive and adaptive user interface will be designed and implemented.

REFERENCES

- [1] "Economic Impact of Travel & Tourism 2014 Annual Update: Summary." World travel & tourism council.
- [2] "Thailand Annual Report 2013."
- [3] E. Pantano and L. D. Pietro, "From e-tourism to f-tourism: emerging issues from negative tourists' online reviews," *J. Hosp. Tour. Technol.*, vol. 4, no. 3, pp. 211–227, 2013.

- [4] B. Pan and D. R. Fesenmaier, "Semantics of Online Tourism and Travel Information Search on the Internet: A Preliminary Study," *Inf. Commun. Technol. Tour. 2002 Proc. Int. Conf. Innsbr. Austria 2002*, pp. 320–328, Jan. 2002.
- [5] E. Pitoska, "E-Tourism: The Use of Internet and Information and Communication Technologies in Tourism: The Case of Hotel Units in Peripheral Areas," *Tour. South East Eur.*, vol. 2, pp. 335–344, Dec. 2013.
- [6] G. Häubl and V. Trifts, "Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids," *Mark. Sci.*, vol. 19, no. 1, p. 4, Winter 2000.
- [7] F. Ricci, L. Rokach, and B. Shapira, "Introduction to Recommender Systems Handbook," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer US, 2011, pp. 1–35.
- [8] P. Resnick and H. R. Varian, "Recommender Systems," *Commun ACM*, vol. 40, no. 3, pp. 56–58, Mar. 1997.
- [9] G. I. Alptekin and G. Buyukozkan, "An integrated case-based reasoning and MCDM system for Web based tourism destination planning," *EXPERT Syst. Appl.*, vol. 38, pp. 2125–2132, 2011.
- [10] R. Anacleto, L. Figueiredo, A. Almeida, and P. Novais, "Mobile application to provide personalized sightseeing tours," *J. Netw. Comput. Appl.*, vol. 41, pp. 56–64, May 2014.
- [11] L. Sebastia, I. Garcia, E. Onaindia, and C. Guzman, "e-TOURISM: A TOURIST RECOMMENDATION AND PLANNING APPLICATION," *Int. J. Artif. Intell. Tools*, vol. 18, no. 5, pp. 717–738, Oct. 2009.
- [12] J. B. Schafer, J. A. Konstan, and J. Riedl, "E-Commerce Recommendation Applications," in *Applications of Data Mining to Electronic Commerce*, R. Kohavi and F. Provost, Eds. Springer US, 2001, pp. 115–153.
- [13] R. Burke, "Hybrid Recommender Systems: Survey and Experiments," *User Model. User-Adapt. Interact.*, vol. 12, no. 4, pp. 331–370, Nov. 2002.
- [14] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*. New York: Cambridge University Press, 2010.
- [15] M. Montaner, B. Lopez, and J. L. de la Rosa, "A taxonomy of recommender agents on the Internet," *Artif. Intell. Rev.*, vol. 19, pp. 285–330, 2003.
- [16] F. M. Santiago, F. A. López, A. Montejó-Ráez, and A. U. López, "GeOasis: A knowledge-based geo-referenced tourist assistant," *Expert Syst. Appl.*, vol. 39, no. 14, pp. 11737–11745, Oct. 2012.
- [17] D. R. Fesenmaier, K. W. Wöber, and H. Werthner, *Destination recommendation systems [electronic resource]: behavioural foundations and applications / edited by Daniel R. Fesenmaier, Karl W. Wöber, Hannes Werthner*. Wallingford, UK; Cambridge, MA: CABI Pub., c2006., 2006.
- [18] N. Leiper, "Tourist attraction systems," *Ann. Tour. Res.*, vol. 17, no. 3, pp. 367–384, 1990.
- [19] K. L. Andereck and L. L. Caldwell, "The Influence of Tourists' Characteristics on Ratings of Information Sources for an Attraction," *J. Travel Tour. Mark.*, vol. 2, no. 2–3, pp. 171–190, Feb. 1994.
- [20] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
- [21] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized Mutual Information Feature Selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.
- [22] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Revised edition edition. Amsterdam; Boston, MA: Morgan Kaufmann Publishers In, 2005.
- [23] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.
- [24] J. R. Quinlan, "Induction of Decision Trees," *MACH LEARN*, vol. 1, pp. 81–106, 1986.

Innovative Developments in HCI and Future Trends

Mohammad S. Hasan¹ and Hongnian Yu²

¹Faculty of Computing, Engineering and Sciences, Staffordshire University, UK

²Faculty of Science and Technology, Bournemouth University, UK

¹m.s.hasan@staffs.ac.uk, ²yuh@bournemouth.ac.uk

Abstract—The recent developments in technology have made noteworthy positive impacts on the human computer interaction (HCI). It is now possible to interact with computers using voice commands, touchscreen, eye movement etc. This paper compiles some of the innovative HCI progresses in the modern desktop and mobile computing and identifies some future research directions.

Keywords – *HCI, Virtual Reality, Augmented Reality, Haptic Feedback Controller, Smart Glass, Smart Lens.*

I. INTRODUCTION

The technological developments in the areas of desktop and mobile (portable) computing in the recent years have changed the Human Computer Interaction (HCI) quite significantly. Current desktop or laptop computers are equipped with speedy and vast processing capabilities e.g. multi-core processors with hyper threading, larger and faster main memory, powerful graphics cards, solid state drive (SSD) based secondary memory as well as built-in input-output devices e.g. web cam, sound card etc. Similarly, mobile devices e.g. smartphones, smart-watch, tablet computers with 8-core processor, 3GB or higher RAM, 32GB or higher SSD storage, multi-touch screen, high resolution cameras, global positioning system (GPS), Near Field Communication (NFC), sensors e.g. proximity, finger print, acceleration, barometer etc. are common nowadays. Many of these advancements have enabled these devices to process information in real time e.g. voice commands, human body pulse rate monitoring etc. Furthermore, the cost of such devices has dropped drastically because of the competitions among the leading manufacturers e.g. Apple, Asus, Dell, HP, Lenovo, Samsung etc. This paper reviews some of the latest developments in hardware and software for desktop and mobile computing HCI. It also attempts to identify future trends, research directions and challenges.

The rest of the paper is organised as follows. Sections II and III discuss some of the recent innovative developments in the desktop or laptop and mobile computing, respectively. And section III.O draws some conclusions.

II. DESKTOP AND LAPTOP COMPUTING

The desktop or laptop computers normally run operating systems such as Linux, Mac OS, Windows etc. The desktop HCI has developed in many areas e.g. affordable augmented and virtual reality, gaming, customer convenience for shopping etc.

A. Curved and Ultra High Definition (UHD) or 4K Resolution Displays

Curved monitors, shown in Figure 1, are already available as consumer products and are manufactured by Dell, LG, Samsung etc. These offer better viewing angle, less reflection, better 3D experience etc. However, these displays also have some drawbacks e.g. higher cost, wall hanging problem etc. [1]. The UHD or 4K resolution displays are able to produce 4 times more than the HD i.e. in the order of 4000 pixels horizontally for clearer and crispier viewing experience as shown in Figure 1. Many manufacturers produce displays that are curved and support UHD resolution.



Figure 1: Curved Display and UHD (4K) display [2].

B. Augmented Reality Headset - Microsoft HoloLens

Microsoft HoloLens is an Optical Head-Mounted Display (OHMD) that offers 3D augmented reality platform [3]. It blends holograms with reality where the user is able to design and customise holograms as shown in Figure 2. It is compatible to Windows 10 which is the next version of the Windows operating system.



Figure 2: The Microsoft HoloLens [3].

C. Virtual Reality Headset

Headsets such as HTC Vive, Oculus Rift etc. are very popular in the world of virtual reality, gaming etc. HTC vive carries 70 sensors, 360 degree head-tracking with a refresh rate of 90Hz to produce lower delay [4]. Oculus Rift [5], shown in Figure 3, offers low latency 360° head tracking capable of detecting subtle movements to produce natural experience, stereoscopic 3D View etc.



Figure 3: The Oculus Rift Virtual Reality Headset [5].

D. Virtual Mannequin and Virtual Fitting Room

Many online clothing retailers are now using the “Virtual Mannequin” technology to reduce the volume of returned goods because of wrong size or fitting [6]. The online shoppers enter some of their basic measurements and a virtual mannequin is generated [7] as shown in Figure 4. Another area of development is in-store Virtual Fitting Room. In this case the customer is able to try the clothing virtually in real time and very quickly without even going into a fitting room [8] as shown in Figure 5.

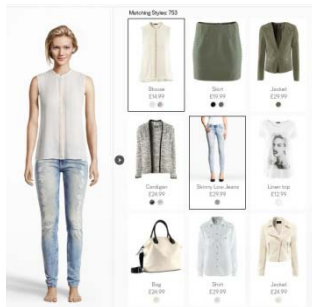


Figure 4: The Virtual Mannequin technology [7].



Figure 5: The Virtual Fitting Room [8].

E. Haptic Feedback Controller

Haptic feedback controllers produce realistic feedback e.g. force, vibration etc. to the user for virtual reality, gaming, tele-robotics, medical applications e.g. computer aided surgery etc. Many current game controllers already support this feature in various forms. For instance, the Steam Controller released by the game developer Valve, shown in Figure 6, offers two trackpads to deliver various physical sensations to the player [9]. Another example is Reactive Grip motion controller [10], shown in Figure 6, that can deliver motion and force feedback to the user using sliding contactor plates.



Figure 6: The Steam controller by Valve [9] and The Reactive Grip motion controller [10].

F. Brain signal Capturing Headset

A brain signal capturing headset is able to detect the changes in voltage when the human brain neurons are working on a thought. The headset normally carries a number of electrodes or sensors that are attached to the human scalp to record the electroencephalographic (EEG) signals and then these signals can be converted into digital form that can be processed by a computer [11]. For example, an epilepsy patient can pick a soft ball using a headset and a robotic arm which is shown in Figure 7. Such systems have been used to drive cars where the driver’s captured brain signals are processed by a laptop to drive the vehicle [12]. As an example, the Emotiv EPOC / EPOC+ headset [13] is quite popular in the research community. It offers a convenient brain computer interface with high resolution, 14 EEG channels and 2 references and is shown in Figure 8. The headset has been used to drive a taxi, a wheelchair [14] etc.

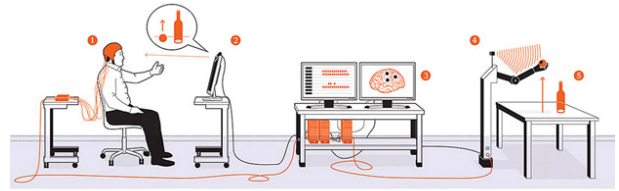


Figure 7: Controlling a robotic arm using a brain signal capturing headset [11].



Figure 8: The Emotiv EPOC / EPOC+ headset [13].

III. MOBILE COMPUTING

The mobile devices generally run operating systems such as Android, Bada, iOS, Tizen, Windows etc. The mobile computing HCI has developed in diverse directions and users can interact with modern devices in many ways e.g. touch, voice, heart pulse, body temperature etc.

A. Personal Assistant

Leading mobile device manufactures have introduced “Personal Assistant” e.g. Apple Siri, Google Now, Microsoft Cortana etc. as shown in Figure 9. These software tools use natural language user interface to interact with the user. The personal assistants can perform various tasks e.g. making phone calls, sending messages, scheduling meetings, launching browser, Internet searching, getting updated traffic information, obtaining weather forecast, answering questions etc.



Figure 9: The Apple Siri [15] and the Google Now [16].

B. Real Time Natural Language Translation Tools

Many smartphones are able to translate one language into another using translation tool. Recently Google has developed a real time translation tool based on images [17] which is shown in Figure 10 and supports a number of languages.



Figure 10: Real time natural language translation on a smartphone [17].

C. Ultrasound Fingerprint sensor

Many mobile devices e.g. smartphones use fingerprint scanner for user authentication. A new type of scanner, shown in Figure 11, has been developed that uses ultrasonic sound waves to scan fingerprints and is able to read finger prints through glass, metal and plastic smartphone covers [18]. It can even scan through sweat, hand lotion, condensation etc.



Figure 11: The ultrasound fingerprint sensor revealed by Qualcomm [18].

D. Curved and Flexible Displays

Mobile device manufacturers have launched devices with curved displays e.g. Samsung Galaxy S6 Edge [19] which can display notifications, text messages, weather information etc. on the curved edge for the convenience of the user as shown in Figure 12. Other areas of developments are flexible and transparent displays as shown in Figure 12 and Figure 13 that have been developed by Samsung, LG etc. Many of the flexible displays use polyimide film as the backplane. Polyimides are strong, flexible plastics that can achieve high degree of curvature by allowing a much thinner backplane than the conventional plastic [20]. Transparent displays with transmittance of 30% have already been achieved which is shown in Figure 12. These displays could be useful in advertising, security etc. applications.



Figure 12: Curved display for smartphone [19] and transparent display [20].



Figure 13: Flexible display [20].

E. Smart Glass e.g. Google Glass

Smart glass e.g. Google Glass is a smart phone like hands free device that is able to take voice commands as shown in Figure 14. It is a heads-up display (HUD) equipped with a camera, microphone and GPS etc. and can perform various tasks e.g. taking and viewing pictures, online searching, reading emails, satellite navigation, taking and making calls etc. [21]. However, Google has stopped commercial production of the Glass in 2015.



Figure 14: Smart glass e.g. Google glass and its application [21].

F. Google Cardboard

Google Cardboard [22] allows users to build a very low cost headset to experience virtual reality using smartphones as shown in Figure 15. As the name suggests, the Google Cardboard comes with cardboard, lenses, straps etc. The smartphone needs to run special application to create the stereoscopic view for both eyes. Various smartphones e.g. Apple iPhone, Google/LG Nexus, HTC Sensation, Huawei Ascend, LG G2, Optimus, Samsung Galaxy, Sony Xperia etc. are compatible to Google Cardboard.



Figure 15: The Google Cardboard [22].

G. Smart Contact Lens

Google has developed a smart contact lens prototype that can measure glucose levels in tears for diabetes patients. The lens is equipped with a tiny wireless chip, a miniaturised glucose sensor and a tiny LED that lights to indicate high glucose level [23], [24] and is shown in Figure 16.

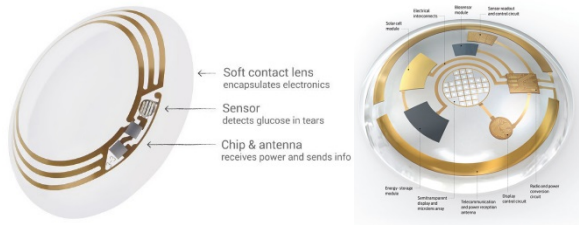


Figure 16: The Google's smart contact lens [23] and the Bionic Contact Lens for augmented reality [25].

H. Bionic Contact Lens

The bionic contact lens has an antenna to collect incoming radio frequency energy from a separate portable transmitter as shown in Figure 16. Solar energy can also be harvested to provide a boost to the lens. The total collected energy is used to power the internal circuits e.g. display to produce the augmented reality environment, communication with an external computer etc. [25]. The user can read emails, view images e.g. using the contact lens [26].

I. Smart Watch

Mobile devices are not limited to only phones, tablet etc. now. Various manufactures e.g. Apple, LG, Motorola, Samsung, Sony etc. have released smart watches as shown in Figure 17. These devices normally communicate with a smartphone using wireless technology e.g. Bluetooth to make or receive phone calls, display messages, notifications etc. Smartwatches can also play music, take photos, display stored photos, perform flight check-in, monitor human body fitness, run simple applications etc. without a smartphone [27], [28].



Figure 17: The iWatch by Apple [27] and the Galaxy Gear by Samsung [28].

J. Fitness Tracking devices

A number of fitness tracking devices e.g. Fitbit Charge HR, Fitbit Surge, Basis Peak, Jawbone UP Move, Swarovski Shine etc. are available nowadays [29]. These devices can monitor and display information about the user's heart rate, calorie burning e.g. running, walking, gym activities etc. Some fitness tracking devices e.g. Swarovski Shine, shown in Figure 18, combines the fashion and technology together as these can track running, cycling, swimming etc. activities and can be powered by solar energy.



Figure 18: The Fitbit Charge HR and the Swarovski Shine [29].

K. Smart Jewellery

HCI has also entered the world of fashion and jewellery in the form of Smart devices [30]. For example, the "CSR and Cellini Bluetooth Pendant", shown in Figure 19, is capable of connecting to a smartphone and is able to change its LED colour or brightness depending on the mood or clothing of the user. Furthermore, it can also generate notification of incoming calls, emails or text messages to the user by changing LED colour or flashing or vibrating. Another example is the "June Bracelet or Brooch", also shown in Figure 19, which looks like a diamond in a metal e.g. platinum, gold etc. or on leather strap or it can be worn as a brooch. The device feeds data to a smartphone and shows summary of the user's exposure to the sun. It produces UV index, weather forecast etc. for sunscreen, sunglasses etc. to protect the user's skin from sun damages or premature wrinkles.



Figure 19: The CSR and Cellini Bluetooth Pendant and the Netatmo June Bracelet or Brooch [30].

L. Smart Shoe, Bluetooth Insoles, Smart Sock

Smart shoe, shown in Figure 20, generates power using two devices – a "shock harvester" and a "swing harvester" which produce power when the heel hits the ground and when the foot is swinging, respectively [31]. The generated power is three to four milliWatt (mW) and can be used to power sensors and an antenna. One of the applications of the shoe is indoor navigation and rescue operation. Similarly Bluetooth insoles, shown in Figure 20, are equipped with a number of sensors, accelerometers etc. and can monitor activity levels, walking health issues, therapy progress [32], communicate with smartphones to give directions using vibrations [33] etc. Another area of development is Smart sock which is shown in Figure 21. The smart sock from Sensoria has embedded sensors and is able to produce feedback on the running techniques via smartphone application for a sportsperson [34].



Figure 20: The smart shoe capable of generating power [31] and the Bluetooth insole [32].



Figure 21: The smart sock from Sensoria Fitness [34].

M. Habit Changing Wristband

Technology can help us to get rid of bad habit as well. For example, the Pavlok, shown in Figure 22, can be

programmed e.g. visiting time-wasting websites, launching a maximum number of tabs in the browser etc. and it will generate an electric shock for the user to remind of bad habit [35].



Figure 22: The Pavlok wristband [35].

N. Baby monitors

A number of smart devices are available to monitor baby activities as shown in Figure 23. For example, the Sensible Baby SmartOne monitor [36] and the Owlet Vitals Monitor [37] etc. allow parents to monitor baby's position, movement, body temperature, heart rate, oxygen level etc. on a smartphone or tablet in real time.



Figure 23: The Sensible Baby SmartOne monitor [36] and the Owlet Vitals Monitor [37].

O. Nervous system or brain controlled Robotics

It is now possible to read the nerve signals of a human body and to control an artificial leg or arm. For example, a patient who lost the lower part of his leg has been fitted with a bionic leg that is controlled by his thoughts [38] which is shown in Figure 24. Another example is brain controlled robotics where the brain signals are used to control a robotic body part e.g. an arm. For example, a patient's body is paralysed from the neck down. Two sensors implanted in his brain can now monitor the brain activities and control a robotic arm. The patient can shake hands of another person, lift a drink, control a computer mouse etc. [39].



Figure 24: First thought-controlled prosthetic leg [38].

IV. CONCLUSION

Latest HCI innovations have made many technologies e.g. virtual reality, personal digital assistant, biometric authentication e.g. finger print scanner etc. available to us and have made our lives convenient and secure. Nowadays, we can monitor our health, manage our daily schedules, navigate from one place to another etc. using different forms of hardware and software HCIs. In the

desktop computing HCI, many areas e.g. real time translation of one natural language into another, object recognition from images, low cost virtual studio, 3D scanner to produce 3D printed models etc. can be identified as the future researches. On the other hand, for mobile computing, e.g. eye movement tracking, hand gesture recognition, obtaining heart rate by putting a finger on the touchscreen for some time, starting a vehicle remotely, feedback on driving behaviour, augmented reality headset for vehicle e.g. engine maintenance etc. can be identified as future researches.

REFERENCES

- [1] "Curved Monitors – Pros and Cons | Ebuyer Blog," Ebuyer, 05-Jan-2015. [Online]. Available: <http://www.ebuyer.com/blog/2015/01/curved-monitors-pros-and-cons/>. [Accessed: 01-Apr-2015].
- [2] "Infographic: High End Display PPI Showdown | Infogram," 21-Oct-2013. [Online]. Available: <https://infogr.am/High-End-Display-PPI-Showdown>. [Accessed: 12-Apr-2015].
- [3] "Microsoft HoloLens," Microsoft HoloLens. [Online]. Available: <http://www.microsoft.com/microsoft-hololens/en-us>. [Accessed: 01-Apr-2015].
- [4] "The best VR headsets," Wareable, 06-Mar-2015. [Online]. Available: <http://www.wareable.com/headgear/the-best-ar-and-vr-headsets>. [Accessed: 01-Apr-2015].
- [5] "Rift," Oculus VR. [Online]. Available: <https://www.oculus.com/rift/>. [Accessed: 01-Apr-2015].
- [6] R. Preston, "Virtual mannequins' promise better fit for online shoppers," BBC News. [Online]. Available: <http://www.bbc.co.uk/news/technology-25812130>. [Accessed: 12-Apr-2015].
- [7] "Virtual Fitting Rooms - are they a gimmick or an essential tool for fashion ecommerce? | App Commerce Platform," Poq Studio | App Commerce Platform. [Online]. Available: <http://poqstudio.com/2013/02/virtual-fitting-rooms-are-they-a-gimmick-or-an-essential-tool-for-fashion-ecommerce/>. [Accessed: 12-Apr-2015].
- [8] "Virtual Fitting Room Courtesy of the Kinect," Daily Bits. [Online]. Available: <http://www.dailybits.com/virtual-fitting-room-courtesy-of-kinect/>. [Accessed: 12-Apr-2015].
- [9] "Valve reveals haptic game controller for release in 2014," BBC News, 27-Sep-2013. [Online]. Available: <http://www.bbc.co.uk/news/technology-24304272>. [Accessed: 01-Apr-2015].
- [10] "Products | Tactical Haptics," Tactical Haptics. [Online]. Available: <http://tacticalhaptics.com/products/>. [Accessed: 01-Apr-2015].
- [11] "How to Catch Brain Waves in a Net," IEEE Spectrum, 21-Aug-2014. [Online]. Available: <http://spectrum.ieee.org/biomedical/bionics/how-to-catch-brain-waves-in-a-net>. [Accessed: 01-Apr-2015].
- [12] "Look, no hands (or feet): Scientists develop car that can be driven just by THINKING," Mail Online. [Online]. Available: <http://www.dailymail.co.uk/sciencetech/article-1359512/Computer-scientists-Germany-invented-car-steered-power-thought.html>. [Accessed: 01-Apr-2015].
- [13] "Emotiv EPOC / EPOC+," Emotiv. [Online]. Available: <http://emotiv.com/epoc.php>. [Accessed: 02-Apr-2015].
- [14] "Emotiv Claims Its Brainwave Scanner Allows People to Control Wheelchairs with Their Minds," Wearable Tech World, 03-Feb-2014. [Online]. Available: <http://www.wearabletechworld.com/topics/from-the->

- experts/articles/369076-emotiv-claims-its-brainwave-scanner-allows-people-control.htm. [Accessed: 01-Apr-2015].
- [15] "Siri, Your wish is its command," *Apple*. [Online]. Available: <https://www.apple.com/uk/ios/siri/>. [Accessed: 02-Apr-2015].
- [16] "Google Now. The right information at just the right time.," *Google*. [Online]. Available: <https://www.google.co.uk/landing/now/>. [Accessed: 02-Apr-2015].
- [17] "Pardon? Testing Google's speedy translation tool," *BBC News*. [Online]. Available: <http://www.bbc.co.uk/news/technology-30824033>. [Accessed: 07-Apr-2015].
- [18] L. Kelion, "Fingerprint sensor revealed by Qualcomm at MWC," *BBC News*. [Online]. Available: <http://www.bbc.co.uk/news/technology-31692988>. [Accessed: 07-Apr-2015].
- [19] "Galaxy S6 & Galaxy S6 edge - Pre-order online." [Online]. Available: <http://www.samsung.com/uk/galaxys6/>. [Accessed: 02-Apr-2015].
- [20] S. Anthony, "LG's flexible and transparent OLED displays are the beginning of the e-paper revolution," *ExtremeTech*. [Online]. Available: <http://www.extremetech.com/computing/186241-lgs-flexible-and-transparent-oled-displays-are-the-beginning-of-the-e-paper-revolution>. [Accessed: 03-Apr-2015].
- [21] "Google Glass UK release date, price and specs: Google Glass will go on sale on 19 January; how to buy Google Glass," *PC Advisor*. [Online]. Available: <http://www.pcadvisor.co.uk/features/gadget/3436249/google-glass-release-date-uk-price-specs/>. [Accessed: 03-Apr-2015].
- [22] "Google Cardboard, Experience virtual reality in a simple, fun, and inexpensive way," *Google*. [Online]. Available: <https://www.google.com/get/cardboard/index.html>. [Accessed: 03-Apr-2015].
- [23] "Inside Google X's Smart Contact Lens," *Re/code*. [Online]. Available: <http://recode.net/2014/01/16/inside-google-xs-smart-contact-lens/>. [Accessed: 03-Apr-2015].
- [24] "Google unveils 'smart contact lens' to measure glucose levels," *BBC News*. [Online]. Available: <http://www.bbc.co.uk/news/technology-25771907>. [Accessed: 03-Apr-2015].
- [25] "Augmented Reality in a Contact Lens," *IEEE Spectrum*. [Online]. Available: <http://spectrum.ieee.org/biomedical/bionics/augmented-reality-in-a-contact-lens/eyesb1>. [Accessed: 03-Apr-2015].
- [26] M. Roberts, "Bionic contact lens 'to project emails before eyes,'" *BBC News*. [Online]. Available: <http://www.bbc.co.uk/news/health-15817316>. [Accessed: 03-Apr-2015].
- [27] "Apple (United Kingdom) - Apple Watch," *Apple*. [Online]. Available: <https://www.apple.com/uk/watch/>. [Accessed: 02-Apr-2015].
- [28] "Samsung Galaxy Gear Watch (Jet Black) - 1.63" Super AMOLED, 1.9MP, 4GB," *Samsung UK*. [Online]. Available: <http://fb.uk.samsung.com/consumer/mobile-devices/wearables/gear/SM-V7000ZKABTU>. [Accessed: 02-Apr-2015].
- [29] "Best fitness trackers 2015: Jawbone, Misfit, Fitbit, Garmin and more," *Wareable*. [Online]. Available: <http://www.wareable.com/fitness-trackers/the-best-fitness-tracker>. [Accessed: 03-Apr-2015].
- [30] S. Hill, "Smart jewelry is proving wearable tech doesn't have to be hideous," *Digital Trends*. [Online]. Available: <http://www.digitaltrends.com/mobile/smart-jewelry-roundup/>. [Accessed: 03-Apr-2015].
- [31] P. Rincon, "Smart shoe devices generate power from walking," *BBC News*. [Online]. Available: <http://www.bbc.co.uk/news/science-environment-30816255>. [Accessed: 05-Apr-2015].
- [32] "Watch your step -- with a Bluetooth-connected insole," *CNET*. [Online]. Available: <http://www.cnet.com/uk/news/watch-your-step-with-a-bluetooth-connected-insole/>. [Accessed: 06-Apr-2015].
- [33] L. Kelion, "CES 2015: Preview of the new tech on show in Las Vegas," *BBC News*. [Online]. Available: <http://www.bbc.co.uk/news/technology-30643396>. [Accessed: 06-Apr-2015].
- [34] "Sensoria Fitness Smart Sock Preview," *CNET*. [Online]. Available: <http://www.cnet.com/uk/products/sensoria-fitness-smart-sock/>. [Accessed: 06-Apr-2015].
- [35] "Pavlok Uses Mild Electric Shock To Help You Break Any Habit," *Pavlok*. [Online]. Available: <http://pavlok.com/>. [Accessed: 06-Apr-2015].
- [36] "No more tiptoeing around," *Sensible Baby*. [Online]. Available: <http://mysensiblebaby.com/>. [Accessed: 06-Apr-2015].
- [37] "Owlet Vitals Monitor," *Thing Alive*. [Online]. Available: <http://thingalive.com/owlet-vitals-monitor>. [Accessed: 06-Apr-2015].
- [38] "NEJM: First thought-controlled prosthetic leg," *MedCity News*. [Online]. Available: <http://medcitynews.com/2013/09/nejm-first-thought-controlled-prosthetic-leg/>. [Accessed: 02-Apr-2015].
- [39] J. Gallagher, "Brain-reading implant controls arm," *BBC News*, 22-May-2015. [Online]. Available: <http://www.bbc.co.uk/news/health-32784534>. [Accessed: 17-Jun-2015].

Message Dissemination Reliability in Vehicular Networks*

Elias Eze C^{1,2,*}, Sijing Zhang^{1,2,†}, Enjie Liu^{1,2,†}

1. Centre for Wireless Research, Institute for Research in Applicable Computing (IRAC)
 2. Department of Computer Science and Technology, University of Bedfordshire, Luton, England
- *elias.eze@study.beds.ac.uk, †{sijing.zhang, enjie.liu}@beds.ac.uk

Abstract— Prior to wide deployment of Vehicular Ad-Hoc Networks (VANETs) in public motorways, amongst the key challenges that must be adequately resolved is safety message dissemination reliability in the presence of error-prone wireless channel, propagation delays, extremely dynamic network topology, frequent network fragmentation and high vehicle mobility. In this paper, we investigate the application of network coding concept to achieve improved message dissemination reliability as well as increased bandwidth efficiency for efficient vehicular communication system. In particular, we proposed an efficient error recovery scheme which employs network coding concept to increase safety message broadcast reliability with a minimized number of encoded packet retransmissions. The benefits of the proposed scheme over simple error correction based on retransmission are clearly shown with detailed theoretical analysis and further validated with simulation experiments.

Keywords— VANET, Broadcast, DSRC, PHY, MAC, XOR

I. INTRODUCTION

Broadcast technique is an important message dissemination mechanism in vehicular communication systems for disseminating identical data information in a one-to-many scenario. It has been widely utilized in many network applications such as device configuration and content delivery [1-2]. This technique is majorly used in a case where it is required that all the information sent by the source node must be correctly received by every intended recipient node. However, wireless links envisioned for vehicle-to-vehicle (V2V) communications are more error-prone than their wired counterparts. Therefore, it is necessary to use some efficient error-recovery schemes to provide some measures of reliability guarantees. The most widely employed error recovery technique for handling packet losses in the network transmission is Automatic Repeat reQuest (ARQ), which uses acknowledgements and timeouts for the loss packet recovery. Though ARQ is very efficient in the unicast communication scenario, it is not as efficient in the broadcast transmissions as required in a saturated one-to-many broadcast communication of vehicular communication systems.

While several research results have been published in different areas of VANETs one intrinsic area that still poses a significant challenge is the packet transmission reliability of V2V communication. The extreme dynamic topology,

intermittent connectivity and high mobility in VANETs wireless environment often lead to possibility of losing data packets meant to deliver life-saving information, thereby making communication reliability a challenge. The traditional problem of reliable packet transmission in wireless channel has been widely discussed by several researchers with many approaches adopted like RTS/CTS, BRTS/BCTS and ACK/N-ACK [3-4]. However, these classical approaches for improving reliability only works efficiently with one-to-one, unicast communication unlike VANETs where road safety relies on each node transmitting its status to all nearby nodes within its transmission range resulting to many one-to-many, broadcast communication. Hence, relying on conventional error recovery mechanisms to ensure reliability in vehicular networks would worsen the problem rather than improving reliability as they will lead to more packet collisions.

Several research results have also been published on retransmission-based loss recovery mechanisms (especially repetition-based error recovery) [5-6] for VANETs during exchange of time-sensitive, safety-related information. The key aim of repetition-based error recovery technique is to allow each node to repeat the transmission of its raw packet(s) within the timeout period thereby giving the nodes within their transmission range multiple chances of receiving the packets not correctly received. Though retransmission-based loss recovery techniques increases the chances of recovering packets not received or correctly received, the repeated packets consume a substantial amount of the channel bandwidth thereby giving rise to excessive increase of network overhead and further loss of packets due to network congestion. Since retransmission increases network overhead and after a given number of consecutive repeats may lead to packet collision due to channel congestion, this paper investigates the possibility of reducing the number of retransmissions while enhancing the efficiency of a combined retransmission to enable all the vehicles within the transmission range to receive the combined original packets without error using the network coding technique.

Network coding [7] is a concept originally designed to enable routers to intelligently mix different packets from different sources in order to increase the information content of each transmission as well as increase the overall network throughput. This paper exploits the manifold benefits of network coding to achieve reliable and efficient communication in vehicular networks by proposing a scheme called CODE-aided Error Recovery (CODER) scheme. The proposed scheme enables each vehicle to perform an exclusive OR operation on

*This work is supported by a Grant-in-Aid for Scientific Research from Ebonyi State Government (EBSG) (No. EBSG/SSB/PS/VII/105)

its *packet pool*² and to retransmit the XORed packets to other vehicles within their vicinity. Broadcasting the XORed version instead of the raw packets creates ample opportunity for high rate of lost packets recovery for each repetition.

The remaining part of the paper is arranged as follows: Section II presents the literature review. Section III introduces the proposed CODER scheme, followed by an analysis of packet recovery probability (PRP) as a function of the total packet loss probability (PLP) in Section IV. Analysis of Packet Collision Probability (PCP) is presented in Section V while Section VI presents the Location-aware algorithm (LAA). Section VII presents the scheme validation and analysis. Section VIII discusses the analytical and simulation results, while Section IX concludes the paper.

II. LITERATURE REVIEW

Li *et al* [8] followed the pioneering work of [7] on network coding with their own work showing that linear codes can be used to achieve the maximum capacity bounds for multicast traffic. Many researchers have built on this foundation to study network coding with omnidirectional antennae and showed that the challenge of minimizing the communication cost (high bandwidth usage) can be expressed as a linear program that can be solved through a distributed approach. Network coding has been employed several times for applications relying on packet broadcasting, most of which are basically for disseminating information across the entire network. Contrarily, VANETs are different because their periodic status and emergency broadcast alerts are exceptionally relevant within the limited radio transmission range that the vehicles' radio or omnidirectional antennae signal can cover. There are existing solutions that can efficiently recover lost packets in one-hop wireless broadcast retransmissions through network coding [9-10]. However, these schemes are not applicable in VANETs because the packets that are XORed for retransmission are generated from a single node over a period of time. In VANETs, all nodes broadcast their packets containing their real-time status where the most recent packets supersede the previous packets due to change in position and other kinematic data. Hence, only currently broadcasted packets are coded for retransmission.

The concept of repetition-based error recovery is applicable to various medium access control (MAC) protocols including slotted ALOHA, TDMA and CSMA [11]. Though the most widely recognized MAC protocol for wireless access in vehicular environments is the CSMA-based IEEE 802.11p, different TDMA-based MAC protocols [12] for VANETs have been studied. These protocols improve the conventional repetition-based retransmission scheme by optimizing the process of nodes time-slot selection. However, more realistic scenario is studied in our work where mobile nodes are always located in a 5x5 Manhattan grid road network which allowed our proposed scheme to adopt CSMA-based protocol.

Recent works by Wang and Hassan [13] and Wu *et al* [14] studied a different type of repetition-based error recovery techniques. In [14], the nodes located at the edge of the radio transmission range relay the raw packets upon reception to nodes far away from the senders. The relay approach is shown

to increase the broadcasted packets coverage range as well as ensure reliable transmission of the packets. However, the transmission of time-constraint safety information in VANETs falls under one-hop broadcast communication mode. A few existing coded retransmission error recovery techniques applicable to vehicular communications that would fall in the category of the CODER scheme proposed in this paper are the research carried out in [14][15]. The protocols they proposed are basically for performing bit-wise exclusive OR operation on two packets, which is retransmitted as a single packet. Given their results, it was clear that their proposed scheme, though without feedback mechanism, out-performed a simple repetition-based error recovery technique in [5]. Our proposed scheme presents an in-depth analytical model of a loss recovery scheme that is not limited to merely two packets but can combine n -packets received from other vehicles and self-generated raw packets.

III. CODE-AIDED ERROR RECOVERY (CODER) SCHEME

A typical vehicular traffic scenario where every vehicle periodically (or upon the event of emergency) transmits and receives broadcasted packets to and from other vehicles within their radio's signal transmission range. Due to the harsh condition of the motorway, high mobility of the vehicles, intermittent connectivity and extremely dynamic network topology of VANETs, some of the broadcasted packets are not reliably received as a result of congestion and collision in the wireless medium. Traditionally, recovering these lost packets will entail that each sender retransmits its raw packets until all the recipients recover every lost packets. Hence, instead of the source nodes simply repeating their various raw packets, our proposed scheme uses a location-aware algorithm (LAA) (see Fig. 2) to determine the closest vehicle to the source node with the highest rate of packet reception probability to perform an exclusive OR (denoted with \oplus) on all the received packets and its own raw packets. The same node retransmits the encoded packets to enable all the vehicles within its radio signal coverage receive, decode and recover any packet(s) lost. The intuition behind the scheme is exemplified in Fig. 1, where vehicle D performs an exclusive OR operation on the packets (P_a, P_b, P_c, P_e, P_f , and P_g) received from vehicles (A, B, C, E, F , and G) and its own raw (internally generated) packet (P_d) and retransmits the encoded packet to enable other vehicles within its one-hop broadcast range recover their lost packet(s). Each node can either broadcast its own generated packets or the XORed packets from node i with node $i + 1$ and or node $i - 1, \{i \oplus (i + 1)\}$ or $\{(i - 1) \oplus i\}$ (so as to accommodate vehicles moving in opposite direction).

I. ANALYSIS OF PACKET RECOVERY PROBABILITY (PRP)

PRP (P_r) is defined as the probability that a raw packet is lost but recovered with CODER scheme after m number of retransmissions. We use analytical study to derive PRP as a function of the total packet error probability to determine the feasibility of lost recovery with CODER as opposed to traditional error recovery schemes such as simple (or classical) repetition-based error recovery (SIRER) scheme without

² Virtual buffer that stores all packets heard in the past T ms.

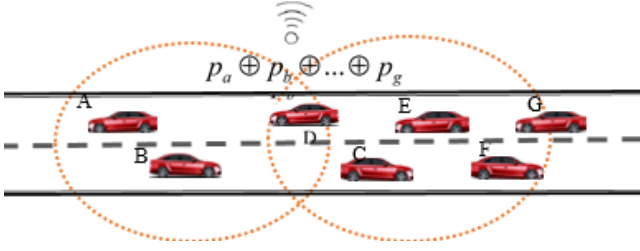


Fig. 1: CODER: Vehicle D perform an exclusive OR operation on packets (P_a, P_b, P_c, P_e, P_f , and P_g) received from other vehicles A, B, C, E, F, G and its own self-generated packets (P_d) and retransmits the XORed version to enable other vehicles within its one-hop broadcast range to recover their lost packet(s). Note that d is the diameter of vehicle i 's radio signal coverage.

network coding. Packet error probability (ϵ_i) is the probability of vehicle i losing either a raw packet or a retransmitted encoded packet. Given that in SIRER scheme, raw packets are retransmitted until the recipients have received and recovered every lost packet(s). Therefore, each node performs a total of $m+1$ number of retransmissions before lost packets are completely recovered.

Let $P_l(S)$ be the packet loss probability (PLP) of the SIRER scheme without network coding. Since in SIRER, node i retransmits its self-generated raw packets couple of $m+1$ times, $P_l(S)$ is the probability of not recovering the lost raw packets even after $m+1$ number of retransmissions. So we have

$$P_l(S) = \epsilon_i^{m+1} \quad (1)$$

With our proposed scheme, each vehicle is expected to transmit its own generated raw packets only once. Then, using LAA, another node is selected to encode and retransmit the XORed packets instead of the source vehicles repeating their raw packets indiscriminately. Here, nodes $i+1$ and $i-1$ retransmit the encoded versions of the packets $\{i \oplus (i+1)\}$ and $\{(i-1) \oplus i\}$ for $m+1$ times respectively to enable every vehicle within their vicinity to recover any packet(s) they may have lost or received incorrectly. Therefore, let $P_l(C)$ be the PLP of CODER. Then, we have

$$P_l(C) = \epsilon_i \{(1 - \alpha_i)(1 - \beta_i)\} \quad (2)$$

Where α_i and β_i represent the probability that node i can successfully recover raw packet i after successfully decoding the encoded XORed retransmitted packets $\{(i-1) \oplus i\}$ or $\{i \oplus (i+1)\}$. $P_l(C)$ implies that: 1) node i lost the raw packet i ; 2) node $i-1$ cannot recover raw packet i from the encoded packet retransmission $\{(i-1) \oplus i\}$; and (3) node $i+1$ cannot recover raw packet i from the encoded packet retransmission $\{i \oplus (i+1)\}$ as well. Considering that node $i-1$ repeats the encoded packets $\{(i-1) \oplus i\}$ for a total of $m+1$ times before all the vehicles within the radio signal transmission coverage of vehicle $i-1$ are able to successfully recover their lost packet(s), it follows that α_i is the probability that at least a single packet was received out of the $m+1$ retransmitted encoded packet $\{(i-1) \oplus i\}$ by vehicle $(i-1)$. Hence, the encoded retransmission $\{(i-1) \oplus i\}$ by vehicle $(i-1)$ can be recovered as shown in the formula:

$$\alpha_i = (1 - \epsilon_i^{m+1})\mu_i \quad (3)$$

Where μ_i represents the probability that vehicle $(i-1)$'s retransmitted XORed packets, $\{(i-1) \oplus i\}$, can be decoded by node i when successfully received. Since vehicle $i-1, i-2, \dots, i-n$ must have at least one of $i-1, i-2, \dots, i-n$ raw packets to be able to decode any of the encoded packets such as $\{(i-1) \oplus i\}, \{(i-2) \oplus 2i\}, \dots, \{(i-n) \oplus ni\}$ repetitions, μ_i as well means that vehicle $i-n$ already has at least one of the $\{(i-1) \oplus i\}, \{(i-2) \oplus 2i\}, \dots, \{(i-n) \oplus ni\}$ raw packets and therefore can decode the retransmissions of the encoded XOR packets. So that we can now have:

$$\mu_i = (1 - \epsilon_{i-n}) + \{(\epsilon_{i-n})(\alpha_{i-n})\} \quad (4)$$

Which becomes:

$$\alpha_i = (1 - \epsilon_{i-1}^{m+1})\{(1 - \epsilon_{i-n}) + \{(\epsilon_{i-n})(\alpha_{i-n})\}\} \quad (5)$$

Using our proposed scheme, the zone of interest for retransmission of the encoded XOR packets is the immediate radio signal transmission range of vehicle i as is exemplified in Fig. 1, where d is the diameter of the vehicle i 's radio signal transmission range and the vehicles outside the radio coverage of the source node. From Eq. (5), it is observed that α_i strongly depends on α_{i-1} which in turn depends on α_{i-2} , which again depends on α_{i-3} and so on till α_{i-n} . However, the difference between $i-1, i-2, \dots, i-n$ and i is trivial and insignificant. Hence, substituting $i-n$ with $i-1$ in (5), gives a linear equation of α_i which can be solved as :

$$\alpha_i = \frac{(1 - \epsilon_{i-1}^{m+1})(1 - \epsilon_{i-1})}{1 - (1 - \epsilon_{i-1})} \quad (6)$$

Given that node $i+1$ retransmits the encoded packets $\{i \oplus (i+1)\}$ for a total of $m+1$ times before all the vehicles within the radio coverage of vehicle $i+1$ can successfully recover their lost packet(s), we consider β_i as the probability that at least a single packet was received out of the XORed packets $\{i \oplus (i+1)\}$ retransmitted by vehicle $i+1$. Therefore, the encoded retransmission $\{i \oplus (i+1)\}$ by vehicle $i+1$ can be decoded by using the following formula:

$$\beta_i = (1 - \epsilon_i^{m+1})X_i \quad (7)$$

Where X_i represents the probability that vehicle $(i+1)$'s retransmitted XORed packets, $\{i \oplus (i+1)\}$, can be decoded by node i when successfully received. Again, since vehicle $i+1, i+2, \dots, i+n$ must have at least one of $i+1, i+2, \dots, i+n$ raw packets to be able to decode any of the encoded XOR packets such as $\{(i-1) \oplus i\}, \{(i-2) \oplus 2i\}, \dots, \{(i-n) \oplus ni\}$, coded packets retransmissions, β_i as well means that vehicle $i+n$ already has at least one of the $\{(i-1) \oplus i\}, \{(i-2) \oplus 2i\}, \dots, \{(i-n) \oplus ni\}$ packets and therefore can decode the retransmissions of the encoded XOR packets. So we can now have:

$$X_i = (1 - \epsilon_{i+1}) + \{(\epsilon_{i+1})(\beta_{i+1})\} \quad (8)$$

However, given that the difference in distance between $i+1, i+2, \dots, i+n$ and i is trivial and insignificant as already

mentioned previously. Consequently, after some algebraic manipulations we obtain:

$$\beta_i = \frac{(1-\epsilon_{i+1}^{m+1})(1-\epsilon_{i+1})}{1-(1-\epsilon_{i+1})} \quad (9)$$

Eq. (9) is as a result of the fact that the radius of vehicle i 's radio signal transmission range is constant. Considering Eq. (2), (6) and (9), the PLP of CODER, $P_i(C)$, can be expressed as a function of PLP and $m+1$, by combining the separate packet loss probabilities of vehicle i , (ϵ_i), vehicle $i-1$, (ϵ_{i-1}) and vehicle $i+1$, (ϵ_{i+1}).

Since, the PLP of CODER is deduced in Eq. (2) as $P_i(C) = \epsilon_i \{(1-\alpha_i)(1-\beta_i)\}$; then, substituting (6) and (9) in (2) will give us:

$$P_i(CODER) = \epsilon_i \left\{ \left(1 - \frac{\{(1-\epsilon_{i-1}^{m+1})(1-\epsilon_{i-1})\}}{1-(1-\epsilon_{i-1})} \right) \times \left(1 - \frac{\{(1-\epsilon_{i+1}^{m+1})(1-\epsilon_{i+1})\}}{1-(1-\epsilon_{i+1})} \right) \right\} \quad (10)$$

The difference in distance between vehicles $i-1, i-2, \dots, i-n$ and $i+1, i+2, \dots, i+n$ and vehicle i is trivial and insignificant given that the radius of the vehicle $i-1, i-2, \dots, i-n$ and $i+1, i+2, \dots, i+n$ and vehicle i 's radio signal transmission range does not change (i.e. 1000m as stipulated in IEEE 802.11p). Therefore, both the PLP of vehicle $i-1$ (i.e. ϵ_{i-1}) and of vehicle $i+1$ (i.e. ϵ_{i+1}) can be represented as ϵ_i . Hence, Eq. (10) can be simplified and expressed as a functions of m and ϵ_i , resulting in

$$P_i(CODER) = \epsilon_i \left\{ \left(1 - \frac{\{(1-\epsilon_i^{m+1})(1-\epsilon_i)\}}{1-(1-\epsilon_i)} \right) \times \left(1 - \frac{\{(1-\epsilon_i^{m+1})(1-\epsilon_i)\}}{1-(1-\epsilon_i)} \right) \right\} \quad (11)$$

Hence, the packet recovery probability (P_r) of SIRER and CODER scheme are given respectively as:

$$P_r(SRR) = 1 - P_i(S) \\ = 1 - \epsilon_i^{m+1} \quad (12)$$

$$\text{and } P_r(CODER) = 1 - P_i(C) \\ = 1 - \left[\epsilon_i \left\{ \left(1 - \frac{\{(1-\epsilon_i^{m+1})(1-\epsilon_i)\}}{1-(1-\epsilon_i)} \right) \times \left(1 - \frac{\{(1-\epsilon_i^{m+1})(1-\epsilon_i)\}}{1-(1-\epsilon_i)} \right) \right\} \right] \quad (13)$$

Consequently, the overall measure of performance improvement in terms of Packet(s) loss recovery potential between our proposed network coding assisted scheme and the simple error recovery scheme can be determined by comparing

the results obtained from both Eq. (13) and Eq. (12) after a given number of packet retransmissions.

I. ANALYSIS OF PACKET COLLISION PROBABILITY (PCP)

The default MAC protocol for channel access in IEEE 802.11p standard is equivalent to the Enhanced Distribution Coordination Function (EDCF) mechanism originally provided by IEEE 802.11e which has four categories of Access Classes (ACs) [4]. The initial contention window is set to W , then given p as packet collision probability, an arbitrary packet will be successfully transmitted with a probability $1-p$ and such packet will have an average backoff window $\frac{(W-1)}{2}$. If the first packet transmission fails, there will be successful retransmission using a second attempt with probability $p(1-p)$. In this case, the average backoff window becomes $\frac{(2W-1)}{2}$. This can be continued until the last $(m+1)^{th}$ permitted retransmissions and the backoff window continued until it reaches the CW_{max} value. Mathematically, the overall average backoff window can be derived as follows:

$$W_{avg} = \gamma \left(\frac{W-1}{2} \right) + \gamma p \left(\frac{2W-1}{2} \right) + \dots + \gamma p^m \left(\frac{2^m W - 1}{2} \right) + \gamma p^{m+1} \left(\frac{2^m W - 1}{2} \right), \quad (14)$$

where $\gamma = \frac{(1-p)}{(1-p^m)}$ and $(1-p^m)$ is a normalization term which enables the probability of each backoff stage to maintain a valid probability distribution. Following some algebraic manipulations, we have:

$$W_{avg} = \frac{1}{(1-p^m)} \times \left(\frac{W(1-p)(1-(2p)^m)}{2(1-2p)} - \frac{1-p^m}{2} + \frac{(2^m W - 1)(1-p^m)}{2} \right) \quad (15)$$

Given eq. (15), the probability that a vehicle attempts to transmit in an arbitrary slot is equal to

$$\frac{1}{W_{avg}} \quad (16)$$

And the probability that there is no other active node transmitting concurrently during the transmission of an arbitrary node is given by

$$\left[1 - \frac{1}{W_{avg}} \right]^{m+1} \quad (17)$$

The probability that there exists another active node (or a hidden terminal) during the transmission of an arbitrary

node leads to packet collision. Hence, from eq. (16) and (17), the packet collision probability, p is given by

$$p = 1 - \left[1 - \frac{1}{W_{avg}} \right]^{m+1} \quad (18)$$

II. LOCATION-AWARE ALGORITHM (LAA)

In VANETs, nodes are aware of their location (or coordinates) by the help of the embedded global positioning system (GPS). Vehicles also discover the location of their neighbor vehicles from the periodic 1-hop messages broadcasted periodically by each node which contain vehicle direction, velocity, position and MAC information. Given that the coordinates of destination vehicle are known, then the direction and distance from the source to the destination nodes is calculated by using vector formulae where the magnitude of vector \overline{AB} is the distance between node A and B as shown in Fig. 2.

Mathematically, the transmitting vehicle calculates the distance to its destination nodes using

$$|\overline{AB}| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (19)$$

Where (x_1, y_1) and (x_2, y_2) stands for initial and final coordinates of the vehicles respectively. In the same manner, the direction of the vector \overline{AB} is given by θ which is the formation of horizontal angle between point A and B.

$$\theta = \tan^{-1} \left\{ \frac{(y_2 - y_1)}{(x_2 - x_1)} \right\} \quad (20)$$

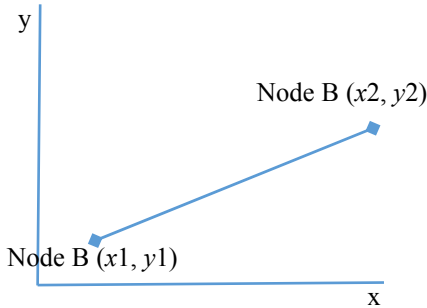


Fig. 2: Graphical representation of Vector \overline{AB}

I. SCHEME VALIDATION AND ANALYSIS

We used NS-2 [3] which is a well-used simulator in analyzing vehicular networks [3] [5] [11] to validate our analytical model. The choice of suitable vehicular movement pattern for the experiments is significant to enable us to achieve close-to-real-life scenarios with dynamic vehicular network topology. We use a 5x5 Manhattan grid road network with 2000m edge length and BonnMotion tool [16] to generate suitable node mobility model. 200 nodes were randomly distributed across the road pattern with a minimum average of 30m space between any given pair of adjacent vehicles that are

in the same lane. Configurations of the parameters for both PHY and MAC layers are according to the specifications of IEEE 802.11p standard [11]. For each vehicular network connections, UDP is used to constantly generate packets of 512 bytes every 2s. Some of the values of the parameters are shown in Table I. The Rayleigh model is used with a maximum velocity of 50mps for the simulation of channel fading effect. Speed of the vehicles ranges from 50km/h to 80km/h.

II. ANALYTICAL AND SIMULATION RESULTS

We analyse and calculate the packet recovery probability as a function of the PEP obtainable in the network. The x-axis of the graphs depicted in Figs. 3 indicates PLP. PRP is defined as the total number of lost packets recovered through m number of retransmissions to the total number of packets lost. Figs. 3 shows both the analytical and simulation results of PRP for both SIRER and CODER schemes for the value of $m = 3$. The PRP of both SIRER and CODER schemes starts to drop drastically when PLP increases towards 10^1 (see Fig. 3). This rapid degradation in loss recovery probability (LRP) caused by increased change of PLP from 10^0 to 10^1 is due to the fact that fast retransmission of raw and encoded packets tends to congest the network thereby resulting to excessive network congestion and overhead. Generally, UDP which resides at the transport layer shows increasing poor performance across the network whenever the overall network loss probability exceeds 10^0 towards 10^1 . As PLP increases, more packets are lost and even their retransmissions tend to be lost as well. In Figs. 3), there is a significant improvement (over 40% in maximum) in LRP of our proposed scheme over traditional (simple repetition) error recovery scheme. This can be explained by the fact that classical error recovery techniques based on retransmission of packets repeat each packet separately thereby congesting the channel excessively as opposed to our proposed scheme which combines two or more packets into one, without increasing the size of the packet.

Hence, with a lesser number of retransmissions, our proposed scheme is able to provide over 40% lost packet recovery compared to simple repetition approach. More interesting observation is also witnessed in Fig. 2 where the results of the analysis and simulation of both SIRER and CODER scheme agree with each other. In other words, the results of our simulation experiments show that the analytical model is accurate in calculating both the recovery of lost packet(s) and packets collision probability for both vehicle's periodic status and emergency packets.

TABLE I. VALUE OF PARAMETERS USED IN OUR SIMULATIONS

Parameter	Value
Frequency	5.9GHz
Bandwidth	10MHz
Modulation and Data rate	BPSK, 3Mbps
Transmission power	2mW
DIFS time	64μs
Packet rate	10pkts/s
Packet size	512Bytes

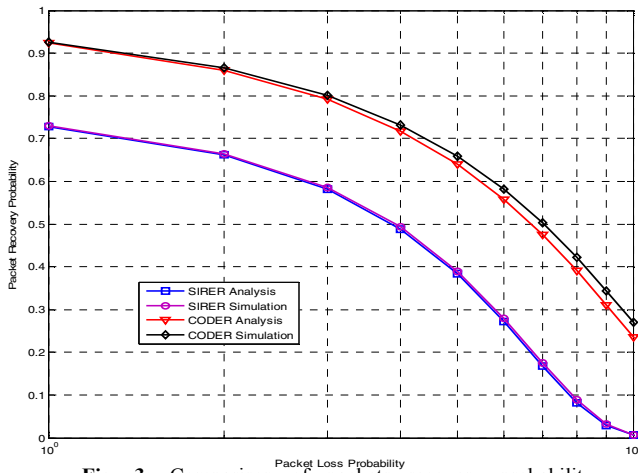
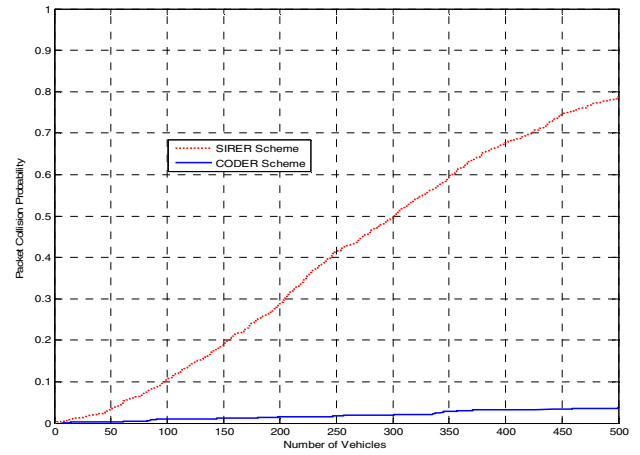


Fig. 3: Comparison of packets recovery probability predicted by simulation and analytical results of CODER and SIRER scheme when $m = 3$



predicted by simulation and analytical results of CODER and SIRER scheme using a dense vehicular network

Unfortunately, in practical vehicular networks, packet loss probability essentially depends on several other factors including the distance existing between the sender and receiver, node density, multi-path fading and network congestion. Hence, we further study the performance gain of CODER over SIRER scheme using packet collision probabilities (PCP) under heavy network density (see Fig. 4). The wide difference that exist between the two results are expected given that an increased vehicle traffic density will lead to increased channel load especially for SIRER scheme when the total number of packet retransmission increases. Consequently, it will result to heavy network congestion and overall increase in packet collision. Hence, Fig. 4 shows that our proposed scheme performs better than simple repetition approach.

III. CONCLUSIONS

In this paper, we presented an analytical model to achieve a reliable and efficient message dissemination in vehicular networks using network coding technology. Our model computes the successful packet recovery probability as a function of packet loss probability and packet collision probability of vehicular communication systems. The results of our simulation experiments show that the analytical model is accurate in calculating both the recovery of lost packet(s) and packets collision probability for both vehicle's periodic status and emergency packets. Enhancing the reliability of relay-based retransmission technique using network coding will form an interesting future work to complement our proposed scheme in order to cover a wider range of V2V communication system.

REFERENCES

- [1] Elias Eze C.; Sijing Zhang; Enjie Liu, "Vehicular ad hoc networks (VANETs): Current state, challenges, potentials and way forward," *2014 20th International Conference on Automation and Computing (ICAC)*, pp.176-181, 12-13 Sept. 2014.
- [2] Bai, F., Krishnan, H., "Reliability Analysis of DSRC Wireless Communication for Vehicle Safety Applications," *IEEE Intelligent Transportation Systems Conference*, Toronto, pp.355-362, Sept. 2006.
- [3] J.W. Lee and Y.H. Lee, "ITB: Intrusion-Tolerant Broadcast Protocol in Wireless Sensor Networks", *Second International Conference on High Performance Computing and Communications, HPCC 2006*, Munich, Germany, pp. 505-514, 2006.
- [4] Yuanguo Bi; Cai, L.X.; Xuemin Shen; Hai Zhao, "Efficient and Reliable Broadcast in Intervehicle Communication Networks: A Cross-Layer Approach," *IEEE Transactions on Vehicular Technology*, vol.59, no.5, pp.2404-2417, Jun 2010.
- [5] F. Farnoud and S. Valaee, "Repetition-based broadcast in vehicular ad hoc networks in rician channel with capture," in *IEEE INFOCOM Workshops*, pp. 1-6, 2008.
- [6] L. Yang, G. Jinhua, and W. Ying, "Piggyback cooperative repetition for reliable broadcasting of safety messages in vanets," in *6th IEEE Consumer Communications and Networking Conference, CCNC*, 2009.
- [7] R. Ahlswede, N. Cai, S.Y.R. Li, and R.W. Yeung, "Network Information Flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204-1216, July 2000.
- [8] Li, S.-Y.R.; Ning Cai; Yeung, R.W., "On theory of linear network coding," 2005. *ISIT 2005 Proceedings. International Symposium on Information Theory*, Adelaide, SA, pp.273-277, 4-9 Sept. 2005
- [9] Wang, Yasong; Zhang, Qinyu, "An Approach on Wireless Broadcasting Retransmission Using Network Coding," *Wireless Communications, Networking and Mobile Computing (WiCOM)*, 2012 8th International Conference on , vol., no., pp.1-4, 21-23 Sept. 2012
- [10] Dong Nguyen; Tran, T.; Thanh Nguyen; Bose, B., "Wireless Broadcast Using Network Coding," *IEEE Transactions on Vehicular Technology*, vol.58, no.2, pp.914-925, Feb. 2009.
- [11] Q. Xu, T. Mak, J. Ko, and R. Sengupta, "Vehicle-to-vehicle safety messaging in DSRC," in *Proceedings of the 1st ACM international workshop on Vehicular ad hoc networks, VANET04*, pp. 19-28, .2004.
- [12] F. Farnoud and S. Valaee, "Repetition-based broadcast in vehicular ad hoc networks in rician channel with capture," in *IEEE INFOCOM Workshops*, pp. 1-6, 2008.
- [13] Zhe Wang; Hassan, M., "Network Coded Repetition: A Method to Recover Lost Packets in Vehicular Communications," *2011 IEEE International Conference on Communications, (ICC)*, pp.1-6, June 2011.
- [14] Y. Wu, L. Yang, G. Wu, and J. Guo, "An improved coded repetition scheme for safety messaging in vanets," in *5th International Conference on Wireless Communications, Networking and Mobile Computing, WiCom '09*, pp. 1-4, 2009.
- [15] Dong Nguyen; Tran, T.; Thanh Nguyen; Bose, B., "Wireless Broadcast Using Network Coding," *IEEE Transactions on Vehicular Technology*, vol.58, no.2, pp.914-925, Feb. 2009.
- [16] Nils A.; Raphael E.; Elmar G.; Matthias S., "BonnMotion: a mobility scenario generation and analysis tool," In *Procs of the 3rd International ICST Conference on Simulation Tools and Techniques*, pp. 51- 60; 2010

Joint Resource Blocks Switching Off and Bandwidth Expansion for Energy Saving in LTE Networks

Kapil Kanwal, Ghazanfar A. Safdar, Shyqyri Haxha

Institute of Research in Applied Computing

University of Bedfordshire, LU1 3JU, Luton, United Kingdom

kapil.kanwal@study.beds.ac.uk, {ghazanfar.safdar, shyqyri.haxha}@beds.ac.uk

Abstract— In the wireless networks community, Long Term Evolution (LTE) facilitates users with high data rate at the cost of increased energy consumption. The base station (BS) also known as eNodeBs are the main energy hungry elements in LTE networks. Power is consumed by different components of BS such as Baseband Unit (BB), Power Amplifier (PA) and other cooling systems. Since power consumption directly affects the Operational Expenditure (OPEX), thus the provision of cost effective services with adequate quality of service (QoS) has become a major challenge. Moreover, the energy consumed by Information and Communication Technology (ICT) appliances contributes 2% to global warming (CO₂ emission), which is another significant problem. This paper presents a joint resource blocks switching off and bandwidth expansion energy saving scheme for LTE networks. Performance analysis of the proposed scheme has revealed that it is around 29% energy efficient as compared to the benchmark LTE systems.

Keywords— Long Term Evolution; OPEX; energy saving; resource blocks switching; bandwidth expansion

I. INTRODUCTION

Energy-saving related research has gained vast attraction in LTE networks. Energy-efficient carrier aggregation algorithms group the component carriers (CC) to provide enhanced energy saving (ES) [1]. In this context, authors proposed an energy-efficient resource allocation technique in Multicast Broadcast Single Frequency Networks (MBSFN) [2]. Energy efficiency could also be achieved in dynamic distance-aware approach which involves switching off base stations depending on distance and time varying load information [3, 4]. On the same lines, optimized resource allocation could also reduce energy consumption [5, 6]. Centralized and distributed schemes which engage users (UEs) migration also improve ES [7, 8]. Another scheme which benefits from bandwidth expansion has also been investigated [9]. Dynamic traffic-aware approach [10], uses time varying traffic information for energy conservation. Each BS divides its cell in different numbers of sectors, then switches off the appropriate sector (with low traffic) providing power saving opportunities. Energy efficient BSs deployment has also improved energy conservation [11]. Researchers have proposed an energy-efficient link adaptation scheme which combines the traditional link adaptation with power control, resulting in enhanced energy efficiency at the BS [12]. The mentioned scheme uses BS's

transmitted power as a new feedback parameter and predicts an optimal set of parameters in order to maximize the BS's energy efficiency and satisfy the Block Error Rate (BLER) constraint for the channel state. Another interesting scheme is presented in [13]; authors proposed an energy-efficient resource allocation that operates in multi-cells Orthogonal Frequency Division Multiple Access (OFDMA) based LTE networks. It combines dynamic resource block allocation with energy-efficient power provision and reduces the overall BS's power consumption. In low load networks, ES is also tackled by addressing factors such as energy efficiency and mobility load balancing. An effective energy efficient resource allocation optimization model has been designed which used a low complexity method called Energy Efficient Virtual Bandwidth Expansion Mode (EE-VBEM) to achieve the objective of optimization [14].

Most of the above discussed schemes offer ES but in light loaded networks only and do not work efficiently during high traffic load. Therefore despite existing research, there is strong need to develop an ES scheme in LTE networks which span across different layers to provide ES during both lightly and heavily loaded network. In this context, a novel ES scheme is proposed in this paper which reduces dynamic power consumption at Downlink BS by employing resource blocks (RBs) switching off and bandwidth expansion through time compression resulting in reduced physical downlink control channel (PDCCH) signaling. The novel feature of our proposed scheme lies in the fact that it performs early handover without waiting for A3 event. Moreover proposed scheme reduces PDCCH's overhead through extended bandwidth, which further reduces power consumption and provides enhanced ES opportunities. The rest of the paper is organized as follows: Section II describes our proposed scheme. It also presents the salient features of our scheme. Section III provides performance analysis of proposed scheme. Paper is finally concluded in section IV with discussion on future work.

II. PROPOSED SCHEME

The proposed scheme while implemented at every BS (a.k.a. eNodeB) enable BSs to achieve load balancing among themselves by relocating users (UEs) from overlapping areas

to the centre cell as shown in Fig (1) and giving the neighbouring cells' BSs an opportunity to switch off RBs and attain increased ES in densely deployed LTE networks. Employing the concept of incremental RBs, compared to benchmark, our proposed scheme combines two RBs per UE thereby resulting into reduced PDCCH overhead transmission. By doing so, it improves energy conservation and better system capacity. Enhancement in system capacity helps even further towards load balancing, thus addressing the problem of BS capacity limitation during high traffic periods. Accordingly, a BS in our proposed scheme could accommodate more UEs from the overlapping area of neighbouring cells; thereby further increasing ES opportunities.

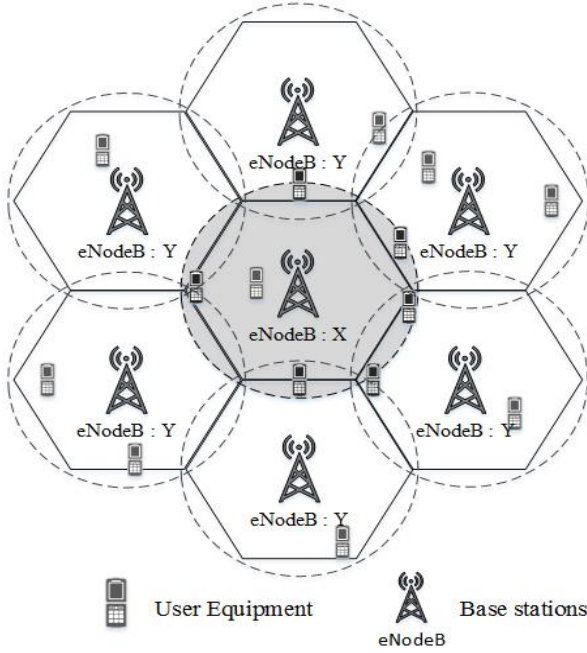


Fig. 1: Proposed Scheme System Model

Fig. 2 show two cells (Centre Cell X and Neighbour Cell Y) to elaborate working of proposed scheme. In densely deployed LTE networks, there could be numerous centres and neighbour cells in cluster of cells, while each centre cell has six neighbour cells as shown in Fig 1. Accordingly in proposed scheme the counter for centre cells X can vary from $X \rightarrow T$, whereas neighbour cells counter varies from $Y \rightarrow Z$. Each BS exchanges load information with neighbour cells through X2 interface, whereas the identification of UEs in the overlapping area is based on distance (D), which is calculated using Signal to Interference and Noise Ratio (SINR)/ positioning reference signal (RS). Based on traffic load, neighbour BSs are arranged in ascending order before any relocation of the UEs can take place, provided the centre cell capacity is below threshold. Proposed scheme iteratively works through neighbour cells (Y) list before total number of centre and neighbour cells are dealt with, T and Z accordingly. Table I clearly provides the description for notations used in Figure 2.

Table I. Notations Description

Y	Neighbour cells among selected centre cell X
Z	Total number of neighbour cells Y in list
X	Centre cells
T	Total number of centre cells
Capacity	Available RBs in current cell X
Threshold	Indicator of resources availability

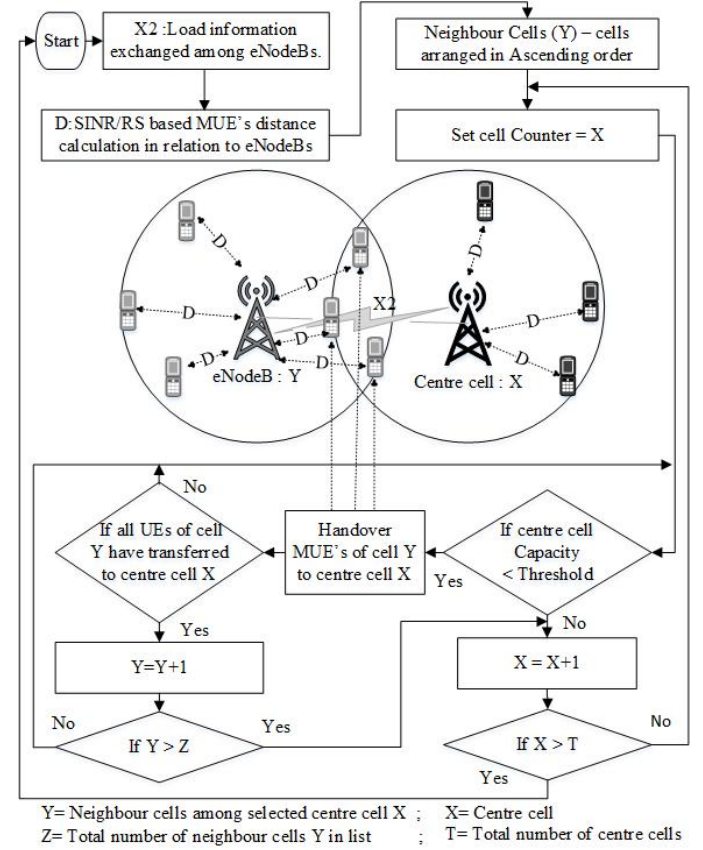


Fig. 2: Proposed Scheme

A. Resource Blocks Switching off:

Considering 3GPP specification for handover process in LTE networks [15]; when UEs enter in overlapping area, they receive cell specific reference signals (RS) from neighbour cell which is used to calculate Reference Signal Receive Power (RSRP) of neighbour cell X. In parallel UEs also calculate RSRP value of serving cell Y. Handover is triggered through A3 event which indicate that RSRP value of neighbour cell has become greater then serving cell. Handover process is optimized through parameter called hysteresis. The role of hysteresis in A3 event is to worsen neighbour's RSRP value than the actual value (Equation 1) to result in reduced call drop probability and ensure seamless connectivity.

$$RSRP_T \geq RSRP_S + \text{Hysteresis} \quad (1)$$

$RSRP_T$ presents RSRP value of target (centre) cell X, while $RSRP_S$ presents RSRP value of current (serving) cell Y while Hysteresis controls early handover process. In our proposed scheme Hysteresis value is reduced to perform early handover while satisfying Radio Link Failure (RLF) requirements at cell edges as shown in Fig. 3. The purpose of early handover is to turn OFF unused RBs of serving cell Y earlier. Worth noting, the novelty of our proposed scheme lies in a fact that, UEs do not have to wait for an A3 event to occur before handing over from serving cell Y to the centre cell X, which is unique from LTE standard [17]. This offers greater opportunities to turn OFF the spared RBs of serving cell Y, resulting in improved ES. Resource blocks switching OFF implementation pseudo code is provided in Fig. 4.

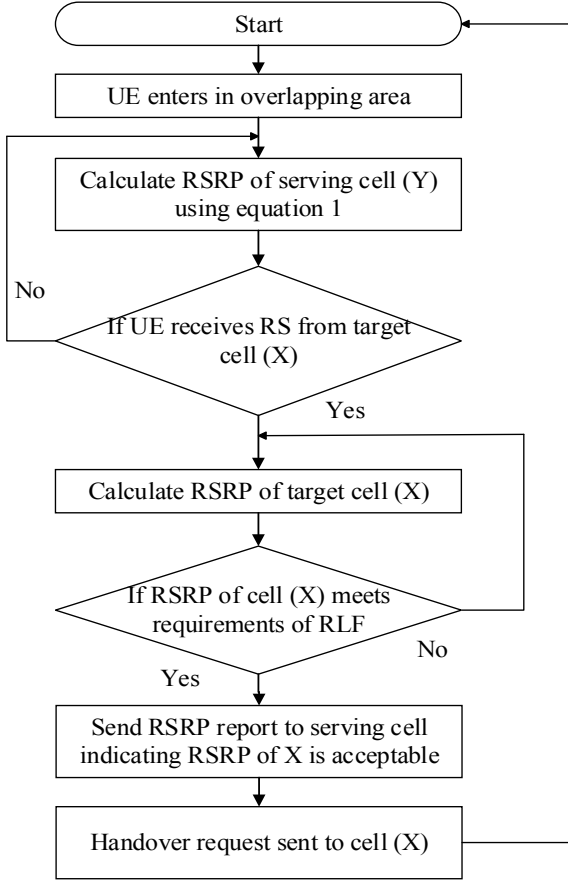


Fig.3: Proposed ES handover flow diagram

B. Bandwidth Expansion:

Each RB carries two parts, user's data part and PDCCH signals with RS as shown in Fig 5. PDCCH signals can occupy the 1st, 2nd or 3rd OFDMA symbols in each time slot of given RB. Our scheme combines two RBs to expand bandwidth through time compression and allocates these joint RBs to single user thus reducing PDCCH's overhead as shown in (Fig 5). Considering 3GPP link budget at downlink, the maximum overhead due to PDCCH is around 26% which is transmitted using the fraction of time slot at constant pre-set of radio frequency power. Hence the dynamic energy cost at BS can be calculated as [9].

$$EC_BEN_{Pdc} = M \cdot TM [(1 - \mu_{Pdc})EN_{Data} + \mu_{Pdc} \cdot EN_{Pdc}] \quad (2)$$

EC_BEN_{Pdc} is the energy required for transmission of RBs in benchmark system. M is total number of allocated RBs, while TM is transmission duration. μ_{Pdc} represent PDCCH's overhead. EN_{Data} is amount of power required for data transmission, and EN_{Pdc} accounts for PDCCH overhead's transmission power. On the other hand, our scheme expands the bandwidth using time compression factor ($B = 2$), hence pooling together two RBs which increases the size of sub-frame. Therefore EN_{Data} in proposed scheme consumes double energy as compared to the one RB in benchmark, however proposed scheme saves considerable energy due to removal of control overhead (equation 4 below). Equation 2 thus can be rewritten as:

$$EC_Proposed_{Pdc} = M \cdot TM [(B \cdot EN_{Data})(1 - \mu_{Pdc}) + \mu_{Pdc} \cdot EN_{Pdc}] \quad (3)$$

$EC_Proposed_{Pdc}$ is energy consumed after time compression. $B \cdot EN_{Data}$ is the power required for data transmission after bandwidth expansion.

Fig .4: Resource Blocks Switching off

Referring to Figure 2:

1. Get T ; Get Z ; U_Y represent UEs of Y
2. Set cell counter = X // X : Set of centre cells;
3. Set cell counter = Y // Y : Set of neighbour cells;
4. **For** each $X \rightarrow T$ **do**, $\forall X \rightarrow T$
5. **For** each $Y \rightarrow Z$ **do**, $\forall Y \rightarrow Z$
6. **IF** ($RSRP_{U_Y} < THRESHOLD_Y \in Y$) // threshold of Y
7. **then** Measure $RSRP_X$ for $X \in U_Y$ // Calculate X 's RSRP
8. Information of U_Y exchanged between X & Y
9. **IF** $X_{CAPACITY} < THRESHOLD_{CAPACITY}$
10. X acknowledge the availability of resources to U_Y & Y
11. **then** $\forall U_Y \leftarrow$ belongs to overlapping area of Y
12. Y send HO request $\rightarrow U_Y$ // HO: Handover
13. U_Y send HO confirmation $\rightarrow Y$
14. $HO \forall U_Y \in Y \rightarrow X$
15. Y send data packet information $\rightarrow X$
16. Y turn off unused RBs of $\forall U_Y$ handover to X
17. $Y = Y + 1$; move to next neighbour cell
18. **While** ($Y \gtrsim Z$) when all neighbour cells have served
19. **do** $X = X + 1$; move to next centre cell in cluster
20. **Endwhile**
21. **Else IF** ($X_{CAPACITY} \geq THRESHOLD$)
22. $X = X + 1$; move to next centre cell X
23. **End IF**
24. **End IF**
25. **End For**
26. **While** ($X \gtrsim T$) when all centre cells have been served
27. **do** Stop/Terminate:
28. **Endwhile**
29. **End For**

Since two RBs jointly transmitted to single user; thereby it reduces the PDCCH's signaling ($\mu_{Pdc} \cdot EN_{Pdc}$), hence following holds true for enhanced ES.

$$\mathcal{B} = 2: EC_Proposed_{Pdc} \text{ must be } < EC_BEN_{Pdc} \cdot \mathcal{B} \quad (4)$$

Equation 4 compares energy consumption of total RBs of benchmark against the total RBs of proposed scheme. Our scheme besides providing improved energy efficiency does not deteriorate in data transmission (equation 5). Where \mathcal{C} is data rate and \mathcal{B} is compression factor in relation to benchmark system.

$$\mathcal{C}Proposed = \mathcal{C}Benchmark \mathcal{B} \quad (5)$$

From SINR calculation formula, the required Radio Frequency (RF) transmission power (EN_{Data}) for RBi can be calculated as:

$$\mathcal{R}'_i = \frac{o' + J_i}{S_{pq}}(Z_i) \quad (6)$$

While Z_i is required target SINR for RBi, O' is the noise floor, S_{pq} is path gain between BS (q) and UE (p), and J_i is interference at RBi. Using equations (2) & (6) the total transmission energy required to deliver the payload on \mathcal{B} for benchmark system can be calculated as:

$$TP_{Benchmark} = TM \sum_{i=1}^B \left[\{(1 - \mu_{Pdc}) \mathcal{R}'_i\} + \mu_{Pdc} \cdot EN_{Pdc} \right] \quad (7)$$

$$TP_{Benchmark} = TM \sum_{i=1}^B \left[\left\{ (1 - \mu_{Pdc}) \times \frac{o' + J_i}{S_{pq}}(Z_i) \right\} + \mu_{Pdc} \cdot EN_{Pdc} \right] \quad (8)$$

While total transmission energy required to deliver the total number of RBs (M) in benchmark system can be calculated as:

$$TP_{Benchmark} = TM \sum_{i=1}^M \left[\left\{ (1 - \mu_{Pdc}) \times \frac{o' + J_i}{S_{pq}}(Z_i) \right\} + \mu_{Pdc} \cdot EN_{Pdc} \right] \quad (9)$$

The total transmission energy required to deliver the total number of RBs (M) in proposed scheme can be calculated using equations (3) and (6):

$$TP_{Proposed} = TM \sum_{i=1}^M \left[\{(\mathcal{B} \cdot \mathcal{R}'_i) \times (1 - \mu_{Pdc})\} + \mu_{Pdc} \cdot EN_{Pdc} \right] \quad (10)$$

$$TP_{Proposed} = TM \sum_{i=1}^M \left[\left\{ \mathcal{B} \left(\frac{o' + J_i}{S_{pq}}(Z_i) \right) \times (1 - \mu_{Pdc}) \right\} + \mu_{Pdc} \cdot EN_{Pdc} \right] \quad (11)$$

Hence using equation (9) and equation (11), Energy Consumption Gain (ECG) for proposed scheme in comparison with benchmark can be calculated as

$$ECG = \frac{TP_{Benchmark}}{TP_{Proposed}} \quad (12)$$

C. Power model:

Proposed scheme adopts power model presented in [16] for energy related calculations at BS.

$$P_{Total} = [P_{Dynamic}] + [P_{Static}] \quad (13)$$

P_{Total} is total power consumption which consists of dynamic power ($P_{Dynamic}$) and static power (P_{Static}) components. P_{Static} is constant part, while $P_{Dynamic}$ corresponds to radio operations depending on traffic load, number of UEs and their distance from associated BS. The $P_{Dynamic}$ is further categorized in two parts; P_{Amp} which represents power consumption by BS during RBs transmission, and P_{Idle} which means that the BS is in idle state as shown in equation (14).

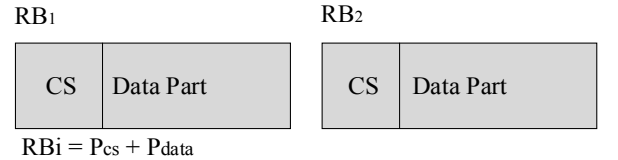
$$\text{Further, } P_{Dynamic} = P_{Idle} + P_{Amp} \quad (14)$$

$$P_{Amp} = \frac{P_{Trx}}{\rho} \quad (15)$$

P_{Amp} is power consumed by power amplifier (PA). P_{Trx} is output transmission power and ρ is efficiency component of PA which depends on electrical input power and RF output power.

$$P_{Static} = [P_{Trans}] + [P_{Dsp}] + [P_{Rect}] + [P_{Ac}] \quad (16)$$

P_{Trans} represents the required transceiver power, P_{Dsp} is the digital signal processing, P_{Rect} is power required for rectification and P_{Ac} accounts for cooling purpose.



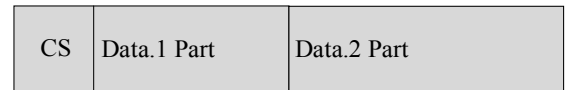
Pcs: Power consumed by control signals.

Pdata: Power consumed by data part.

RBi presents resource block.

After Bandwidth expansion by B=2

RB1,2



$$RB1,2 = Pcs + (Pdata1 + Pdata2)$$

$$RB1,2 = Pcs + 2(Pdata)$$

Fig.5: Bandwidth expansion through time compression

III. PERFORMANCE ANALYSIS

Performance analysis of proposed scheme is done using system level simulators in MATLAB with parameters according to the 3GPP specifications [17] as shown in Table II. Proposed scheme is compared with LTE network (benchmark) with no ES based on 3GPP specifications. Benchmark model consists of core network also known as Evolved Packet Core (EPC), the UEs and the Evolved Universal Terrestrial Radio Access Network (E-UTRAN) based on 3GPP LTE specifications [17]. The considered LTE network topology is a densely deployed scenario which consists of 7 hexagonal cells with overlapping neighbour cells as shown in Fig 1. Each cell has radius of around 700 meters while it contains 10 mobile UEs randomly distributed within the coverage area. Power consumption is calculated at downlink transmission on eNodeB.

Table II: System Parameters

Parameters	Value	Parameters	Value
Bandwidth	20 MHz	RB per UE	1,2
Carrier frequency	2.14 GHz	Time compression	$\mathcal{B} = 2$
Traffic model	Full Buffer	BS Max. power	46 dBm
Target SINR, Z_i	6.3, 12.1 dB	Overhead, μ_{pdc}	14.99%
UE Speed	2.8 m/s	Total number of cells	7

A. Users Acceptance and Expansion Factor \mathcal{B} .

Proposed Scheme employs RBs expansion factor value of $\mathcal{B} = 2$ & 3 and is compared with benchmark system ($\mathcal{B} = 1$) in Fig. 6. It is evident that there is direct relationship between power consumption and data rate. Although RB expansion factor of $\mathcal{B} = 3$ is energy efficient compared to $\mathcal{B} = 2$, it significantly reduces overall available capacity per eNodeB, increasing the unaccepted user rate (Fig. 7).

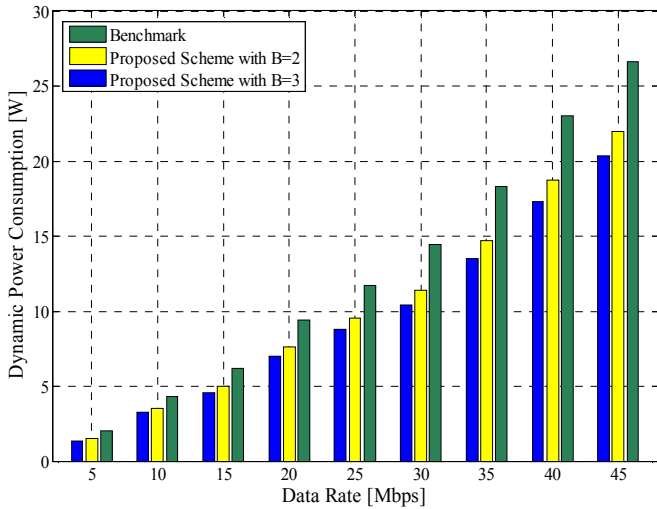


Fig. 6: Data Rate vs Power Consumption

Figure 7 shows that with $\mathcal{B} = 3$, proposed scheme rejects around 40% users due to lack of resources which is clearly reduced user satisfaction and undermines QoS, while user's rejections rate with $\mathcal{B} = 2$ is around 10%. Therefore $\mathcal{B} = 2$ is considered best expansion factor for proposed scheme to keep good balance of ES, overall eNodeB capacity and users acceptance.

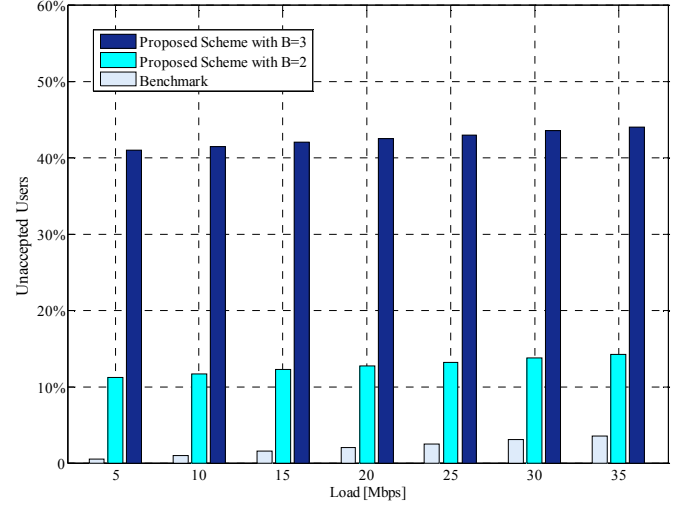


Fig. 7: Unaccepted Users

B. Power Consumption and Energy Consumption Gain.

Fig. 8 demonstrates CDF plot for the dynamic power consumption in our system and benchmark. Our scheme is around 29 % efficient while compared to the benchmark. It offers this energy efficiency by employment of joint RBs switching off and bandwidth expansion.

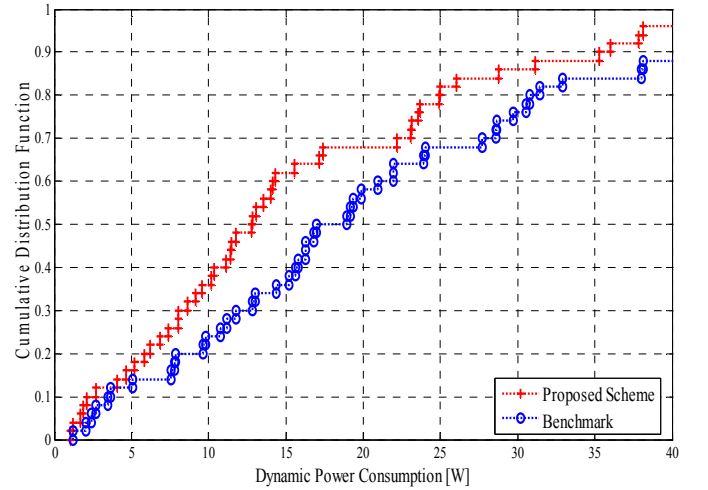


Fig. 8 : Dynamic Power Consumption

Energy Consumption Gain (ECG) is in direct relationship with the energy saving, therefore 29% energy saving can provide 29% improved ECG over Benchmark as shown in Fig. 9.

C. Throughput.

The fact that expansion factor $\mathcal{B} = 2$ accounts for availability of increased bandwidth results in improved throughput for

proposed scheme (6.1 Mbps) as compared to benchmark systems (5.5 Mbps), Fig. 10.

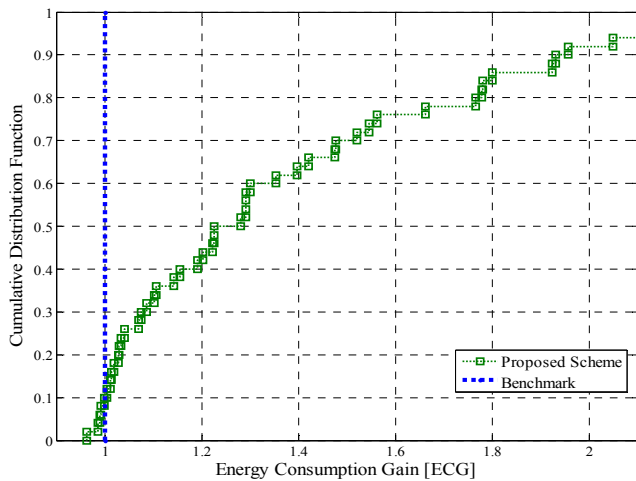


Fig. 9: Energy Consumption Gain [ECG].

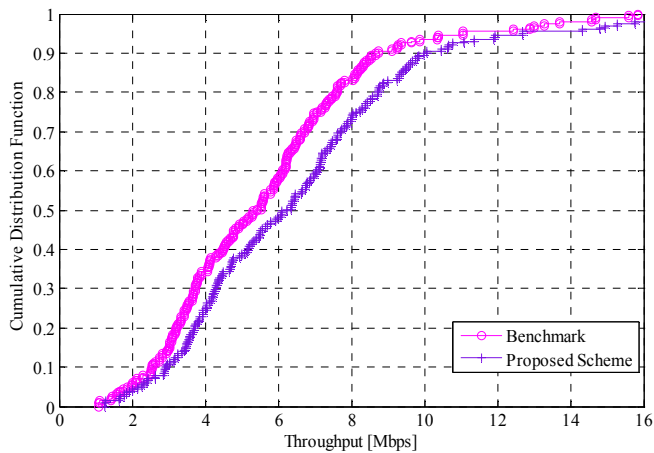


Fig 10: Throughput

IV. CONCLUSION

Due to the increased OPEX and global warming, energy saving has become a major challenge for future telecommunication systems. The eNodeBs are the main energy hungry elements in LTE networks. This paper presents a joint resource blocks switching off and bandwidth expansion ES scheme for LTE networks. The performance of proposed scheme was analyzed using system level simulations for a densely deployed LTE scenario. Proposed scheme was found to provide improved throughput and reduced energy consumption compared to benchmark systems. On the other hand due to the early handovers without considering A3 event, UEs served by centre cell suffer from weak signals quality at the cell edges which is the limitation of our proposed scheme.

Future work involves large clusters which contain two or more centre cells. Additionally the proposed scheme will be compared with existing state of art.

REFERENCES

- [1] Fei Liu; Kan Zheng; Wei Xiang; Hui Zhao, "Design and Performance Analysis of An Energy-Efficient Uplink Carrier Aggregation Scheme," *Selected Areas in Communications, IEEE Journal on*, vol.32, no.2, pp.197,207, February 2014
- [2] Migliorini, D.; Stea, G.; Caretti, M.; Sabella, D., "Power-Aware Allocation of MBSFN Subframes Using Discontinuous Cell Transmission in LTE Systems," *Vehicular Technology Conference (VTC Fall)*, 2013 IEEE 78th, vol., no., pp.1,5, 2-5 Sept. 2013
- [3] Bousia, A.; Kartsakli, E.; Alonso, L.; Verikoukis, C., "Dynamic energy efficient distance-aware Base Station switch on/off scheme for LTE-advanced," *Global Communications Conference (GLOBECOM)*, 2012 IEEE, vol., no., pp.1532,1537, 3-7 Dec. 2012
- [4] Bousia, A.; Antonopoulos, A.; Alonso, L.; Verikoukis, C., "'Green' distance-aware base station sleeping algorithm in LTE-Advanced," *Communications (ICC)*, 2012 IEEE International Conference on, vol., no., pp.1347,1351, 10-15 June 2012
- [5] Han, C.; Armour, S., "Energy efficient radio resource management strategies for green radio," *Communications, IET*, vol.5, no.18, pp.2629,2639, Dec. 16 2011
- [6] Yun Li; Wenjing Liu; Bin Cao; Man Li, "Green resource allocation in LTE system for unbalanced low load networks," *Personal Indoor and Mobile Radio Communications (PIMRC)*, 2012 IEEE 23rd International Symposium on, vol., no., pp.1009,1014, 9-12 Sept. 2012
- [7] Samdanis, K.; Kutscher, D.; Brunner, M., "Self-organized energy efficient cellular networks," *Personal Indoor and Mobile Radio Communications (PIMRC)*, 2010 IEEE 21st International Symposium on, vol., no., pp.1665,1670, 26-30 Sept. 2010
- [8] Taleb, T.; Brunner, M.; Kutscher, D.; Samdanis, 2011 "Self organized network management functions for energy efficient cellular urban infrastructures," 2011
- [9] Videv, S.; Haas, H.; Thompson, J.S.; Grant, Peter M., "Energy efficient resource allocation in wireless systems with control channel overhead," *Wireless Communications and Networking Conference Workshops (WCNCW)*, 2012 IEEE, vol., no., pp.64,68, 1-1 April 2012
- [10] Hossain, M.F.; Munasinghe, K.S.; Jamalipour, A., "Toward self-organizing sectorization of LTE eNBs for energy efficient network operation under QoS constraints," *Wireless Communications and Networking Conference (WCNC)*, 2013 IEEE, vol., no., pp.1279,1284, 7-10 April 2013
- [11] Coskun, C.C.; Ayanoglu, E., "Energy-Efficient Base Station Deployment in Heterogeneous Networks," *Wireless Communications Letters, IEEE*, vol.3, no.6, pp.593,596, Dec. 2014
- [12] Guo Li; Shi Jin; Fuchun Zheng; Xiqi Gao; Xiaoyu Wang, "Energy Efficient Link Adaptation for Downlink Transmission of LTE/LTE-A Systems," *Vehicular Technology Conference (VTC Fall)*, 2013 IEEE 78th, vol., no., pp.1,5, 2-5 Sept. 2013
- [13] Dabing Ling; Zhao Ming Lu; Wei Zheng; Xiangming Wen; Ying Ju, "Energy efficient cross-layer resource allocation scheme based on potential games in LTE-A," *Wireless Personal Multimedia Communications (WPMC)*, 2012 15th International Symposium on, vol., no., pp.623,627, 24-27 Sept. 2012
- [14] Yun Li; Wenjing Liu; Bin Cao; Man Li, "Green resource allocation in LTE system for unbalanced low load networks," *Personal Indoor and Mobile Radio Communications (PIMRC)*, 2012 IEEE 23rd International Symposium on, vol., no., pp.1009,1014, 9-12 Sept. 2012
- [15] 3GPP, "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Handover procedures" TS 23.009, v.12.0.0.
- [16] Margot Deruyck, Emmeric Tanghe, Wout Joseph and Luc Martens, "Modelling the Energy Efficiency of Microcell Base Stations" 1st international conference on smart grids, green communications and IT energy-aware technologies, p.1-6, 2011
- [17] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Networks (E-UTRAN): Overall description", TS 36.300, V10.4.0.

Energy Harvesting Wireless Sensor Node for Monitoring of Surface Water

Zivorad Mihajlovic, Ana Joza, Vladimir Milosavljevic, Vladimir Rajs, Milos Zivanov
Department of Power, Electronic and Telecommunication Engineering
Faculty of Technical Sciences, University of Novi Sad
Novi Sad, Serbia
zivorad@uns.ac.rs

Abstract—Monitoring of surface water areas before, during and after the flood is important to collect useful information for future risk assessment and for the purpose of learning about future preventive measures. This paper describes the development of the wireless sensor network (WSN) node capable of collecting water level measurement data on remote locations that are covered by surface waters, either by flooding or seasonal environmental impacts. Our main goal was to create a flexible, easy-to-deploy and easy-to-maintain, adaptable, low-cost and low power WSN node for monitoring of water level. Measuring node is based on the ultra low power components, with the most important MSP430 series microcontroller and radio frequency (RF) module based on nRF24L01+ integrated circuit. Each node contains energy harvesting subsystem and supercapacitor for energy storage. Water level measurement is performed by a separate module, which is based on PIC12F1822 microcontroller that supports small capacitance measurements. Water level measurement results for different situations were collected, analyzed and graphically presented.

Keywords—wireless sensor networks; environmental monitoring; water level sensor; energy harvesting; capacitive sensing module

I. INTRODUCTION

Although the progress of technology has contributed to development of systems for prevention and protection, natural disasters continue to have a very negative impact on humanity. Floods are an example that shows that despite of the developed infrastructure and protective systems, they cause huge losses in both human lives and material damage each year. Traditionally, design standards and structural flood protective measures were the dominant flood management approaches. Structural flood protection measures, such as dikes and retention basins, were designed in order to control up to a certain, predefined design flood, e.g. a 100-year flood [1]. New concepts have been developed, usually referred to as “flood risk management”. The level of protection is determined by broader considerations than some predefined design flood while more emphasis is put on non-structural flood mitigation measures.

Besides floods, the problem of ground water in the plain areas drastically affects agricultural production. By monitoring levels, watercourses and canal systems it is possible to collect information which analysts can use to make optimal decisions. Thus, the monitoring of water level changes, which are caused by flooding or

groundwater, could contribute to the overall risk assessment.

The concept of WSN can be of great benefit in this case as evidenced by numerous examples [2]-[5]. A large number of sensors in flood-affected environment can create a real picture of which area is under water, in what period of the year and how the water level changes over time.

WSN node adapted for use in the field conditions with ultra low power consumption and energy harvesting support has been developed for the purposes of environmental monitoring. A more detailed description is given in Section 2. This basic node can be easily upgraded in accordance with the desired application. In this case the module has been upgraded with subsystem for measurement of very small capacitance which is used to measure the water level. More details about the procedure are given in Section 3. Section 4 shows the practical results of water level measurements, together with mathematical and graphical results. We conclude in Section 5.

II. DESIGN OF WSN NODE WITH ENERGY HARVESTING SUPPORT

The use of WSN in the field of environmental protection is increasingly considered, in practice only rich countries spend significant resources in this sphere. That is the reason why the cost of these systems must be as low as possible. For this to be fulfilled during design of WSN node, the following requirements are defined:

- Ultra low power consumption.
- Low cost design with commercial and easily available components with reasonable prices.
- Miniature design and low cost maintenance.
- Easy functionality upgrade.
- Energy harvesting support.

A. Hardware Architecture of WSN Node

Based on the defined requirements, WSN module is designed, and its hardware structure is shown in the block diagram of Fig. 1. The ideas and reasons that led to this implementation will be discussed in more detail. Control and communication systems are based on commercial components MSP430G2553 microcontroller and nRF24L01+ RF transceiver, because they already proved to be an excellent choice for applications where ultra low power consumption and low cost are required [6].

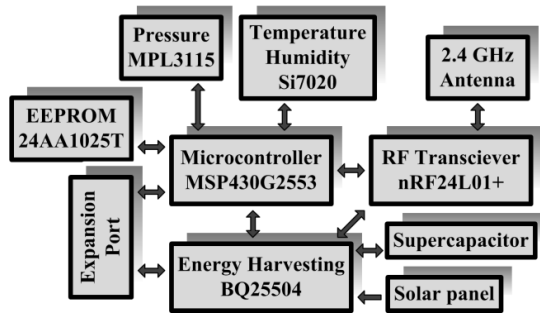


Figure 1. Important hardware subsystems and components incorporated into environmental WSN node

MSP430G2553 is a mixed signal microcontroller that is characterized by the following features: very low power consumption (Table 1), ultra-fast wake-up from standby mode in less than 1 μ s, and plenty of peripherals. In combination with MSP-EXP430G2 development board, this microcontroller is used by a large number of embedded system designers, which enables easy prototype development with lots of documentation available.

The nRF24L01+ is a single chip RF transceiver for the global, license-free 2.4 GHz ISM band. The low cost nRF24L01 is capable to merge very high speed communications (up to 2Mbit/s) with extremely low power (the RX current is just 12.5 mA). Output power, frequency channels, and protocol setup are easily programmable through serial peripheral interface (SPI) of MSP430 microcontroller. Current consumption for different operation modes is given in Table 1.

Since the WSN module is designed for applications in the field of environmental protection, sensors for pressure, temperature and humidity, as well as EEPROM for storing additional data are added to the basic module. The presence of these sensors makes writing firmware and module testing in real conditions easier. The sensors are selected in accordance with the requirements of minimum consumption and low cost (Table 1).

TABLE I. CONSUMPTION OF THE MOST IMPROTANT COMPONENTS OF WSN NODE AND THE APPROXIMATE PRICES

Power and Price table of designed WSN node			
Component	Working Mode	Consumption	Price
MSP430G2553	Active	230 μ A (1 MHz)	0.95€ ^a
	Standby	0.5 μ A	
NRF24L01P	Active	11.3 mA TX at 0 dBm 13.5 mA RX at 2 Mbps	1.44€
	Standby	26 μ A	
Si7020	Active	150 μ A	3.12€
	Standby	60 nA	
MPL3115	Active	2 mA	1.82€
	Standby	2 μ A	
24AA1025T	Active	450 μ A	2.57€
	Standby	5 μ A	
BQ25504	Quiescent	330 nA	2.45€

a. All prices are from Farnell and Mouser distributors calculated for highest quantity

The Si7020-A20 sensor of Silicon Labs is characterized by the following features: precision relative humidity sensor ($\pm 4\%$), high accuracy temperature sensor ($\pm 0.4^\circ\text{C}$), I²C interface and 3x3 mm package size. The Xtrinsic's MPL3115A2 employs a MEMS pressure sensor with an I²C interface to provide accurate pressure/altitude and temperature data. The Microchip's 24AA1025 EEPROM has been developed for advanced, low-power applications such as personal communications or data acquisition. For additional sensors there is expansion port. All microcontroller interfaces, except SPI, are available on the connector. This port is used to attach expansion board for water level measurement.

With a consumption of less than 50uA in standby mode and average consumption of about 3mA (without communication task), compared to commercial solutions is [7], implemented WSN node can be considered to be low power. With a total cost of less than 20 Euro could be considered less expensive compared to existing commercial solutions.

B. Subsystem for Energy Harvesting

In addition to low prices and consumption, one of the important requirements when designing WSN node is to minimize maintenance after final installation. Commercial WSN nodes generally use some form of battery power, which, depending on the modules, can last for a very long time, even up to several years. The batteries as limited energy supply must be optimally used for both processing and communication tasks. Since communication task tends to dominate over the processing task in order to make optimal use of energy, the amount of communication should be minimized as much as possible. In practical real-life applications, the wireless sensor nodes are usually deployed in hostile or unreachable terrains where they cannot be easily retrieved for the purpose of replacing or recharging the batteries, therefore the lifetime of the network is usually limited [8]. When the battery is finally depleted, it is necessary to invest additional funds for their replacement and expensive paid work in field conditions with a relatively large number of devices. The batteries themselves increase the cost of the device, especially if the special battery with high capacity and low self-discharge current is used. Also, as the network expands by employing many sensor nodes, the problem of powering the nodes becomes critical and even worse when one considers the prohibitive cost of providing power through wired cables to them or replacing batteries.

Given the above, it can be concluded that in order to reduce maintenance costs, it is necessary to find a way for the WSN node to work as long as possible. This can be achieved in two ways: (1) it is necessary to increase the capacity of energy sources or in our case to increase the energy density of the power source in order to minimize the design, and (2) to achieve minimum consumption, either by selecting the appropriate components or by firmware optimization. Hardware realization is already optimized for ultra low power application, while by firmware optimization only slight improvements can be achieved. Therefore the third option is increasingly considered, to develop energy harvesting techniques that

allows a WSN node to generate its own power by harvesting energy from the ambient. Energy harvesting are techniques that capture, harvest or scavenge unused ambient energy (such as vibrational or solar) and convert the harvested energy into usable electrical energy which is stored and used for performing sensing or actuation.

Energy harvesting system consists of: energy source, energy harvester circuit, power management circuit and energy storage element. Block diagram of energy harvesting subsystem is shown on Fig. 2. More information about energy harvesting technologies, models and possible applications can be found in [8]-[10].

In this paper, harvesting subsystem where the energy source is solar panel (photo-voltaic cell) was used. Energy harvesting from photo-voltaic cells is popular and well studied [8]. While photo-voltaic cells are most popular for energy harvesting systems because of the availability of solar irradiation, the efficiency of these cells is poor. In order to improve the efficiency of solar cell harvesting, methods such as maximum power point tracking (MPPT) are popularly used. A solar panel will generate different voltages depending on the different parameters like: the amount of sun light, the connected load and the temperature of the solar panel.

There are many types of MPPT techniques. These techniques differ in many aspects, including simplicity, convergence speed, hardware implementation, sensors required, cost, range of effectiveness and need for parameterization. Comparison of different MPPT algorithms from the energy production point of view is given in [11]. MPPT circuit operates on the basis of the maximum power transfer theorem to extract as much power as possible by impedance matching, in order to compensate for the varying characteristic resistance of solar panels (due to varying levels of insolation), thus providing higher power outputs.

Our energy harvesting circuit is based on a Texas Instrument's integrated chip BQ25504. The BQ25504 is an ultra low power charging controller intended for interfacing DC sources such as solar cells, thermal harvesters and high-impedance batteries.

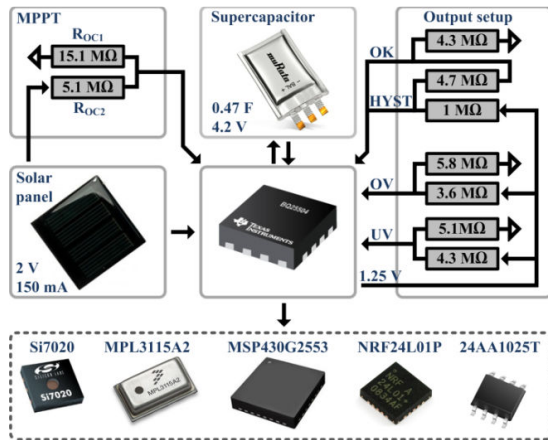


Figure 2. Energy harvesting architecture based on BQ25504 IC with supercapacitor as energy storage element ; important adjustments via external resistors are available to the user

The BQ25504's industry-leading low-quiescent current and high charger efficiency make it an ideal choice for charging batteries and supercapacitor. The BQ25504 charging controller uses the open-circuit voltage (OCV) technique to regulate the solar cell to its maximum power point (MPP). The MPPT circuit obtains a new reference voltage every 16 s (typical) by periodically disabling the charger for 256 ms (typical) and sampling a fraction of the harvester's open-circuit voltage. For solar harvesters, the maximum power point is typically 70%-80% [12].

For the available solar panel with the nominal voltage of 2 V and current of 150mA, the power and current as a function of voltage are shown in Fig. 3. The exact ratio for MPPT can be optimized to meet the needs of the input source being used by connecting external resistors R_{OC1} and R_{OC2} between output of the solar panel and ground with mid-point tied to pin of charger controller. In our case, the MPP is set to 68%.

Among other parameters it is possible to adjust the undervoltage (UV) threshold to prevent storage component to be deeply discharged or damaged. Also, overvoltage (OV) protection is configurable.

As the main information source about the state of harvester controller a digital output signal (OK) is used. This signal can awake microcontroller from sleep by using interrupt service. Both signal level and hysteresis (HYST) are configurable via external resistors. The load is not directly connected to the storage element, but internally via an internal MOSFET inside the controller. The controller on the basis of the external resistors values and internal references regulates the flow of power to and from the storage element. Voltage references of interest and actions that are executed according to these are shown in Fig. 4. Charging the storage element begins when the output of the controller reaches $V_{BAT_UV_HYST}$ (2.3 V + 80 mV), the load being supplied only when it reaches $V_{BAT_OK_HYST}$ (2.9 V) and the controller turns off when it reaches V_{BAT_OV} (3.1 V), when the overvoltage protection is activated. In the opposite direction, when the voltage drops below overvoltage threshold with hysteresis $V_{BAT_OV_HYST}$ (3.1 - 35 mV), charging resumes and the load is powered up, and when the voltage drops below V_{BAT_OK} (2.6 V), the load is separated from the controller, but the charging continues until the voltage falls below undervoltage threshold V_{BAT_UV} (2.3 V)

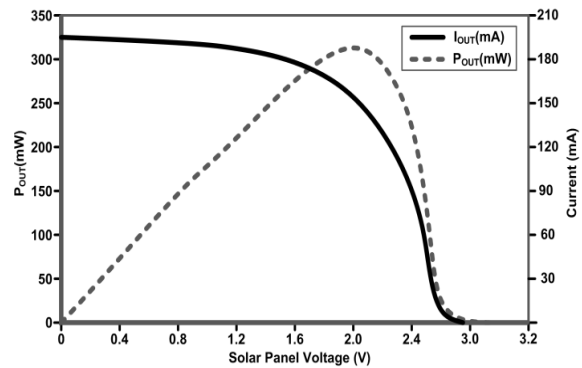


Figure 3. Voltage versus current, and voltage versus power, of the solar panel with the nominal voltage of 2 V and current of 150 mA

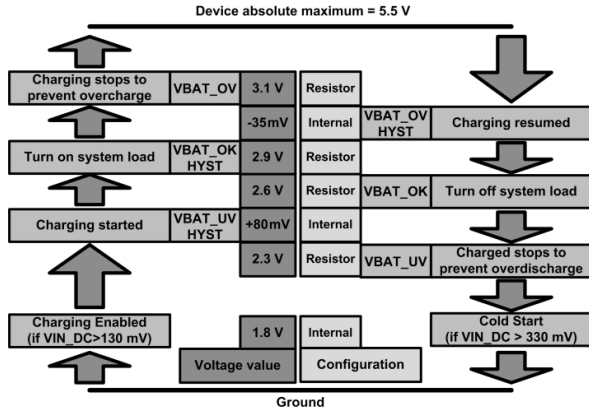


Figure 4. Relative position of the various threshold voltages that affects on charging/discharging of supercapacitor and supplying of load

C. Why supercapacitor?

WSN node described in this paper is designed for use in the field conditions to monitor environmental parameters. Most of the environmental parameters are changing slowly so the measurement frequency does not have to be high. Taking this into account, we have considered the use of supercapacitor as the main power source of WSN node. More details about the analysis and application of supercapacitor in WSN can be found in [13]-[15]. In the final hardware implementation Murata supercapacitor with capacitance of 0.47 F and nominal voltage of 4.2 V is used (Fig. 5).

Based on the values in Table 1, we calculate that the consumption in the active state is about 15 mA with activated communication task (in standby less than 50 μ A, with sensors). If we choose communication task duration of one second (less in practice) to occur every 60 minutes the average consumption will be 0.3 mA. For the mentioned supercapacitor of 0.47 F for voltage change from 3.1 V to 2.6 V (Fig. 4), we obtain that WSN node can operate for about 13 minutes. Real tests showed that this value is actually around 10 minutes because of additional consumption of other passive components that are not listed in Table 1. Note that the used supercapacitor has a planar structure to reduce the total node volume.

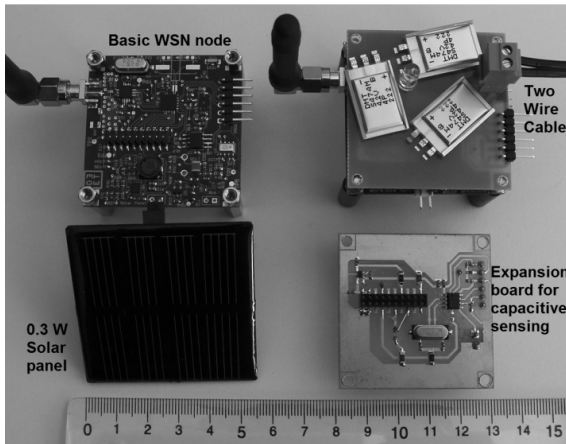


Figure 5. Photo of environmental WSN node with energy harvesting and expansion board for capacitive sensing.

On the basis of these data may be concluded that the WSN node is not applicable to the measurement in night conditions and also in daily conditions if the solar panel is unable to provide output current greater than 0.3 mA which rarely happens because the panel with a nominal voltage of 2 V and maximum current of 150 mA is used. The conclusion is that the designed WSN node can be used without restrictions in the daily conditions, which further implies that it can be used only for measuring parameters that do not oscillate drastically during the day and night, otherwise their measurement is relevant only during the day. Monitoring of surface water levels is slowly varying process, and thus developed WSN node makes appropriate choice for these applications.

Since each WSN node is equipped with energy harvesting system, consumption in different network topologies is not considered. Lower consumption in this case affects size of the design and smaller and cheaper solar panel can be used.

III. CAPACITIVE WATER LEVEL MEASUREMENT

For water level measurement simple expansion board is designed based on Microchip's PIC12F1822 microcontroller. In order to reduce the consumption, a special Texas Instrument's TPS22944 load switch, with enable signal controlled by main MSP430 microcontroller is used. In this way, during the long periods of time when measuring the level of water is not performed, the consumption of the part of hardware which executes this task is practically reduced to zero. In the conditions of using limited resources of alternative energy sources, this represents a significant advantage in node's performance.

PIC12F1822 has integrated capacitive sensing (CPS) module. Isolated cable is connected microcontroller (Fig. 6). Capacitance of isolated cable partially immersed in water is read by microcontroller's CPS oscillator module [16]. Cable acts as the capacitor (C_{sen}), whose capacitance changes linearly with length of the cable immersed in water. Dielectric constant around cable changes, which results in the equivalent cable capacitance change. The time dependence of the capacitor voltage (V_{sen}) can be described with

$$V_{sen}(t_1) = V_{sen}(t_0) + \frac{1}{C_{sen}} * \int_{t_0}^{t_1} i(t) * dt \quad (1)$$

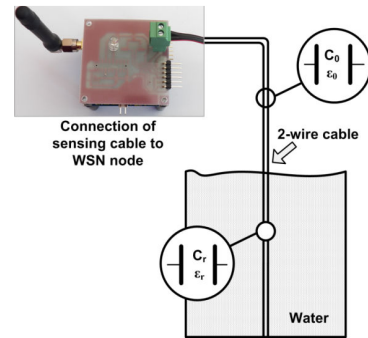


Figure 6. Typical installation of WSN node where the capacitance of 2-wire cable immersed in water is measured

By using constant current, the equation (1) is simplified to

$$V_{sen}(t_1) = V_{sen}(t_0) + \frac{I_{cps}}{C_{sen}} * (t_1 - t_0) \quad (2)$$

If we take into account the fact the charging and discharging currents (I_{cps}) are the same, by measuring the frequency of oscillation of the CPS oscillator, we can express rise (t_{rise}) and fall time (t_{fall}) as

$$t_{fall} = t_{rise} = \frac{1}{2 * f_{CPS}} \quad (3)$$

Cable capacitance per unit length is specified by manufacturer: in the air $C'_a = 55.13 \frac{pF}{m}$, in the water $C'_w = 296.33 \frac{pF}{m}$.

The total capacitance to be measured is

$$C_{ekv} = C_0 + L_a * C'_a + L_w * C'_w \quad (4)$$

$$C_{ekv} = C_0 + (L_c - L_w) * C'_a + L_w * C'_w \quad (5)$$

where C_0 denotes parasitic capacitance of the PCB tracks and the cable connector, L_a length of the cable in the air, L_w length of the cable immersed in the water and L_c total length of the cable. The firmware measures the operating frequency of the CPS oscillator. We can determine the total capacitance of the cable (C_{cable}) using

$$C_{cable} = \frac{I_{CPS} * (t_1 - t_0)}{V_c(t_1) - V_c(t_0)} \quad (6)$$

where t_1 represents the moment of reaching a positive reference voltage of CPS oscillator, while t_0 is the moment of reaching negative reference voltage of CPS oscillator. Thus formula (6) comes down to

$$C_{ekv} = \frac{I_{CPS} * t_{rise}}{V_+ - V_-} = \frac{I_{CPS}}{(V_+ - V_-) * f_{CPS} * 2} \quad (7)$$

In this paper the following configuration of CPS oscillator is used: $V_+ = 2048$ mV which is the internal reference voltage and $V_- = \frac{10}{32} * 2048$ mV.

The negative reference is obtained by using an internal 5-bit DAC converter which also uses the internal reference voltage of 2048 mV. The current source of CPS oscillator is software adjustable and is set to 9 μ A.

Substituting these values in (7), we get the expression for the capacitance of the sensor cable

$$C_{ekv} = \frac{3.196 * 10^{-6}}{f_{CPS}} [F] \quad (8)$$

Using (5), we can calculate the water level, i.e. the length of cable immersed in water.

$$L_w = \frac{C_{ekv} - C_0 - (L_c * C'_a)}{C'_w - C'_a} \quad (9)$$

The reference capacitor of 470 pF capacitance and 5% tolerance that is soldered on the PCB is used for testing purposes. In this case due to the short conductive track lines in this case C_0 can be neglected. By measuring the frequency of the CPS oscillator for this channel, the value

of 6.885 kHz is obtained and using (8) the measured capacitance value of 464.2 pF is calculated, which represents a deviation of 1.27% which is within the tolerance value of testing capacitor used.

IV. RESULTS

The device was tested in a calm section of the Danube River. Sensing cable is marked every 10 cm and immersed into the water. For each marked point minimum 10 measurement results are recorded. Substituting the known parameters of the sensor cable, a graph in Fig. 7 is produced. It shows the deviation of the read measurements from expected capacitance value. During measurements the adjustment of the level to which the cable is immersed in water was done manually, so there is a significant impact of human factor on the measurement result errors. Measurement error in percentage of full scale is shown in Fig. 8.

This graph shows that the largest percentage error occurs at higher values of the immersion depth of the cable into the water. This can be explained by the influence of human error during the device calibration in inconvenient field conditions and difficult access to water in the measurement range. The main disadvantage of this solution is the existence of recovery time after removing a sensor cable out of the water. Due to the remaining water drops and moisture on the cable surface after extraction a false reading may occur.

The impulse response of the system is determined by immersing the entire length of the sensor cable into the water and subsequent rapid withdrawal of the entire cable. This impulse response is shown in Fig. 9.

Considering that this slow response is noticeable only when the cable is removed from water and only for drastic water level changes, we come to the following conclusions:

- System response during immersion in water is detectable immediately.
- When removing the cable from the water the response is exponential.
- The measured signal reaches 95% of the asymptotic value in the first 5 seconds of measurement.
- This system is efficient for slowly varying surface water levels.

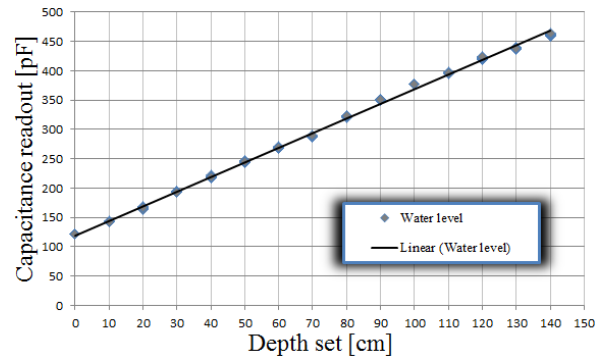


Figure 7. Measured and expected values versus immersion depth

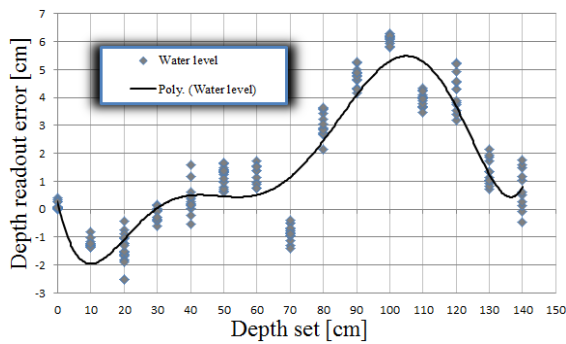


Figure 8. Absolute measurement error during calibration

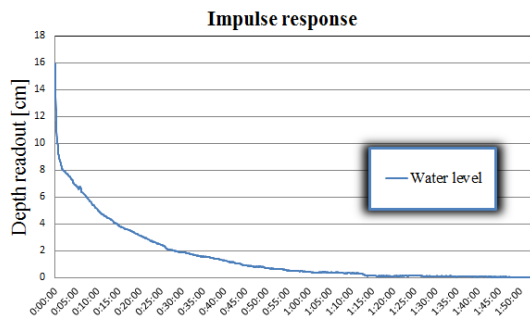


Figure 9. Water sensor response for sudden water level change

The purpose of this sensor node is monitoring of water level in slowly varying systems, and the slow response during the decline of the water level has no greater influence on the measurement results. When water level changes fast, this method is not recommended.

V. CONCLUSION

The proposed wireless sensor network node ensures reliable water level measurement using energy harvesting from solar panels as the energy source and supercapacitor as the storage element that enables long-lasting autonomous operation and minimal maintenance. Wireless communication provides transmission of measured parameters to a remote location and easy integration into existing systems for data logging. The low cost of production opens the possibility of employing larger number of sensor nodes and their interconnection. The measurement results in real field conditions showed satisfactory precision and accuracy of measurements for the intended purpose.

ACKNOWLEDGMENT

This paper is part of the project "Development of the methods, sensors and systems for monitoring quality of water, air and soil", III43008. Project has been carried out with the financial support of the Ministry of Science and Education of the Republic of Serbia, to which authors are very grateful.

REFERENCES

[1] B. Merz, H. Kreibich, R. Schwarze, A. Thieken, "Assesment of economic flood damage", Natural Hazards

and Earth System Science, vol. 10, pp. 1697-1724, August 2010.

[2] M. Rossi, D. Brunelli, "Ultra low power Wireless Gas Sensor Network for environmental monitoring applications", Environmental Energy and Structural Monitoring Systems (EESMS), 2012 IEEE Workshop on, pp. 75-81, September 2012, Perugia.

[3] M. Mukherjee, "Early Warning System deploying wireless sensor network for flood management", International Journal of Innovative Technology & Adaptive Management (IJITAM), vol. 1, December 2013.

[4] N. Ahmad, M. Hussain, N. Riaz, F. Subhani, S. Haider, K. S. Alamgir, "Flood Prediction and Disaster Risk Analysis using GIS based Wireless Sensor Networks, A Review", Journal of Basic and Applied Scientific Research, vol. 3, pp. 632-643, 2013.

[5] C. Alippi, R. Camplani, C. Galperti, M. Roveri, "A Robust, Adaptive, Solar-Powered WSN Framework for Aquatic Environmental Monitoring", IEEE Sensors Journal, vol. 11, pp. 45-55, January 2011.

[6] S. S. Sonavane, V. Kumar, B. P. Patil, "MSP430 and nRF24L01 based Wireless Sensor Network Design with Adaptive Power Control", ICGST-CNIR Journal, vol. 8, pp. 11-15, January 2009.

[7] J. M. Gilbert, F. Balouchi, "Comparison of Energy Harvesting Systems for Wireless Sensor Networks", International Journal of Automation and Computing, vol. 5, pp. 334-337, October 2008.

[8] R. V. Prasad, S. Devasenapathy, V. S. Rao, J. Vazifehdan, "Reincarnation in the Ambiance: Devices and Networks with Energy Harvesting", Communications Surveys & Tutorials, IEEE, vol. 16, pp. 195-213, July 2013.

[9] Z. G. Wan, Y. K. Tan, C. Yuen, "Review on energy harvesting and energy management for sustainable wireless sensor networks", IEEE 13th International Conference on Communication Technology (ICCT), pp. 362-367, September 2011.

[10] M. Hassanaliagh, T. Soyata, A. Nadeau, G. Sharma, "Solar-supercapacitor harvesting system design for energy-aware applications", 27th IEEE International System-on-Chip Conference (SOCC), pp. 280-285, September 2014, Las Vegas.

[11] A. Dolara, R. Faranda, S. Leva, "Energy Comparison of Seven MPPT Techniques for PV Systems", Journal of Electromagnetic Analysis and Applications, vol. 1, pp. 152-162, July 2009.

[12] S. Zhou, L. Kang, J. Sun, G. Guo, B. Cheng, B. Cao, Y. Tang, "A Novel Maximum Power Point Tracking Algorithms for Stand-alone Photovoltaic System", International Journal of Control, Automation, and Systems, vol. 8, pp. 1364-1371, January 2011.

[13] S. Kim, P. H. Chou, "Size and Topology Optimization for Supercapacitor-Based Sub-Watt Energy Harvesters", IEEE Transactions on Power Electronics, vol. 28, pp. 2068-2080, June 2012.

[14] C. Park, P. H. Chou, "AmbiMax: Autonomous Energy Harvesting Platform for Multi-Supply Wireless Sensor Nodes", Sensor and Ad Hoc Communications and Networks - SECON, vol. 1, pp. 168 -177, September 2006, Reston.

[15] J. F. Christmann, E. Beigne, C. Condemine, J. Willemin, "Energy harvesting and power management for autonomous sensor nodes", 49th IEEE Design Automation Conference (DAC), pp. 1049-1054, June 2012, San Francisco.

[16] (Meco)V. Milosavljevic, Z. Mihajlovic, V. Rajs, M. Zivanov, "Implementation of low cost liquid level sensor (LLS) using embedded system with integrated capacitive sensing module", Mediterranean Conference on Embedded Computing (MECO), pp. 58-61, June 2012, Bar.

Insider Threats in Information Security

Categories and Approaches

Nebrase Elmrabit¹, Shuang-Hua Yang¹, Lili Yang²

¹Department of Computer Science, ²School of Business & Economics

Loughborough University, UK

N.Elmrabit@lboro.ac.uk

Abstract—The main concern of most security experts in the last years is the need to mitigate insider threats. However, leaking and selling data these days is easier than before; with the use of the invisible web, insiders can leak confidential data while remaining anonymous. In this paper, we give an overview of the various basic characteristics of insider threats. We also consider current approaches and controls to mitigating the level of such threats by broadly classifying them into two categories.

Keywords— *Insider threats; data leaking; privileged user abuse; insider attacks; insider predictions.*

I. INTRODUCTION

Organisations nowadays depend on computers in every aspect of their daily operating, and because more than 80% of companies use remotely hosted services on the cloud[1], most governments have started to centralise citizens' information in huge data service centres, while the citizens themselves also rely on cloud computing to store their confidential data. All this makes data theft easier. Most of the decision makers in organisations and government are focusing on external cyber-attacks such as unauthorised access to their networks, denial of service attack, viruses, Trojan Horse, Worm, etc. In order to protect such networks from external attack, they spend around 10% of their IT budget on securing their assets[1]. However, new evidence shows that both external attacks and insider threats are significant[1], while the damage caused by insider attack is more damaging than that of outsider attacks[2]. This means that anyone who has authorisation to access organisation's data assets is more dangerous than any other security threat.

Insider attacks are the most expensive form of information security breach, in that the average cost per insider incident is £250,000 according to a recent report by the INSA[3]. This is because the insider has knowledge of, and access to, their employer's assets. This has come about because such an individual has had the trust of the organization causing him or her to be supplied with authorised access so that it is possible to bypass all physical and electronic security measures. However, the number of insider threat incidents has continued to increase to significant extent. In fact, a recent study by the Ponemon Institute found that 88% of IT experts believe that the risk of insider threat will stay the same or increase in the next two years[4]. But more than three-quarters of these incidents usually go unreported and are handled internally, with few referrals to law enforcement agencies and no legal action been taken[5].

Over the last ten years, there have been numerous studies that have tried to define insider threat problems in order to come with one solution to solve current security

data breaches. In addition, security research incorporating survey results in the last three years has shown an increase in insider threat breaches, with a strong level of incident effects on organisations from the activities of insiders.

To comprehend the definition of an insider threat, we should know what an insider is and what a threat means in relation to information security. *An insider* - "Is a person that has been legitimately empowered with the right to access, represent, or decide about one or more assets of the organization's structure"[6], simply as: an individual who has authorised access to an IT system. *A Threat* - refers to anything that has the potential to cause serious harm or damage to an organisation's IT systems or assets.

A definition of Insider Threats is—

(a) Any malicious activities that cause damage to an organisation's IT and network infrastructure, applications, or services - (b) On the part of an employee (current or former), contractor, subcontractor, supplier, or trusted business partner- (c) Who has or has had authorised access to the organisation's IT assets - (d) And poses a significant negative impact on the information security elements (confidentiality, integrity, and availability) of the organization.

Information Security can be defined as the process by which digital information assets are protected in order to ensure the main security goals. These are : a) *Confidentiality* - to ensure that information assets are not disclosed to individuals or systems that are not authorised to receive them[7]. It is also defined as the process of making sure that data assets remain secret and confidential, and that they cannot be viewed by unauthorised users, b) *Integrity* - To ensure that information assets cannot be modified by any other party without authorisation. Integrity could also be described as the process that ensures that data assets are the same as they were when they were originally created, without any change over time, and c) *Availability* - To ensure that information assets are available when requested. It could also be described as a situation in which data assets should be accessible for legitimate users when needed.

Reason for misuse - Based on Wood's assumption[8], an insider threat requires three factors when it comes to the attacker misusing his privileges: a) an insider attacker must have the motivation to attack "*a motive*", b) he must identify a target "*an opportunity*" and c) he must be able to launch an attack "*a capability*". A recent study by Colwill[9] reports that "insider attacks are made with varying degrees of motivation, opportunity and capability. Motivation will come from internal, personal drivers, whereas opportunity and capability will be given to insiders overtly by your organisation to perform their role, or may be attained covertly once they are on the inside".

II. INSIDER THREAT CATEGORIES

It can be divided into seven sub-categories as shown in Table 1, based on the manner in which they affect the organisation's information security goals, and the human factors which lead an insider to act in a malicious manner. We can also name the insider threat categories in term of the impact and the actions that the insider use to achieve his aims. These are: a) insider IT sabotage, b) insider IT fraud, c) insider theft of intellectual property, d) insider social engineering, e) unintentional insider threat incident, f) insider in cloud computing, and g) insider national security[5][10][11][12]. However, organisations could be affected by more than one category of malicious insider threat at the same time.

A. Insider information technology Sabotage

Insider IT sabotage are attacks in which the insider uses his/her IT experience and knowledge to launch an attack on an individual or an organisations. In general the attacker mainly targets the availability of the IT and network infrastructure, applications and services, when they feel they are under pressure or stress from their organisation or from colleagues. In general, insider IT saboteurs are former employees, working remotely, without authorised access to target systems, working outside normal hours, who prepare themselves and plan the attacks, and use tools to launch such attacks. The main targets are databases, systems, and network devices.

From the CERT insider threat cases database, an employee spread rumours across his organisation, that annual bonuses would be smaller than in previous years. This drove a malicious IT employee to design and program a logic bomb from a remote distance. He used authorised VPN access to move the malicious program to all company servers as the foundation for his revenge if the rumour is proved to be true. After he found out that the company was going to reduce the annual bonuses of all staff, he resigned, and then set the logic bomb to go off two weeks later. This deleted company files and disrupted thousands of servers across the USA. However, the insider was convicted and sentenced to more than eight years in prison[5]. It is clear that this piece of IT sabotage was caused by an employee wanting revenge on his organisation in order to achieve self-satisfaction. Usually the employee has high stress levels caused by his organisation, or is aware of the danger of losing his job.

B. Insider IT Fraud

Insider IT fraud is when an insider uses authorised access for personal gain. This abuse can be in the form of creating, modifying, deleting or, in some cases, selling confidential data assets. This fraud also affects data asset confidentiality and integrity. Insider fraudsters in general are current employees, working in an office, who has authorised access to information assets, is in a non-technical position, who operates during normal hours, and who has no need for tools to launch the attack. The main insider target is information assets.

A case study of insider IT fraud in the UK [1], shows how a malicious insider working for a large utility company, having authorised access to sensitive company information could harm the organisation's confidentiality and profits by selling customer data asset to competitors.

However, the organisation accidentally discovered this breach after month following a huge financial impact on their business. It is clear that IT fraud is caused by the greed of employees who work to benefit themselves for financial gain. Usually the employee is suffering from high financial pressures caused by the outside environment, and is unable to solve the problem through legitimate means. This is what motivates the fraud crime in the first place.

Table 1. Insider Threat Categories

Impact	Effect to Organisation Information Security			Human Factors to Act a Threat		
	Confidentiality	Integrity	Availability	Motive	Opportunity	Capability
IT Sabotage	Low	Med	High	High	Med	Med
Fraud	Low	High	Low	High	High	Med
Theft of Intellectual Property	High	Low	Low	High	High	Med
Social Engineering	High	High	High	High	Low	Low
Unintentional	Med	Med	Med	No	Low	Low
Cloud Computing	Med	Low	Low	High	High	Low
National Security	High	High	High	High	High	High

C. Insider Theft of Intellectual Property

Insider theft of intellectual property (IP) is when an insider uses the IT infrastructure to engage in espionage or steal information created and owned by the organization which employs him. Insider thieves of IP in general are current employees, or employees working in their resignation notice period, working in the office, who has authorised access to IP. They tend to hold technical positions such as scientists, programmers, engineers, or sales, during normal hours, and do not need tools to launch an attack. The main insiders targets are source codes, business plans, strategic plans, product information, and customer information[5].

In a case study of the theft of intellectual property in September 2013, a mobile telecommunication company in Germany suffered a data breach caused by an insider who had close knowledge of their IT infrastructure and system, he managed to take a copy of more than two million customers' records, such as customer names, customer addresses, date of birth and bank account details[13]. The theft of IP is usually done by someone who has been a part of the process that creates the organisation's IP. They think that the information asset belongs to them. Other types of people who steal IP are those who want financial gain for themselves.

D. Insider Social Engineering

Insider social engineering (SE) is when malicious insiders act to psychologically manipulate another innocent employee without their knowledge to disclose confidential information or perform an action to harm the organisation's IT, network infrastructure, applications or services. However, insider SE occurs when the insider or outsider does not have the authorisation to access part of, or all of, the organisation's assets. Insider SE in general involves an employee or outsider, using psychological manipulation, working inside normal hours, preparing themselves and planning before the attack, involving a human-based and technology-based attack. It may be a multiple-stage attack, on the part of individual who do not have authorisation access to target systems, and uses phishing tools to launch the attacks. The main targets are access usernames and passwords to a database, systems, services, and network devices.

From the CERT threat cases database, government organisations have been the target to insider SE, in that employees have been tricked by a phishing email sent to them regarding human resource benefits that exploited zero-day vulnerability and downloaded malicious code. The code hides itself on the target system and acts as the back door for the outsider allowing the malicious outsider to transfer government information[11]. It is apparent that insider SE is caused by someone who has no authorised access to the target systems, and whose main reason for SE is to sabotage the IT system, steal intellectual property, or commit fraud.

E. Unintentional Insider Threat Incident

An unintentional insider threat incident is one in which an authorised user accidentally performs an action to harm the organisation's IT and network infrastructures, applications or services, without the motive or intention to mount a malicious attack[14]. Unintentional insiders in general are current employees, working in the organisation's office during normal hours, who have authorised access to the target system, who causes an unplanned incident, without a target or malicious motive.

A mistake by an accounts manager working in a pharmacy company in the USA drove her company to fire her after performing an accidental security breach. The unintentional insider downloaded a file containing the prescription information of 6,000 patients with full patient details onto a memory stick, which she then lost. There is no doubt that unintentional insider threat incidents occur when the victim has no security awareness training, poorly understands organisation security policy, poor management systems, work under high job pressure or stress, is involved in difficult tasks with a lack of knowledge, and uses drugs[14][10].

F. Insider In Cloud Computing

Insider in cloud computing or insider in service providers, are those working inside service provider company environments, who perform malicious insider actions without the client's knowledge in order to harm their data asset confidentiality. However, there are neither possible ways of detecting such an attack during or even after the breach, as the client has no control over service provider infrastructures or any effective method and tools

to prevent such an attack. Insiders in the cloud in generally current employees, working in a technical position, during normal hours, who have fully authorised access to target infrastructure, who are well planned, and have a malicious motive. The main insider targets are data assets such as databases, source codes, business plans, and strategic plans[12][15][16].

In a case study, an experienced IT administrator, working for a cloud computing server provider, used his skills to act as a malicious insider. He managed to take a copy of a client's virtual machine file as part of his duties, and then he broke into the client's administrator account by using password cracking tools. This gave him full access to the client's operating system on the virtual machine without the client's knowledge. Malicious threats from inside the cloud computing providers and caused by their employees are increasing. Using their authorised access rights to the environment, they commit security breaches such as file recovery, coping virtual machine files, and removing disks from a RAID.

G. Insider National Security

Insider national security (NS) threats involve an insider using their authorised access to represent a threat or do harm to a country's NS. This threat can include damage to the country through espionage, sabotage, disclosure of NS information, or through the loss or degradation of departmental resources or capabilities. Their main targets are the NS secret information.

The biggest intelligence leak in U.S. history was launched by a malicious insider (a trusted IT contractor) who worked for the NS Agency (NSA). Edward Snowden managed to download millions of documents on classified intelligence collection programs, as he had the authorised access to mass electronic surveillance data as part of his job. Then he leaked classified material to media outlets. Since then he has released details of unwarranted NSA hacking of friends and foe alike, the fallout damage U.S. relations abroad and putting a spotlight on current security issues facing the U.S. Insider NS threats usually come from insiders as they have the trust of the government. The motivations for their malicious actions are money, psychology, accident, revenge or, as in Snowden's case, "My sole motive is to inform the public as to that which is done in their name and that which is done against them, I do not want to live in a world where everything I do and say is recorded"[17].

III. RESEARCH LITERATURE REVIEW ON INSIDER THREAT APPROACHES

In this section, various approaches towards insider threats and controls are presented in order to explain how we could mitigate insider threat. These approaches can be broadly classified into two categories: *a) technical mitigation approaches* and *b) non-technical mitigation approaches*. However, most malicious insider threat activities are still detected by individuals who are not part of the organisation's security staff, with only one in five activities detected using a combination of automated tools for logging, monitoring and flagging suspicious activity, along with manual diagnosis and analysis [18].

A. Technical Controls to Identify Insider Threats.

In general technical controls are divided into two main categories: a) those that look for *unauthorised malicious activity*, and b) those that look for *changing in behaviour* that may indicate a malicious insider. In addition to this, technical control tools could be implemented to concentrate on: a) network-based activities, b) host-based activities, or c) cloud-based activities.

1) Intrusion Detection Systems

The National Institute of Standards and Technology (NIST) [19] defines *Intrusion Detection (IDS)* as “the process of monitoring the events occurring in a computer system or network and analysing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices”.

IDS are deployed to detect malicious intruders in real time originate from external threats, and are based on monitoring networks or endpoint devices through analysing activities and traffic patterns from any abnormal behaviour in the network and endpoint, or through matching the activities and traffic with a database of attack signature. When IDS detects abnormal behaviour or an attack signature it displays a security alert. As IDS gathers information over different platforms in real time, it is a helpful tool for discovering a malicious insider by analysing information of any change of user behaviour or activity that may lead to data breaches[20].

However, IDS has its limitations in dealing with insider threats such as: a high number of false alarms, a huge database log file size, and requiring an administrator to analyse the traffic and behaviour. In addition, it cannot monitor encrypted traffic[18]. Furthermore, Cyber-Security Centre at the University of Oxford[12] concluded that “IDSs are far from ideal for detecting insiders as they are primarily focused on external attackers and have a tendency to identify false positives”.

2) Security Information and Event Management

Security Information and Event Management (SIEM) is a tool that responsible for centralising and analysing logging in one management platforms. It does this by collecting information through secure network channels from various security-related logs (ranging from client workstations and servers to application servers, antivirus software, network devices, honeypots, firewalls, IDSs), and any other sensors in the network, then correlating the events among them in a database by matching any related characteristics and events[2][19][21][22].

This approach allows the information security administrator to quickly search for events and possibly identify malicious insider activity before it occurs, or as data-mining and evidence for forensic investigations after the accident occurs[19] [23] [24].

3) Data Loss Prevention

Data Loss Prevention (DLP) is a technology responsible for the early detection of data exfiltration attempts by an insider. It is performed in three steps:

a) *system discovery entails* scanning storage devices, capturing network data flow, and watching user behavior on endpoint devices. b) *leaked confidential data identification* information discovered in the first step could

be identified as secret information in three ways: keyword matching, regular expressions, or hashing fingerprinting. c) *organisation policy enforcement* - this step prevents any action that could cause any security breach in identified confidential data in the previous step[22].

The benefit of using a data loss prevention approach is that we can use it to protect three types of data in an organisation, or just a part of any type, depending on business need. These types are:

a) *Data at rest* - refers to inactive data or static data that is stored physically on enterprise devices. b) *Data in motion* - refers to data captured in the moment of data traffic flow. c) *Data in use* - refers to active data assets under constant change, “data in operation” as they are processed by applications or endpoint agents. However, these days research groups use this technology to deploy new insider threat potential approaches such as: web traffic inspection[25]; Virtual Private Network (VPN) data flow monitoring[5]; and Correlating Events from Multiple Sources such as Universal Serial Bus (USB)[23].

4) Access Control System

Access control (AC) is the system that manages and controls the access credentials to specific electronic resources based on a) *authentication* “who you are”, and b) *authorisation* “what you are authorised to do” components, in relation to the security policy of an organisation. The rules are based on different principles such as: a) *least privilege*, b) *privilege escalation*, and c) *separation of task duties*[26][27].

Whether using Role-Based Access Control (RPAC), Mandatory Access Control (MAC), or Discretionary Access Control (DAC) models, the insider threat is granted access by system authentication and is authorised to perform the necessary tasks. An AC system ensures that an organisation’s security administrators have control of their asset and they can change the authorisation access level, or deny access at any time, when needed [28].

5) Honey-tokens

A *honey-token* is a method used to attract malicious insiders, and helps to detect, identify and confirm a malicious insider threat[29]. Moreover, it may be effective in catching insiders who are snooping around a network. The honey-token is a technique that is a part of honeypot technology. However, it is different to other types because it could be any interactive digital entity, such as a Microsoft Office document, rather than a hardware device or software. The main concept is that no one should interact with the trap, and any interaction with the digital entity will indicate to the security administrator that there could be the threat of a malicious insider.

As an example, if a company general manager (GM) suspect that one of his IT staff is checking his emails, owing to the fact that an IT employee has full authorisation to access to emails, then they could use the honey-token approach to generate an email to the GM. This email should contain interesting information to attract an insider. Then, this honey-token leads the insider to use a user-name and password within the email to access the honey-token, as no one else has the username and the password. When a malicious insider accesses the URL, insider information such as the IP address, will be sent to the IT security team to deal with this breach.

B. Non-Technical Approaches

From the fact that insider threat “Is a people problem” and “The trust we give”, mitigating the threat level of a malicious insider is a difficult issue that requires dealing with human behaviour, instead of only dealing with this issue by using a technical approach. As we have seen in the past five years whistle-blowers have managed to avoid all major technical controls. At this point, institutions and researchers should start to look into the problem of insider threats from different points of view, such as: prediction, training and awareness, and security policy.

1) Psychology Prediction Model

Based on the psychology of user behavioural, researchers have found psychology indicators related to a malicious insider threat. These three factor are: a) *a motive*, b) *an opportunity* and c) *a capability*[30][31].

Axelrad et al.[32] proposed a model to predict insider threats. The motivation behind their approach is to define 83 psychological variables potentially associated with insider threats. The approach was to analyse these variables and estimate a score power to each variable. Variables include: (a) *dynamic environmental stress*, such as life and job stress; (b) *personal characteristics*, such as job satisfaction; (c) *insider actions*, such as personal attitude; and finally, (d) *the degree of interest*, such insider threat profile. Greitzer et al.[33][34] proposed another classification method for malicious insider threats based on the case studies of previous insider crimes. Their approach began with setting 12 indicators associated with insider threats. Figure 1 shows Greitzer’s risk indicators classified by the weights of the indicator to risk levels.

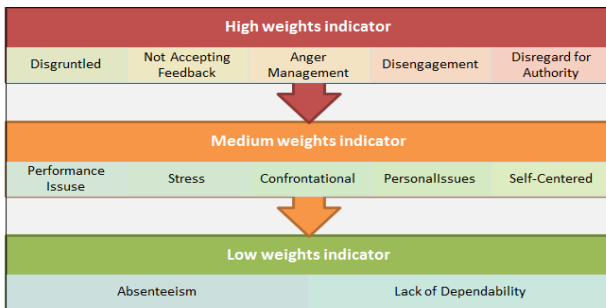


Figure 1. Greitzer’s risk indicators [33].

Both models in last two sections help decision makers to determine whether the user is a potential malicious insider threat or not, based on scoring indicators[35].

2) Security Education and Awareness

Insider threat accidents could be avoided by the appropriate security education and awareness training[36], especially the category of the unintentional insider threat. The Ponemon Institute[4] reports that 62% of organisations conduct regular privileged user training programmes as part of their efforts to protect the organisation from insider threats, with 11% of the IT budget allocated to security education and awareness.

Educational and awareness training could include the following area: a) presentations by outside speakers, b) classroom courses, c) on-line training courses, d), and e) printed leaflets. Training objectives may include: a) incident reporting procedures and responsibilities, b) consequences and sanctions, c) handling of sensitive

information, d) intellectual property protection, e) insider threat indicators, f) social engineering scams, and g) unintentional leaking[3][37][11].

3) Information Security Policy

An organisation’s information security policies deliver the framework that sets the most critical controllers within the organisation once the organisation’s objectives have been identified. It comes in a detailed statement of employees’ expectations of an organisation, and what is expected from them in terms of information security, and the acceptable behaviour and culture within the organisation [38] [39] [40].

A recent paper by the Cyber Security Centre at the University of Oxford[10] focused on the ability of an organisation’s information security policies to mitigate the level of a malicious insider threat. In their paper they pointed out the fact that the risk of an unintentional insider threat is potentially more pressing than that posed by other malicious insider categories. From this point, they found that 45% of employees do not follow security policies for two main reasons: a) the policy was incomplete or poorly defined; or b) the employee was not aware of the security policy. If not following information security policy the insider threat level will increase.

IV. FINAL REMARKS AND FUTURE RESEARCH

In previous years, organisations, governments and armed forces all around the world, have failed to mitigate malicious insider threats through their regular security measures. Moreover, these kinds of security breaches have started to affect our entire society. In this paper we have reviewed and presented various characteristics and categories of insider threats, by dividing the insider threat category into seven sub-categories, based on the manner in which they affect the organisation’s information security goals, and the human factors which lead an insider to act in a malicious manner. We have also considered some of current approaches and controls associated with mitigating the level of insider threat, by classified them into two main categories: technical mitigation and non-technical mitigation approaches.

We have found that there is no solution which can fully eliminate insider threat within organisations. Also, a technical approach by itself may not be the most effective way to prevent and/or detect malicious insider threats.

Dealing with trusted is a difficult issue that involves researchers looking at the problem from a holistic perspective in terms of: a) Human behaviour, b) Technologic controls and c) Organisational aspects. Future research is highly recommended by different institutions to cover the following points:

- More in-depth research is needed, to discover the cause behaviour that drive privileged users to act malicious insider threat. This will lead to the identification of the mitigating factors and indicators of malicious or accidental insider threats.
- To development a comprehensive and useful prediction model, which can convert the raw data inputs from the above holistic perspective to the risk level associated with a malicious insider threat.
- Finally, research is required to identify the effects of information security policy to organisations in term of insider threat risk levels.

REFERENCE

- [1] C. Potter and A. Miller, "Information Security Breaches Survey," Dep. Business, Innov. Ski., 2013.
- [2] S. Gorniak, D. Ikononou, P. Saragiotis, I. Askoxylakis, P. Belimpasakis, B. Bencsath, M. Broda, and C. Vishik, "Priorities for Research on Current and Emerging Network Technologies," Eur. Netw. Inf. Secur. Agency, 2010.
- [3] "a preliminary examination of insider threat programs in the U.S. private sector." [Online]. Available: http://www.insaonline.org/i/d/a/b/InsiderThreat_embed.aspx. [Accessed: 30-Apr-2015].
- [4] Raytheon Company, "Privileged User Abuse & The Insider Threat Commissioned," Ponemon Inst. May, p. 32, 2014.
- [5] D. Cappelli, A. P. Moore, and R. Trzeciak, The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes, 1st ed. Addison-Wesley Professional, 2012.
- [6] M. Bishop, D. Gollmann, J. Hunker, and C. W. Probst, "Countering insider threats," in Dagstuhl Seminar Proceedings 08302, 2008, pp. 1–18.
- [7] M. Swanson, "Security Self-Assessment Guide for Information Technology Systems," NIST Spec. Publ., vol. 800, no. 26, 2001.
- [8] W. Bradley, "AN INSIDER THREAT MODEL FOR ADVERSARY SIMULATION," SRI Int. Res. Mitigating Insid. Threat to Inf. Syst. 2, pp. 1–3, 2000.
- [9] C. Colwill, "Human factors in information security" Inf. Secur. Tech. Rep., vol. 14, no. 4, pp. 186–196, Nov. 2009.
- [10] O. Buckley and J. Nurse, "Reflecting on the Ability of Enterprise Security Policy to Address Accidental Insider Threat," Work. Socio, 2013.
- [11] CERT Insider Threat Team, "Unintentional Insider Threats: Social Engineering," Carnegie Mellon Univ, vol. CMU/SEI-20, no. January, 2014.
- [12] A. Duncan, S. Creese, and M. Goldsmith, "An overview of insider attacks in cloud computing," Concurr. Comput. Pract. Exp., 2014.
- [13] M. Lennon, "Insider Steals Data of 2 Million Vodafone Germany Customers," 2013. [Online]. Available: <http://www.securityweek.com/attacker-steals-data-2-million-vodafone-germany-customers>. [Accessed: 30-Apr-2015].
- [14] F. L. Greitzer, J. Strozer, S. Cohen, J. Bergey, J. Cowley, A. Moore, and D. Mundie, "Unintentional Insider Threat: Contributing Factors, Observables, and Mitigation Strategies," 2014 47th Hawaii Int. Conf. Syst. Sci., pp. 2025–2034, Jan. 2014.
- [15] M. T. Khorshed, a. B. M. S. Ali, and S. a. Wasimi, "A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing," Futur. Gener. Comput. Syst., vol. 28, no. 6, pp. 833–851, Jun. 2012.
- [16] Z. M. Yusop and J. Abawajy, "Analysis of Insiders Attack Mitigation Strategies," Procedia - Soc. Behav. Sci., vol. 129, pp. 581–591, May 2014.
- [17] G. Greenwald, "Edward Snowden: the whistleblower behind the NSA surveillance revelations," The Guardian, 2013. [Online]. Available: [Accessed: 20-Oct-2014]. <http://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance>.
- [18] S. Zeadally, B. Yu, D. H. Jeong, and L. Liang, "Detecting Insider Threats: Solutions and Trends," Inf. Secur. J. A Glob. Perspect., vol. 21, no. 4, pp. 183–192, Jan. 2012.
- [19] K. Scarfone and P. Mell, "Guide to Intrusion Detection and Prevention Systems," Natl. Inst. Stand. Technol, 2007.
- [20] M. Salem, S. Hershkop, and S. Stolfo, "A survey of insider attack detection research," Insid. Attack Cyber Secur., pp. 1–20, 2008.
- [21] C. D. Lori Flynn, Greg Porter, "Cloud Service Provider Methods for Managing Insider Threats: Analysis Phase II , Expanded Analysis and Recommendations," January, 2014.
- [22] G. Silowash, D. Cappelli, and A. Moore, "Common Sense Guide to Mitigating Insider Threats 4th Edition," Dec, 2012.
- [23] T. B. Lewellen, "Insider Threat Control: Using Universal Serial Bus (USB) Device Auditing to Detect Possible Data Exfiltration by Malicious Insiders," no. January, 2013.
- [24] CERT Insider Threat Center., "Insider Threat Control: Using a SIEM signature to detect potential precursors to IT Sabotage," Carnegie Mellon Univ., no. April, 2011.
- [25] J. W. B. George J. Silowash, Todd Lewellen, "Detecting and Preventing Data Exfiltration Through Encrypted Web Sessions via Traffic Inspection," Carnegie Mellon Univ, vol. CMU/SEI-20, no. March, 2013.
- [26] J. Hunker and C. Probst, "Insiders and insider threats—an overview of definitions and mitigation techniques," J. Wirel. Mob. Networks, Ubiquitous , pp. 4–27, 2011.
- [27] J. Crampton and M. Huth, "Towards an access-control framework for countering insider threats," Insid. Threat. Cyber Secur., pp. 1–23, 2010.
- [28] R. S. Sandhu, H. L. Feinstein, and C. E. Youman, "RoleBased Access Control Models," IEEE, 1996.
- [29] L. Spitzner, "Honeypots: catching the insider threat," 19th Annu. Comput. Secur. Appl. Conf., no. Acsac, 2003.
- [30] E. E. Schultz, "A framework for understanding and predicting insider attacks," Comput. Secur., vol. 21, no. 6, pp. 526–531, 2002.
- [31] M. Kandias, A. Mylonas, and N. Virvilis, "An Insider Threat Prediction Model," pp. 26–37, 2010.
- [32] E. T. Axelrad, P. J. Sticha, O. Brdiczka, and J. Shen, "A Bayesian Network Model for Predicting Insider Threats," 2013 IEEE Secur. Priv. Work., pp. 82–89, May 2013.
- [33] F. L. Greitzer and R. E. Hohimer, "Modeling Human Behavior to Anticipate Insider Attacks," J. Strateg. Secur., vol. 4, no. 2, pp. 25–48, 2011.
- [34] F. L. Greitzer, L. J. Kangas, C. F. Noonan, A. C. Dalton, and R. E. Hohimer, "Identifying At-Risk Employees: Modeling Psychosocial Precursors of Potential Insider Threats," 2012 45th Hawaii Int. Conf. Syst. Sci., pp. 2392–240.
- [35] J. R. C. Nurse, P. A. Legg, O. Buckley, I. Agraftiotis, G. Wright, M. Whitty, D. Upton, , and S. Creese, "A critical reflection on the threat from human insiders" Hum. Asp. Inf. Secur. Privacy, Trust, vol. 8533, p. pp 270–281, 2014.
- [36] E. D. Shaw and L. F. Fischer, "Ten Tales of Betrayal: The Threat to Corporate Infrastructures by Information Technology Insiders Analysis and Observations," Sep, 2005.
- [37] K. Roy Sarkar, "Assessing insider threats to information security using technical, behavioural and organisational measures," Inf. Secur. Tech. Rep., vol. 15, no. 3, pp. 112–133, Aug. 2010.
- [38] G. "Gus" Jabbour and D. a. Menasce, "The Insider Threat Security Architecture: A Framework for an Integrated, Inseparable, and Uninterrupted Self-Protection Mechanism" 2009 Int. Conf. Comput. Sci. Eng., pp. 244–251.
- [39] M. E. Palmer, C. Robinson, J. C. Patilla, and E. P. Moser, "Information Security Policy Framework: Best Practices for Security Policy in the E-commerce Age," Inf. Syst. Secur., vol. 10, no. 2, pp. 1–15, May 2001.
- [40] F. Rocha and M. Correia, "Lucy in the sky without diamonds: Stealing confidential data in the cloud," 2011 IEEE/IFIP 41st Int. Conf. Dependable Syst. Networks Work., pp. 129–134.

Energy Efficiency in Smartphones: A Survey on Modern Tools and Techniques

¹Munam Ali Shah, ²Naila Naheed, ³Sijing Zhang

^{1,2}Department of Computer Science, ³Department of Computer Science & Technology

^{1,2}COMSATS Institute of Information Technology, Islamabad, Pakistan

³University of Bedfordshire, Luton, UK

¹mshah@comsats.edu.pk, ²nailanaheed028@yahoo.com, ³sijing.zhang@beds.ac.uk

Abstract— Smartphone is nowadays in the use of the majority of the people. It has different features. It offers power consuming technologies like GPS, 3G, games, apps etc. which creates power management problem faced by the users. In this paper, our contribution is twofold. Firstly, we evaluate different tools and techniques that can be used to optimize the usage of smartphone battery. Secondly, we survey the latest research that has been published in 2010 to 2014 which helps users to find the best technique or tool to achieve energy efficiency in smartphones.

Keywords :Smartphone, Wi-Fi, HBI, GPS, 3G, energy efficiency

I. INTRODUCTION

Human life has become more easy and luxurious with the development of science. It provides many technologies, like robots which are classy, satellites are intellectual and cell phones are now smartphones. These devices work efficiently and give accurate results besides it improve performance. However, common problem of a smartphone is that it runs on limited battery for a longer period of time and the smart features which consumes more battery [1]. Mobile smartphone operates according to their functionality and operating system with a small sized battery. The device remains functional and operable as long as power remains supplied [2]. Modern phones use smart features like monitoring health through sensors, knowing the geo-location of the user and auto adjusting other features. This terms them a smartphone. Smartphone also provides advance technologies like GPS, 3G, Wi-Fi, audio and video playback, web browsing, CPU utilization, media downloads, games and email etc. These are powerful technologies which consume more battery as compared to other features like video streaming. One reason of poor energy management in a smartphone is its size, and it is running different high power-hungry applications on a small device [3].

Another reason of high battery consumption is applications which are regularly used by the user and applications which are opened automatically and consume smartphone battery. It has been noticed that different phones have different power consumption. A large number of users are unaware that certain applications are more power-hungry and consume lots of mobile energy as compared to other applications. This

results in the user assuming that his/her mobile phone is not giving a good battery time and may be the battery needs to be replaced. Many designers of smart phones are working that the applications could automatically adjust themselves with power consumption according to their functionality and performance but unfortunately software developers have limited experience with power usage [4]. Multimedia contents have become important part of a smartphone due to the success of YouTube and other video streaming apps. Basic need for effective power management is that one should know where and how the power is used and which component or app uses more energy. Energy is needed to encode and decode an algorithm, e.g., Wi-Fi consumes three times more power as compared to a normal application and 3G services uses five times more power for encoding and decoding [5]. These services consume more power as compared to other because traffic keeps on flowing during streaming.

In ad hoc network, the most common function is content sharing. Different jobs are needed for content sharing among smartphone like registration, authentication, uploading files to server, downloading files from server and searching files from server. For all these tasks the files should be updated frequently and shared time to time. With smartphone ad hoc network can easily be constructed [6]. Mobile ad hoc networks are dynamic, distributed, and self-governing wireless systems which connect to other wireless devices in the vicinity. The different portable system can interconnected with each other through wireless links or multi hop routing. In wireless ad hoc network power consumption is a serious issue because energy strength is bound to electronic devices [7].

Wi-Fi needs a high bandwidth for working and it also consumes high power of the device. In cellular radios there is delay and overhead in switching from low power of idle state to high power running state [8]. The data sends repeatedly in cellular radio that is the reason for overhead gain.

Users of smartphone face the problem of limited battery lifetime known as *human battery interaction (HBI)*. Human battery interaction is an inverse process. User can save the battery by changing the power saving setting; user can reduce the screen brightness of screen or close process after its use. Power management is important for energy efficiency in smartphones [9].

In this paper, we survey and analyze the techniques that could be used to save mobile energy. The rest of the paper is structured as follows. Section II provides the related work. Section III contains the comparison and analysis of different techniques, protocols, algorithms, devices and applications used for energy efficiency (EE). In section IV, we conclude our survey and discuss our findings.

II. RELATED WORK

In this section, we introduce the background of our work. We discuss how the energy efficiency in smartphones can be improved. The existing work is categorized in different sections mentioned below:

A. Application power consumption

In smartphone, there are a lot of applications which consume power. Boci *et al* [10] survey different architectures and software which improve energy efficiency in smartphones. Energy Efficient Engine [11] is a system which automatically measures the scrolling speed and adjusts speed according to user need.

Most of the users don't know which application consume more energy and how they can save their smartphone battery [12]. Due to this reason, different companies and operating system designers are working on the improvement of the battery life of smartphone. The energy efficient hardware for smartphone is also being designed which consumes low power. Policing is another research area in smartphone and it is being investigated that such policies should be implemented in a smartphone which can help save mobile energy. CIST [13] saves energy by keeping in mind the requirement of application instead of examining the behavior of resources. Han *et al* [14] proposed eDiscovery to improve energy efficiency in smartphone by measuring time duration of Bluetooth. Wang *et al* [12] tried to find out transaction time between the energy consumption and missing probability. They design a probing algorithm called Short Time Arrival Rate (STAR) which finds the arriving rate which falloff slowly and calculates the probing frequency.

The power consumption in 3G cellular phones is influenced by resource control protocol [15]. This protocol is specifically designed and implemented by Radio Network Controller (RNC) for 3G networks to control energy power utilization. The throughput and response time of user equipment depends on its state. The user equipment states are Dedicated Channel (DCH), Forward Access Channel (FACH) and Paging Channel (PCH), stored in ascending order [15]. Now there is worry between the phones having high consumption of batteries with more applications and the phones with high battery life time. Koala [16] is another approach used for energy efficiency. It assign resources according to the requirement of user or according to the performance.

B. Smartphone power model

Smartphones are basically used for communication which uses Wi-Fi, Bluetooth, web browsing and other

applications for communication. For every application Internet or Wi-Fi is required and Wi-Fi consumes more power as compared to other technologies. For this reason, researchers pay a lot of consideration to wireless modules. Different researchers have shown a lot of effort to improve the energy efficiency of smartphones. Earlier energy of mobile phones depended on hardware but now energy in smartphone is dependent on a number of applications which are used by end-users. When users use different applications in a smartphone, each application consumes power which results in the high power consumption of the CPU.

Carrol *et al* [17] investigate that the main source of high energy consumption is the hardware component in the device. Internal and external storages of a smartphone are also components which consume mobile energy. In each of the smartphone's component, current and voltage are used resulting in power consumption. Smartphones users for scientific purpose also adopted similar model and discussed that in [18]. As previously mentioned, 3G is another technology which consumes more power in a smartphone when compared to Bluetooth and infrared. Energy can be saved by the use of backbone network. 3G Radio Resource Control protocol describes three stages for smartphones which are: *IDLE state* in which smartphones do not have any signaling connection with backbone network; *DCH state* in which dedicated channel is assigned to the smartphone by backbone network; and *FACH state* in which there is no dedicated channel for smartphones [19]. Other than above mentioned techniques there are some other sources of energy consumption in a smartphone.

Ananad *et al* [20] present two methods for smartphone power model, i.e., *Potentially Visible Set* (PVS) and *Visual Perception based Localization* (VPL). These models save the energy of a smartphone by evaluating and estimating the non-critical game state. David *et al* [21] present different storage techniques which influence energy consumption in a smartphone. Yazti *et al* [22] describe different techniques for energy management in smartphones. GPS is used frequently nowadays in different applications because it gives us accurate result but it also consumes much power of smartphones. The usage of accelerometer between GPS and motion of device can also help in saving the mobile energy. When motion is sensed, the system automatically switches on the GPS module in a smartphone and when there is no motion it keeps GPS module off [23].

C. Smart battery models and energy cost midels

To use smartphone competently and efficiently, operating system and application requires some computable resources. For this purpose, two models are presented for battery attributes which are: *smart battery model* and *energy cost model* [24]. Different techniques have been purposed to observe energy estimation and measurement. Cycle-accurate simulator shows CPU consumption at low level therefore they can calculate the energy usage but on the other side it is very slow [24].

Some other models, such as the energy consumption of software at instructions level [16], have also been designed, however, the disadvantage of this system is that it only provides energy usage at software and not at instructions level. Also, these type of systems does not take operating system under consideration. System state energy model reflects the energy consumption for hardware components in embedded system [18].

D. Avoiding energy waste

When screen is on smartphone consumes power even user does not use any application. Study has been made on frame rate. It has been investigated that how to improve the frame rate through hardware and software. Energy can be saved if we put the whole system or some components of the system in sleep state for some time. Device named *µsleep* [25] for smartphones which improves the energy efficiency of smartphone. *µsleep* achieves energy efficiency by putting user's smartphone in sleep mode for a short interval of time without disturbing the user's applications. Another technique named *waked on wireless* also improves the energy efficiency of a smartphone. It reduces the smartphone power by turning off the smartphone and some of its interfaces when the user is not using phone and turns it on when there is ongoing traffic [25]. Users never lose the full network connection when the phone is not used. Most of the existing literature about achieving energy efficiency in smartphones discuss applications which consumes more power and how to save power, however, an interesting discussion has been made in [26] which presents the idea of creating some energy efficient interface(s) which can be used for power creation.

E. Communication related energy saving

Many techniques have been purposed to increase energy efficiency at position level of a smartphone. RAPS uses activity of user combining with cell tower to only turn on GPS when it is needed. It also purposes to share readings for position of a Bluetooth to minimize the use of the GPS in order to reduce energy usage. Whilst CAPS uses location history to provide user with GPS services without turning on the GPS, the limitation is that it can only be functional if location based history exists. Cell-ID helps the smartphone for their location based system. By using cell-ID, sequence matching techniques such as CAPS can estimate the current location of user by using cell-ID and GPS history [27].

A lot of research has been carried on minimizing the power consumption of network communication. Reference [28] presents a method called *system level power management* which vigorously turns off the Wi-Fi and radio interfaces to save mobile energy of a smartphone. There is another system, called *coolspots*, which automatically switches between Wi-Fi and Bluetooth according to the need of user to increase the energy efficiency. PSM saves the energy of a smartphone by minimizing the awake time but on the other side, it maximizes the packet delay. Dozy Access Point claims

that when Wi-Fi is put in lower power state or is turned on then power consumption of the mobile access point (MAP) also reduces. Client should not send traffic when MAP is in sleep mode. Dozy AP introduced two new messages: *sleep request* and *sleep response* and a new protocol which performs on both ends clients and AP. Low power "wimpy" is also introduced to improve the energy efficiency of smartphones. Wimpy cluster is more efficient than traditional clusters [28].

F. Computation offloading based energy saving

Computation offloading is a process of migrating the computational task from smartphone to server to save energy in smartphones. Das *et al* [29] proposed a method for content sharing, privacy and energy efficiency management

G. Scheme for optimizing inactivity period of a mobile devices

Yang *et al* [30] propose a new mechanism for Universal Mobile Telecommunication system (UMTS) named as Discontinuous Reception (DRX). Two parameters, i.e., threshold and DRX cycle are controlled by DRX. These parameters improve the energy efficiency in UMTS. They also develop an application dependent protocol called Power Saving Mode (PSM) which redirects the TCP traffic into periodic burst with the same throughput as sever transmission rate. Agrawal extends the current study proposed in [30] and develop a new algorithm called *Opportunistic Power Save Mode* (OPSM). This algorithm works efficiently when all connected clients are busy in web browsing which is considered by download over TCP.

H. Prototypes used to reduce energy overhead

Different prototypes are discussed which reduce the energy overhead. These prototypes are *Haromni*, *Cenceme*, *Medially* and *Little Rock* [31]. Jigsaw proposed a pipeline stream processing architecture that generates different sensors at altered rates to encounter the accuracy needed by different applications. According to the IEEE 802.11 standard, PSM saves the power of a smartphone by turning off Wi-Fi of smartphone when not in use because of Beacon Interval (BI). BI frames contain the identifiers of PSM clients for which there are buffered packets in AP. This is called Traffic Indication Map (TIM). If PSM client does not address TIM then it will go back to sleep mode and will remain in awake state and consume the power of device [32].

I. Protocols used for Energy Efficiency

Dutta and Culler *et al* [33] proposed an asynchronous neighbor discovery protocol for mobile sensing named *Disco*. U-connect is another asynchronous discovery protocol for mobile sensing that selects the time slot and improves the performance. Bakht [34] proposed another protocol called *searchlight* which combines both probabilistic and deterministic approaches to reduce the latency in a smartphone. Another synchronous architecture known *FlashlinQ* [35] is proposed by Qualcomm for direct communication. FlashlinQ is more

Table 2. Comparison and Analysis of Different Protocols for Energy Efficiency in Smartphones

Protocols	Energy Efficiency	Cost	Speed	Operating System	Environment friendly	Deployment	Estimated % of energy efficiency
Disco	High	High	Low	Android, symbian	yes	Nokia N900, HTC	44%
U-Connect	High	High	Low	Mobile OS	yes	FireFly Badge	38%
Searchlight	High	High	Low	Symbian	No	N900	45%

effective when compared with Bluetooth and Wi-Fi. Salondis identifies the delay of Bluetooth and introduces a random symmetric protocol to minimize the delay. Drula tries to find as how to select Bluetooth device to reduce power consumption according to the mobility.

J. Energy localization techniques

Abdesslem used accelerometer to save power of smartphone. It finds user's state of mode. If user is in stationary mode then GPS is not used during this period of time. Sendra *et al* [36] linked S-MAC and T-MAC which introduces a duty cycle to improve the energy efficiency. It is a significant reason of consumption of energy in sensor network where nodes are not continually communicated. They find that T-MAC is better than S-MAC. Their study also discusses the problem of identifying the user which faces power management in term of costs. For performance evaluation of power, cost is very important parameter. It is described that different techniques are used for power and energy management. The power consumption mechanism is split into two parts based on their prime objectives, i.e., *passive PCM* and *active PCM*. Passive PCM is further divided into: *i)* physical layer PCM; *ii)* fine grain PCM; and *iii)* coarse-Grain PCM. Distributed and backbone approaches are used in implementation of Coarse-Grain PCMs. Active PCMs are based on the layer of data structure, Network layer and Transport layer. Different algorithms are considered for each type. The objectives, mechanism, performance and application are studied in these techniques.

Localization can be divided into three unique modes, i.e., *network based*; *client based*; and *hybrid mode*. Each mode has its own pros and cons that are applied in different technologies like positioning for cellular network and GPS. If we talk about location based services, it means it is power hungry because we need

regular updates for tracking communication which results in load on server nodes. Accelerometer is used to reduce the power consumption. Reason for low use of power consumption at GPS and Wi-Fi is due to that they switch on and off to scan for access points and base stations.

K. Energy consumption of background traffic

Smartphone creates background traffic even if the user is not using the device. The volume of these backgrounds is not much higher but this type of traffic consumes energy. User should ensure the termination of the processes properly after using the application. This will improve power consumption of a smartphone to some extent. The authors in [37] studied the IEEE 802.11 standard for energy efficiency using PSM. This study focuses on MAC sub-layer of data link layer. At the network layer, energy efficiency of routing protocol is studied and lastly, the study also examines the energy efficiency of TCP at transport layer. The study in [29] also discusses different components of a smartphone which consume more power during the streaming. The researchers focus on link layer solutions while working on network interface. Systems works on behavior of users for Android phone. System uses a passive method for a large scale of deployment which does not need the active participation of users.

Different researches has been carried out on power modeling techniques for smartphone and for computers. Modeling techniques requires basic knowledge of relationship between functional unit of processor and results of power consumption. On the other hand, some researcher proposed a new microprocessor power model named *Black box* which does not require knowledge about the hardware components for the implementation. These models are based on the assumption of linear relationship between CPU energy consumption and some of hardware performance.

Table 1. Energy Efficient Operating Systems [34]

Names	Description
Ecosystem	It is Hybrid operating system for smart deices which depend on energy efficient scheduler.
Odyssey	It is Hybrid Linux operating system which embraces applications like QoS according to energy requirement.
Cinder	It is Hybrid operating system for smart deices which is assembled on topmost of HiStar operating system.
ErdOS	It is extension of Android OS, centralized and energy efficient for smart devices.
CondOS	It is context efficient operating system which manages resources efficiently.

III. PERFORMANCE COMPARISON AND ANALYSIS

In this section, we compare and analyze different techniques and tools used for energy efficiency in smartphones. We provide our analysis and findings in form of different tables. Table 1 compares energy efficient operating systems. Table 2 provides comparison of different protocols used for energy efficiency in smartphones. Table 3 discusses different techniques used by the user which improves battery timing. Table 4 names energy efficient applications. Algorithms that can be used to save mobile energy are discussed in Table 5. Lastly, there are different devices available that can be used to save mobile energy and are reviewed in Table 6.

Table 3. Comparison and Analysis of Different Techniques for Energy Efficiency in Smartphones

Techniques	Energy Efficiency	Cost	Speed	Operating System	Environment friendly	Deployment	Estimated % of energy efficiency
Koala	High	Low	Low	Ubuntu	yes	Windows	26%
CIST	High	High	High	Linux	yes	Windows mobile	40%
Cycle-accurate	High	High	Low	Linux	Yes	Embedded System	5%
Waked on wireless	High	Low	High	Android, sambian	Yes	Nokia, Samsung	Upto 115%
Accelerometer	High	Low	High	Android, iOS	Yes	Smartphones	53%

Table 4. Comparison and Analysis of Different Applications for Energy Efficiency in Smartphones

Applications	Energy Efficiency	Cost	Speed	Operating System	Environment friendly	Deployment	Estimated % of energy efficiency
Powerscope	High	Low	High	Linux	yes	odyssey	46%
Coolspot	High	High	Low	Linux	yes	Embedded system	50%
Clean master	High	Low	Low	Andriod	yes	HTC,Sony, Samsung	36%

Table 5. Comparison and Analysis of Different Algorithm for Energy Efficiency in Smartphones

Algorithms	Energy Efficiency	Cost	Speed	Operating System	Environ ment friendly	Deployment	Estimated % of energy efficiency
PSM	High	High	Low	Andriod, iOS	yes	NS-2	90%
OPSM	High	High	Low	Linux, Solaris, MAC OS	Yes	Ns-2.33	92%
STAR	High	High	Low	Linux, Solaris, MAC OS	Yes	NS-2	39%

IV. CONCLUSION

Smartphones are widely being used by most of us because of its salient features such as GPS, 3G, Wi-Fi and games etc. These features are power hungry because a smartphone is running on small battery. Energy efficiency in smartphones is a major concern. In this paper, we reviewed different papers and made comparison of different applications, techniques, protocols, devices and algorithms which are used for achieving energy efficiency in a smartphone. It is believed that energy efficiency in smartphones could only be achieved on cost of something. A single solution/techniques cannot be considered a the best choice because there is always a tradeoff between energy efficiency and different parameters such as speed, performance, operability and availability.

REFERENCES

- [1] S. S. Oyewobi and E. N. Onwuka, "Mobile terminals energy: a survey of battery technologies and energy management techniques," *International Journal of Engineering and Technology*, vol. 3, no. 3, pp. 282–286, 2013.
- [2] S. Harizopoulos and S. Papadimitriou, "A case for micro cellstores: energy-efficient data management on recycled smartphones" *proc 7th Int. workshop on Data Management on New Hardware, New York*, pg 50-55-2011.
- [3] R. Bala, "Battery power saving profile with learning engine in android phones," *International Journal of Computer Applications* vol. 69, no. 13, pp. 38–41, 2013.
- [4] B. Aggarwal, N. Spring, and A. Schulman, "stratus \square : energy-efficient mobile communication using cloud support," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4, pp. 477-478, 2011.
- [5] M. A. Hoque, M. Siekkinen, and J. K. Nurminen, "Energy efficient multimedia streaming to mobile devices: a survey," in *journal Communication Surveys and Tutorials, IEEE*, vol-16, pp. 579-597, 2014.
- [6] S. Nirjon, R. F. Dickerson, P. Asare, Q. Li, D. Hong, A. John, P.

Table 6. Comparison and Analysis of Different Devices for Energy Efficiency in Smartphones

Devices	Energy Efficiency	Cost	Speed	Operating System	Environment friendly	Deployment	Estimated % of energy efficiency
μ sleep	High	Low	High	Android, Symbian	yes	Pocket PC	60%
DRX	High	High	Low	Linux, Solaris	Yes	LTE	40%

- Hu, G. Shen, and X. Jiang, "Auditeur: A mobile-cloud service platform for acoustic event detection on smartphones", *11th international conf on mobile systems, applications, and services*, ACM, pg 403-416, june-2013.
- [7] H. Liu, F. Xia, Z. Yang, and Y. Cao, "An energy-efficient localization strategy for smartphones" *Journal on Computer Science and Information Systems*, vol.8, No.4, pg 1117-1128, 2011.
- [8] K. H. Kim, A. W. Min, D. Gupta, P. Mohapatra, and J. P. Singh, "Improving energy efficiency of Wi-Fi sensing on smartphones," *2011 Proc. IEEE INFOCOM*, pp. 2930-2938, Apr. 2011.
- [9] D. Li and W. G. J. Halfond, "An investigation into energy-saving programming practices for Android smartphone app development," *Proc. 3rd Int. Work. Green Sustain. Softw. - GREENS 2014*, pp. 46-53, 2014
- [10] S. Maloney and I. Boci, "Survey: techniques for efficient energy consumption in mobile architecture" in *Power MW* "vol. 16. no. 9, pp. 7-35, 2012.
- [11] H. Han, J. Yu, H. Zhu, and Y. Chen, "E3: energy-efficient engine for frame rate adaptation on smartphones," in *Proceedings of 11th ACM on Embedded Networked Sensor Systems*, pp. 15, nov-2013
- [12] S. Zöller, A. Reinhardt, M. Wachtel, R. Steinmetz and Z. Sebastian, "Integrating wireless sensor nodes and smartphones for energy-efficient data exchange in smart environments" *Conference on PERCOM Workshops, 2013 IEEE*, pg 652-657, March-2013.
- [13] R. Palit, K. Naik, and A. Singh, "Anatomy of WiFi Access Traffic of Smartphones and Energy Saving Techniques" *International Journal of Energy, Information and Communications* vol. 3, Issue 1, February, 2012.
- [14] B. Han, J. Li, and A. S. Fellow, "On the Energy Efficiency of Device Discovery in Mobile Opportunistic Networks: A Systematic Approach" *Journal on Mobile Computing, IEEE*, vol.14, pg 786-799, July 2014.
- [15] E. J. Vergara, J. Sanjuan, and S. Nadjm-Tehrani, "Kernel level energy-efficient 3G background traffic shaper for android smartphones," *2013 9th Int. Wirel. Commun. Mob. Comput. Conf.*, pp. 443-449, Jul. 2013
- [16] N. S. Deshmukh, "Energy Efficient Content Sharing in Smart Phones using Wi-Fi Networks," *International Journal of Emerging Science and Engineering (IJESE)*, ISSN: 2319-6378, Volume-2 Issue-9, July 2014.
- [17] A. Carroll and H. Gernot, "An Analysis of Power Consumption in a Smartphone" *Conference on USENIX annual technical*, vol. 14, June-2010
- [18] L. Zhang, R. P. Dick, Z. M. Mao, Z. Wang, and A. Arbor, "Accurate online power estimation and automatic battery behavior based power model generation for smartphones." *In 8th Int. Conf. on Hardware/software codesign and system synthesis, ACM*, pg. 105-114, Oct-201
- [19] N. Lathia, V. Pejovic, K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow, "Smartphones for large-scale behaviour change interventions," vol. 12, no. 3, pp: 66-73, IEEE Computer Society
- [20] A. Bhojan, Z. Qiang, and A. L. Akkihebbal, "Energy efficient multi-player smartphone gaming using 3D spatial subdivision and pvs techniques," *Proc. 3rd ACM Int. Work. Interact. Multimed. Mob. portable devices - IMMPD '13*, pp. 37-42, 2013.
- [21] D. T. Nguyen, "Evaluating impact of storage on smartphone energy efficiency," *Proc. 2013 ACM Conf. Pervasive ubiquitous Comput. Adjunct. Publ. - UbiComp '13 Adjunct.*, pp. 319-324, 2013
- [22] D. Zeinalipour-yazti, "Energy Efficient Data Management in Smartphone Networks," *Department of Computer Science, University of Cyprus* pp. 8-9.
- [23] T. Graf, "Power-efficient positioning technologies for mobile devices," SNET2 Seminar, Dec- 2011
- [24] W. Jung, K. Kim, and H. Cha, "UserScope: A fine-grained framework for collecting energy-related smartphone user contexts," *2013 Int. Conf. Parallel Distrib. Syst.*, pp. 158-165, Dec. 2013.
- [25] M. Waseem, "Energy efficient mobile operating systems," vol. 1817, pp. 1812-1817, 2013.
- [26] A. Misra and L. Lim, "Optimizing sensor data acquisition for energy-efficient smartphone-based continuous event processing," *IEEE 12th Int. Conf. Mob. Data Manag.*, pp. 88-97, Jun. 2011..
- [27] K. Kim and J. P. Singh, "Energy-Efficient Positioning for Smartphones using Cell-ID Sequence Matching Categories and Subject Descriptors." *Conference on Proceedings of the 9th international conference on Mobile systems, applications, and services*, ACM, pg 293-306, june-2011.
- [28] H. Han, Y. Liu and G. Shen, "DozyAP: Power-Efficient Wi-Fi Tethering" *MobiSys '12, June 25-29, 2012, Low Wood Bay, Lake District, UK*. ACM 978-1-4503-1301-8/12/06.
- [29] P. K. Das, A. Joshi, and T. Finin, "Energy Efficient Sensing for Managing Context and Privacy on Smartphones," *The First Conf. Society, Privacy and the Semantic Web (2013)*, vol. 11, no. 21, Oct-2013.
- [30] B. Zhao, Q. Zheng, G. Cao, and S. Addepalli, "Energy-Aware Web Browsing in 3G Based Smartphones," *2013 IEEE 33rd Int. Conf. Distrib. Comput. Syst.*, pp. 165-175, Jul. 2013.
- [31] D. P. Lymberopoulos and J. Liu, "Little Rock: Enabling Energy Efficient Continuous Sensing on Mobile Phones." *Journal on Pervasive Computing*, IEEE, vol.10, no: 2, pg 12-15, april-june 2011.
- [32] Z. Yan, V. Subbaraju, D. Chakraborty, A. Misra, and K. Aberer, "Energy-Efficient Continuous Activity Recognition on Mobile Phones: An Activity-Adaptive Approach," *2012 16th Int. Symp. Wearable Comput.*, pp. 17-24, Jun. 2012.
- [33] N. Tantubay, D. R. Gautam, and M. K. Dhariwal, "A Review of Power Conservation in Wireless Mobile Adhoc Network (MANET)," vol. 8, no. 4, pp. 378-383, 2011.
- [34] N. Vallina-rodriguez and J. Crowcroft, "Energy Management Techniques in Modern Mobile Handsets," *journal on IEEE Communications Surveys & Tutorials*, IEEE, pp. 1-20, no:99, Feb- 2012.
- [35] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "FlashLinQ: A synchronous distributed scheduler for peer-to-peer ad hoc networks," in *Proc. 48th Annu. Allerton Conf. Commun. Control Comput.*, Sep./Oct. 2010, pp. 514-521.
- [36] S. Sendra, J. Lloret, M. Garcia, and J. F. Toledo, "Power Saving and Energy Optimization Techniques for Wireless Sensor Networks (Invited Paper)," *J. Commun.*, vol. 6, no. 6, pp. 439-459, Sep. 2011.
- [37] K. H. Jung, Y. Qi, C. Yu, and Y.-J. Suh, "Energy efficient Wifi tethering on a smartphone," *IEEE INFOCOM 2014 - IEEE Conf. Comput. Commun.*, pp. 1357-1365, Apr. 2014.

Dynamic Analysis of An Underwater Leveling-Gripping System of An Jacket Platform Under Offshore Environmental Loads

Peiran Jiang¹, Liquan Wang¹, Xichun Luo²

¹College of Mechanical and Electrical Engineering, Harbin Engineering University, Harbin, China

²DMEM, University of Strathclyde, Glasgow, UK
jiangpeiran@hrbeu.edu.cn

Abstract—This paper concerns dynamic analysis of an underwater leveling-gripping system which is mounted on a jacket under the influence of offshore environmental loads. Based on the Shinozuka theory, the wave load is calculated in the time domain while the ocean current and wind load on the jacket structure are calculated as constant loads. The main environmental loads and its combination which jacket withstand in leveling process are therefore defined. Using SACS software, according to the South China Sea conditions, a platform bottom dynamic response is calculated under extreme environmental loads in different return period. ADAMS software is also used to dynamically analyze the contact force of key clamping contact parts of leveling-gripping system in leveling process. With the result of analysis, the influence of environmental loads on leveling-gripping system, changes with time, can be obtained accurately, which is an important basis for the design of key parts of the leveling-gripping system.

Keywords- dynamic analysis; leveling-gripping system; offshore environmental load; time domain ; contact

I. INTRODUCTION

Jacket platform is widely used in the world's offshore oil fields. The installation technology of jacket platform is one of the key technologies used for offshore oil and gas production^[1]. Currently, a large number of jackets are installed in East China Sea and South China Sea, water depth ranging from a few meters to a few hundred meters. When installing a platform, in order to achieve the required leveling precision, the jacket must be adjusted by leveling-gripping system after it falls into the seabed. During the leveling process, jacket will withstand environment loads include wind, wave, ocean current and so on^[2], so leveling-gripping system of jacket have obvious dynamic response^[3]. In order to ensure the stability for operating the system, dynamic analysis on leveling-gripping system of Jacket becomes an important task. The paper will carry out dynamic analysis of the gripping system during the leveling process propose study its response in the time domain under the environmental loading.

II. THEORETICAL CALCULATION OF ENVIRONMENT LOADS

A. Wave load

The effect of wave load on offshore structure is studied based on the Morrison formula and the Stokes five order wave theory^[4,5].

The components of jacket legs are space tilt rods. Morrison formula is adjusted to calculate the wave force acted on the offshore inclined structures bar. It can be described as:

$$\begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} = \frac{1}{2} \rho C_D D |\omega_n| \cdot \begin{bmatrix} u_{nx} \\ u_{ny} \\ u_{nz} \end{bmatrix} + \rho C_M A \begin{bmatrix} \dot{u}_{nx} \\ \dot{u}_{ny} \\ \dot{u}_{nz} \end{bmatrix} \quad (1)$$

Where F is the wave force acted on pile, while u , \dot{u} , ρ , D , C_D , C_M are velocity of water particle, acceleration of water particle, density of water, cross section diameter of the pile, drag coefficient $C_D = 0.6 \sim 1.2$ and inertial force coefficient, $C_M = 1.3 \sim 2.0$.

The wave generated by the wind is a kind of highly irregular phenomenon and will not be repeated, so the wave is actually a kind of random wave^[6,7]. At present, the analysis of offshore structures is mainly based on spectrum analysis. In order to study the dynamic response of the key parts of the leveling-gripping system in the leveling process, the dynamic analysis of jacket structure under wave load must be carried out in the time domain.

The harmonic wave superposition method is used to simulate the random wave^[8]. According to the Shinozuka theory, the surface of wave $\eta(t)$ can be simulated as follows:

$$\eta(t) = \sum_{j=1}^N \sqrt{2\mathcal{Q}(\omega)} \Delta\omega \cos(\omega t + \phi_j) \quad (2)$$

Where N is sufficiently large integer, while $\mathcal{Q}(\omega)$, $\Delta\omega$, ϕ_j are Random wave spectrum, frequency increment and random variable distributed in $(0, 2\pi)$, respectively.

The single parameter I.T.T.C P-M spectrum is selected to describe the wave:

$$S(\omega) = 0.78\omega^{-5} \exp\left(\frac{-3.11}{H_s^2\omega^4}\right) \quad (3)$$

Where the effective height of wave H_s equals to 8.5m. The wave spectrum curve is shown in Figure 1. In the figure 1, density of wave spectrum mainly distribute in (0, 1.5), and reached the maximum when $\omega=0.48$. 500 seconds of wave height curve is illustrated in Figure 2, simulate the height of random wave changes with time, height of wave reaches the minimum and maximum values 4-5 times respectively, accords with the change process of height of real wave.

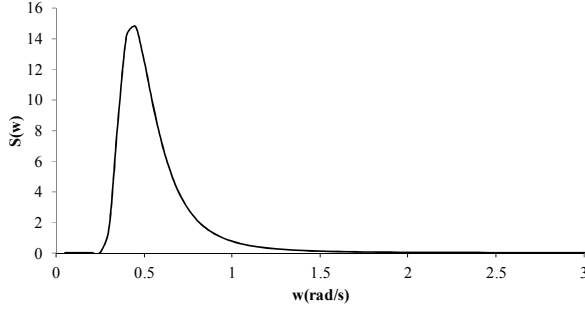


Figure 1 The I.T.T.C wave spectrum curve

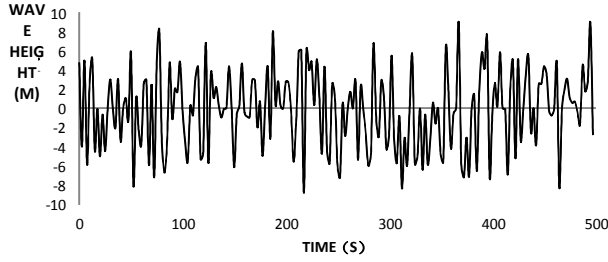


Figure 2 The curve of wave height

B. Current load.

When the current and wave are in the same direction, the current will result in the increase of the total load. On the contrary, the current will weaken the total load when it is in the opposite direction of the movement of wave. Therefore, the current has obvious effects on the wave load.

The current force can be expressed as:

$$F_c = \frac{1}{2} C_D \rho D v_c^2 \quad (4)$$

Where v_c is velocity of current, m/s.

The wave velocity vector is set to be v . The drag force vector, i.e. the joint action of wave and current can be calculated by Morrison formula, shown as follow:

$$F_D = \frac{1}{2} C_D \rho D (v + v_c) |v + v_c| \quad (5)$$

C. Wind load.

Wind load is usually can be regarded as a constant force. By using the calculation method of wind in API RP 2A, the wind load can be calculated as:

$$F = (\omega / 2g) V^2 C_s A \quad (6)$$

Where F is the force of wind, while ω , g , V , C_s and A are density of air, acceleration of gravity, velocity of wind, shape coefficient and windward area of structure.

III. THE BASE SHEAR OF JACKET UNDER OFFSHORE ENVIRONMENTAL LOADS

The jacket leveling-gripping system is installed to connect between jacket legs and steel pipe pile. This connected will withstand significant shear force caused by the environmental loads in the horizontal direction. Therefore, the base shear of the jacket under environmental loads is analyzed in this study.

A. Analysis model of jacket

A jacket platform located in the South China Sea has been selected as the object to analyze. This platform is a four legged jacket platform, fixed in the seabed by steel pipe piles and pile skirt structure. The depth of water is about 354 feet. The analysis model is shown in Figure 3, the jacket platform is modeled and presented by using SACS (Seastate Analysis Computer System software, Engineering Dynamics Company in USA).

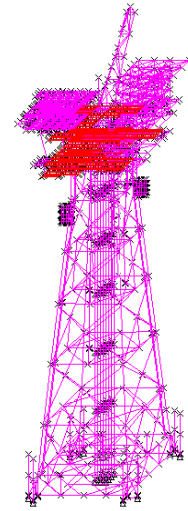


Figure 3 A jacket platform SACS analysis model

B. Determination of environmental loads factor

The working time of leveling-gripping system is generally a few days, and the offshore environment loads always include wind, wave and current. In order to ensure the performance of leveling-gripping system, the environment loads of wave, current and wind are analyzed in different return period, i.e. for one year, ten years and one hundred years of the South China Sea^[9], Respectively. The Environmental loads factor data are shown in table 1.

TABLE I. ENVIRONMENTAL LOAD FACTOR DATA TABLE

Environmental loads factor	unit	Return period (years)		
		1	10	100
1 minute average wind speed	m/s	33.2	38.3	51.7
3 second wind	m/s	37.8	43.6	58.7
effective wave height	m	8.5	10.7	13.9
effective period	s	11.9	13.7	15
Zero mean period	s	9.6	10.8	12.3
maximum wave height	m	14.2	17.9	23.3
period of wave spectrum peak	s	12.7	14.6	16
surface current velocity	m/s	1.09	1.61	2.19
middle current velocity	m/s	0.7	1.13	1.82
Bottom current velocity	m/s	0.33	0.49	0.93

C. The maximum base shear force of jacket

To consider the most dangerous situation, jacket is analyzed and calculated with all environmental loads acting in the same direction. The incident angle of environmental loads action to jacket is analyzed in eight directions, and the relationship between the maximum base shear and incident angle under different return period is shown in Figure 4.

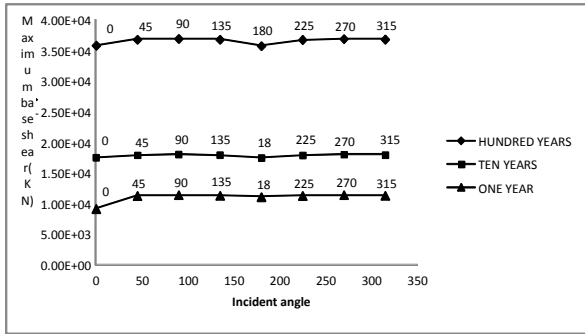


Figure 4 The relationship between base shear and incident angle

IV. DYNAMIC RESPONSE OF LEVELING-GRIPPING SYSTEM KEY PARTS

In the gripping system, the clamping contact parts between steel pipe pile and clamping claw are the key parts. The contact force directly determines the performance of the leveling-gripping system leveling and its stability. The contact forces are in the horizontal plane, same with the offshore environment loads include the waves, currents and wind, resulting in contact forces are directly influenced by the environment loads, and have obvious dynamic response.

The whole process of the system leveling under environmental loads is analyzed by using ADAMS software, include analysis of dynamics and kinematics, to get the dynamic response of the clamping contact parts. The system dynamic model is shown in Figure 5.

Figure 6 shows the different contact position of upper sleeve clamping claws and direction of environment load.

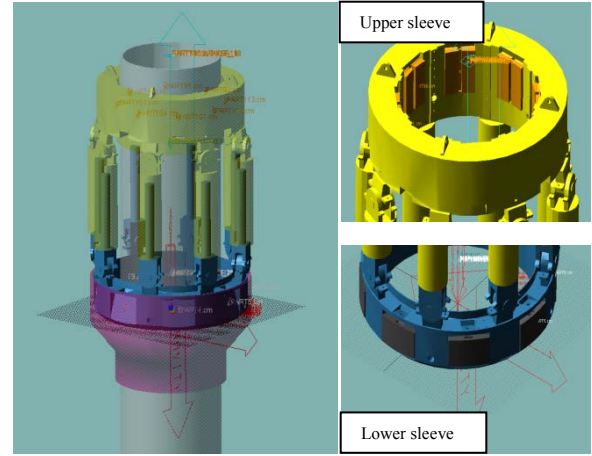


Figure 5 Leveling-gripping system dynamic model

Under one year return period of sea conditions, the upper sleeve clamping position is analyzed. At leveling, the time curve of the contact forces between the gripper and steel pipe pile in 1#, 3#, 7# position are shown in Figure 7.

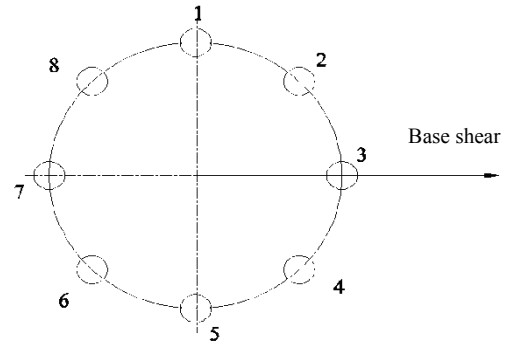


Figure 6 The contact position and loading direction

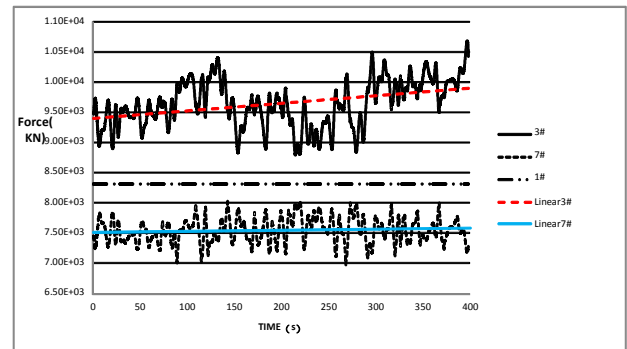


Figure 7 The time curve of each contact position force in upper sleeve

Because the direction of pressure and environment load is vertical, the contact force at 1# position is not affected by the environment load. The contact force at 3# position increased significantly because of the direction of the clamping force is same with the environmental loads, while the 7# pressure decreased significantly because of the direction of the clamping force is contrary. To view the overall leveling period, with the hydraulic cylinder extending, the contact forces at 3# and 7# positions are increasing, because the arm of force growth.

Under the ten years return period and one hundred years extreme sea conditions, time curves of contact force at 3# position in the upper sleeve are shown in Figure 8 and Figure 9.

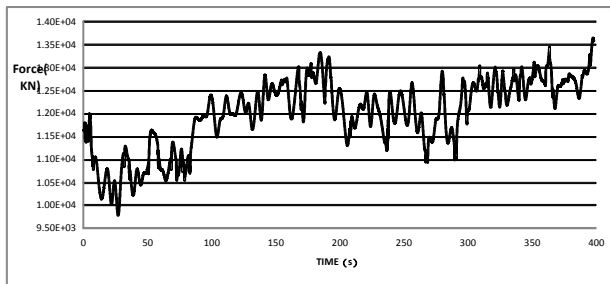


Figure 8 Upper clamping claw 3# position force time curve(10years)

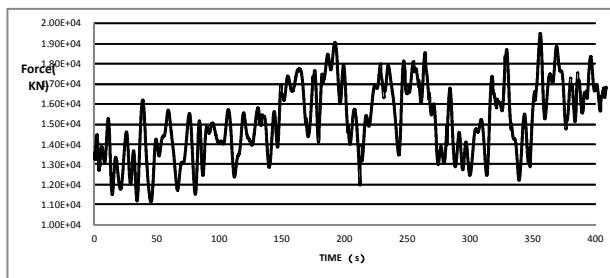


Figure 9 Upper clamping claw 3# position force time curve(100years)

The curves show that, the contact force at 3# position is between 9740KN and 13600KN under the ten years return period extreme sea conditions, and the contact force is between 11000KN and 19500KN under the one hundred years return period extreme sea conditions. If the contact area isn't sufficient, it will exceed the yield limit of pile and crush pile.

CONCLUSIONS

This paper presented dynamic analysis on the leveling-gripping system of a jacket platform under the influence of wave, current and wind. The analysis models of base shear of the jacket and contact force between piles and clamps of leveling-gripping system have been built. Through the calculation, the contact forces of the key parts which can be increased or decreased obviously by extreme environmental loads have been obtained. This dynamic analysis can be used to find the relationship between offshore environmental loads and the dynamic responses of leveling-gripping system, the data is important for the design of key contact part of leveling-gripping system.

REFERENCES

[1] Qian Yalin, Various World Offshore Oil&Gas Production Systems, Offshore Oil, Vol.25, No.2, pp.86-92, 2005.

[2] Cossa Nelson J., Potty Narayanan S., Idrus Arazi B., Hamid Mohd Foad Abdul, Nizamani Zafarullah. Reliability analysis of jacket platforms in Malaysia-environmental load factors, Research Journal of Applied Sciences, Engineering and Technology, Vol.4, No.19, pp.3544-3551,2012.

[3] Dousti M., Gharabaghi A.R.M., Chenaghlou M.R. Study of nonlinear dynamic behavior of a typical jacket type platform under environmental loading (wave & current) and blast overpressure, Offshore Mechanics and Arctic Engineering - OMAE, Vol.1,pp. 901-909P, 2004.

[4] Morison J.R., O' Brien M.P., Johnson J.W, et al. The forces exerted by surface waves on piles, Petro. Trans., Am.Inst. of Mining Eng., Vol.189, pp.149-154, 1950.

[5] L.Skjelbreia and J.A. Hendricksen, Fifth-order Gravity Wave Theory, Proceedings of Seventh Conference on Coastal Engineering, pp.184-196,1961.

[6] Hou Yijun, Guo Peifang, Song Guiting, Song Jinbao, Yin Baoshu, Zhao Xixi, Statistical distribution of nonlinear random wave height, Science in China, Series D: Earth Sciences, Vol.49, No.4, pp.443-448, 2006.

[7] Su Houde, Li Jinbo, Ren Yongquan, Fan Jianling, Dynamic Response Analysis of Jacket Platform Under Complex Loads, PETRO-CHEMICAL EQUIPMENT, Vol.40, No.6, pp.35-37, 2011.

[8] Yan Shi, Zheng Wei, Wind Load Simulation by Superposition of Harmonic, Journal of Shenyang Arch. And Civ. Eng. Univ., Vol.21, No.1, pp.1-2, 2005.

[9] Qi Yiquan, Zhang Zhixu, Shi Ping, Extreme wind, wave and current in deep water of south china sea, International Journal of Offshore and Polar Engineering, Vol.20, No.1, pp.18-23, 2010.

Design of compliant parallel grippers using the position space concept for manipulating sub-millimeter objects

Guangbo Hao*, Ronan Hand

School of Engineering
University College Cork
Cork, Ireland

*Corresponding author: G.Hao@ucc.ie

Xianwen Kong

School of Engineering and
Physical Sciences
Heriot-Watt University
Edinburgh, EH14 4AS, UK

Wenlong Chang[†], Xichun Luo

Department of Design, Manufacture
and Engineering Management
University of Strathclyde
Glasgow, G1 1XJ, UK

[†]Presenter

Abstract—*The structure or configuration of compliant mechanisms can be reconfigured through changing the positions of each compliant module thereof within their position spaces. A number of 1-DOF 2-PRRP compliant parallel grippers (CPGs) can be obtained using the structure re-configurability for manipulating sub-millimeter objects. Even with the geometrical parameters for the system's pseudo-rigid-body model (PRBM) and each compliant module kept at the same values, the position of each compliant joint can be anywhere within its position space. The performance of the resulting CPG varies with the position of the compliant joint. In this paper two typical CPG designs are presented and analyzed. Comparisons between FEA simulation results and analytical models show that the input-output kinematic relationship of the non-compact design agrees better with that of the PRBM due to its better load transmissibility. One can design different structures based on specific design requirements.*

Keywords—*gripper; compliant mechanisms; sub-millimeter objects; position space; conceptual design*

I. INTRODUCTION

Manufacturing is an essential part of the EU economy which contributes major portions of Gross Domestic Products (GDP), exports and economic resilience. Increasing competitiveness of high value manufacturing sector through high-level process control or monitoring [1] has been set up as an important agenda in the EU Horizon 2020 programme. Using robotic gripper to automatically handle sub-millimeter objects (such as 1mm dimension of micro-lens) for high-precision manufacturing has become an important research topic for competitive manufacturing [2].

Traditional rigid-body robotic grippers often suffer from poor resolution and repeatability due to the joint's backlash and friction, and are therefore not suitable for manipulating sub-millimeter objects. However, compliant mechanism based designs, transferring force or displacement through the elastic deformation of one or more flexible members within the structure (i.e., jointless), can overcome the above problems. Due to their advantages of reducing the number of parts (thereby

raising the system reliability), reducing the assembly and fabrication cost, and increasing the system performance, such grippers have been successfully used in the applications of precision engineering, biomedical devices and MEMS [3-11].

This paper focuses on the design of compliant parallel grippers (CPGs) for manipulating sub-millimeter objects. A CPG is generally a 1-DOF compliant parallel mechanism composed of a base, compliant members, and two or more jaws. The jaws of a CPG are often indirectly driven by a linear actuator to grasp an object within the jaws. CPGs, by fine control, can achieve micro/nano-manipulation precision specified in terms of motion repeatability, accuracy (lack of error) and actual resolution (actual minimum incremental motion). There are generally two manners of gripping for the jaws: angular and parallel [5]. Angular one may lead to a sliding motion between the gripped object and the jaw, while this is maximally avoided with the parallel gripping arrangement. The parallel gripping can also provide an even distribution of the gripping force over the manipulated sample and minimize the stress distribution on the grasped object [6].

Most of emerging CPGs are based on the well-known kinematics-based substitution methods using traditional slider-crank mechanisms, parallelogram mechanisms, and/or straight-line four-bar linkages [6, 8-10]. Topology optimization based methods have also been employed to design CPGs [11]. Displacement amplification mechanisms are usually involved in these designs. However, how to improve CPGs with regard to compactness, simplicity, and/or motion range is still an open issue.

This paper aims to design new CPGs with distributed compliance using the structure re-configurability through the concept of position spaces for compliant joints. These new designs can provide a variety of selections for specific requirements such as compactness and simplicity. This paper is organized as followed. Detailed design is described in Section II. Section III conducts the analytical kinetostatic modeling for the CPG based on the pseudo-rigid-body model (PRBM) followed by case studies for two typical designs in Section IV. Further considerations

are discussed in Section V and conclusions are drawn in Section VI.

II. POSITION-SPACE-BASED DESIGN

As shown in Fig. 1, a CPG is proposed based on a 1-DOF 2-PRRP mechanism [12]. Here and throughout this paper, P and R stand for a prismatic joint and a revolute joint, respectively. The CPG is obtained by replacing each joint in the 2-PRRP mechanism with the compliant counterpart. Here, the compliant P joint is a basic parallelogram mechanism and the compliant R joint is an isosceles trapezoidal flexure mechanism with its remote rotation center coinciding with the center of the corresponding R joint in the PRBM (Fig. 1(d)). This 2-PRRP mechanism itself is a displacement amplification mechanism for amplifying the input displacement. Therefore, a separate displacement amplification mechanism is not needed.

The structure of compliant mechanisms is reconfigurable through changing the positions of each compositional compliant module thereof within the

position space [13]. The position space of a compliant module is the combination of all permitted positions in a system where the constraint of this compliant module in the system remains unchanged when the position of the compliant module changes relative to its adjacent compliant module rather than being considered in isolation. The position space can be identified using the screw theory based method as reported in [13], which is not only useful for gripper design but also for general compliant mechanism design [13-15].

The CPG can therefore have a variety of structures/configurations through changing the positions of each compliant P/R joint (Fig. 1(a)). Here, the P joint's position can be translated and/or rotated, and the R joint's position can be rotated. Note that even through the geometrical parameters (Fig. 1(d)) of the system's PRBM and each compliant joint/module are kept at the same values, the position of each compliant joint can be anywhere within its position space (Fig. 1(a)). Two typical designs, compact one and non-compact one, are shown in Figs. 1(b) and 1(c).

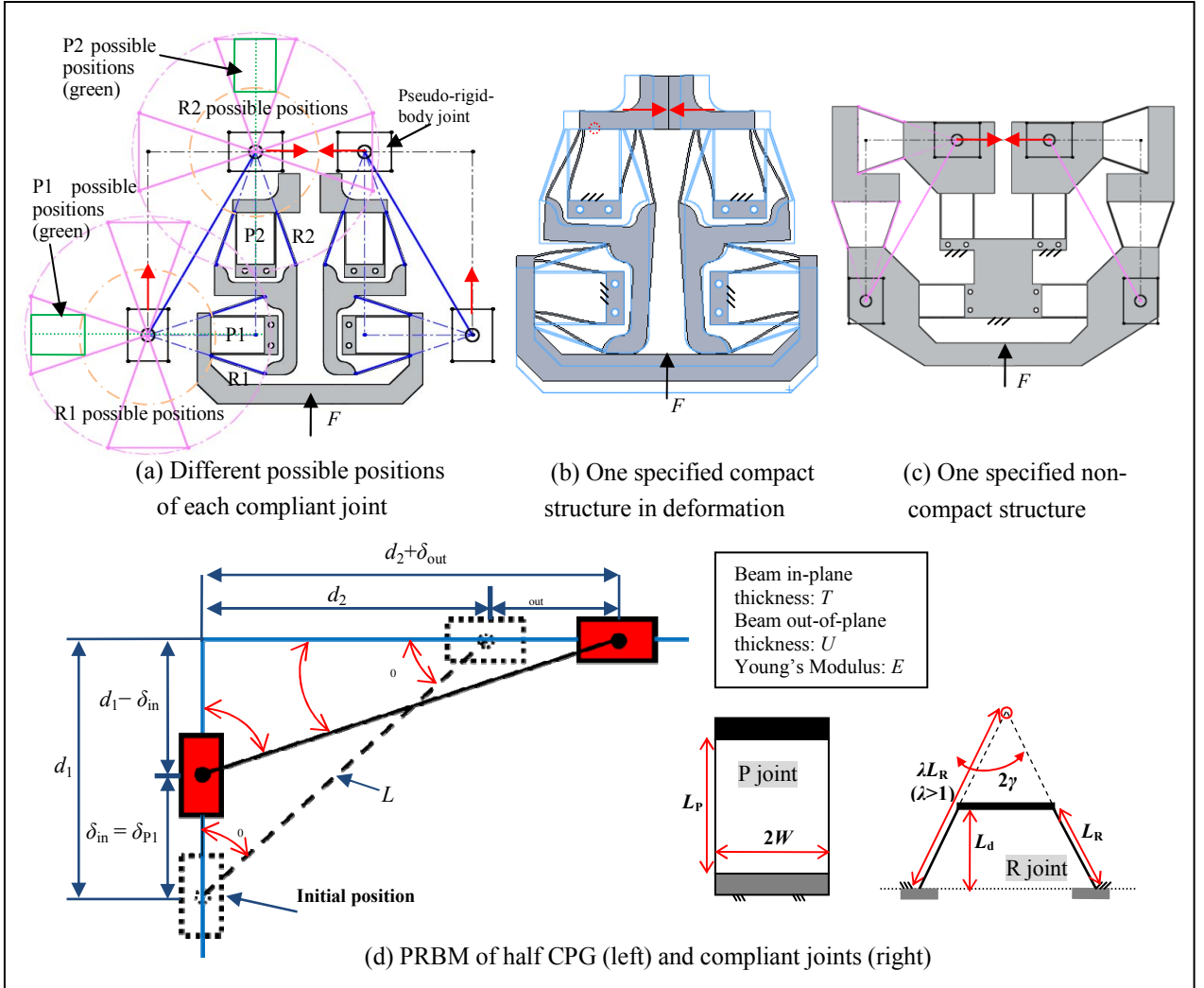


Figure 1. Conceptual design of planar 2-PRRP CPGs

III. PRBM-BASED APPROXIMATE ANALYTICAL KINETOSTATIC MODELING

The primary motion of each compliant joint associated with the input motion can be firstly derived as

$$\begin{cases} \theta_{R1} = \alpha - \alpha_0 = \arccos\left(\frac{d_1 - \delta_{in}}{L}\right) - \alpha_0 \\ \theta_{R2} = \beta_0 - \beta = \beta_0 - \arcsin\left(\frac{d_1 - \delta_{in}}{L}\right) \\ \delta_{P1} = \delta_{in} \\ \delta_{P2} = \delta_{out} = L \sin(\arccos\left(\frac{d_1 - \delta_{in}}{L}\right)) - d_2 \\ \quad = L \sqrt{1 - \left(\frac{d_1 - \delta_{in}}{L}\right)^2} - d_2 \end{cases} \quad (1)$$

where θ_{R1} and θ_{R2} are the rotational angle of the compliant R joints, and δ_{P1} and δ_{P2} is the translational displacement of the compliant P joints. δ_{in} is the input motion from the linear actuator, and δ_{out} is the primary output motion of the jaw where the bottom-surface center of the motion stage of the P2 joint is specified as the output point of the jaw. The other symbols are the geometrical parameters as indicated in Fig. 1(d).

Using Eq. (1), the amplification ratio between the output displacement and the input displacement is obtained as:

$$\frac{\delta_{out}}{\delta_{in}} = \frac{L \sqrt{1 - X^2} - d_2}{d_1 - LX} \quad (2)$$

where $X = \frac{d_1 - \delta_{in}}{L}$.

The linear stiffness of each compliant joint is further obtained as follows [12, 16]:

$$\begin{cases} K_{P1} = K_{P2} = 24 \frac{E(UT^3/12)}{L_P^3} \\ K_{R1} = K_{R2} = 8(3\lambda^2 - 3\lambda + 1) \frac{E(UT^3/12)}{L_d / \cos \gamma} \end{cases} \quad (3)$$

where K_{P1} and K_{P2} is the translational stiffness of the compliant P joints, and K_{R1} and K_{R2} are the rotational stiffness of the compliant R joints. The other symbols denote the geometrical parameters and material property as indicated in Fig. 1(d).

The use of Eqs. (1) and (2) yields the potential energy of the system with regard to the input displacement:

$$\begin{aligned} U &= 2 \times \frac{1}{2} K_{P1} \delta_{in}^2 + 2 \times \frac{1}{2} K_{R1} \theta_{R1}^2 + 2 \times \frac{1}{2} K_{R2} \theta_{R2}^2 + 2 \times \frac{1}{2} K_{P2} \delta_{out}^2 \\ &= K_{P1} \delta_{in}^2 + K_{R1} \left[\arccos\left(\frac{d_1 - \delta_{in}}{L}\right) - \alpha_0 \right]^2 \\ &\quad + K_{R2} \left[\beta_0 - \arcsin\left(\frac{d_1 - \delta_{in}}{L}\right) \right]^2 \\ &\quad + K_{P2} \left[L \sqrt{1 - \left(\frac{d_1 - \delta_{in}}{L}\right)^2} - d_2 \right]^2 \end{aligned} \quad (4)$$

The input force is finally obtained using principle of virtual work [3]:

$$\begin{aligned} F &= 2 \frac{\partial U}{\partial \delta_{in}} \\ &= 2K_{P1}(d_1 - LX) + 2K_{R1}[\arccos(X) - \alpha_0] \left(\frac{1}{\sqrt{1 - X^2}} \right) / L \\ &\quad + 2K_{R2}[\beta_0 - \arcsin(X)] \left(\frac{1}{\sqrt{1 - X^2}} \right) / L \\ &\quad + 2K_{P2}[L\sqrt{1 - X^2} - d_2] \left(\frac{X}{\sqrt{1 - X^2}} \right) \end{aligned} \quad (5)$$

IV. CASE STUDIES

In this section, the two presented designs (Figs. 1(b) and 1(c)) are studied in details. Both designs have the same geometrical parameters for the CPG's PRBM (Fig. 1(d)) and each compliant joint/module. All these parameters are assigned valued as listed in Tables 1 and 2 for the case studies.

Nonlinear Finite element analysis (FEA) software, Comsol, is used to simulate the two cases with comparison to the analytical model obtained in Section III. Here, we set up the simulation as follows: free 10-node tetrahedral element and extra fine meshing with maximum element size of 10.50 mm and minimum element size of 0.45 mm. Material properties are: Young's modulus $E=69$ GPa and yield strength $\sigma_s=276$ MPa. The input displacement is limited to less than 1 mm to ensure that the material deformation is within the yield strength.

TABLE 1. GEOMETRICAL PARAMETERS OF CPG'S PRBM

L	162.69 mm
d_1	141.49mm
d_2	80.29 mm
α_0	29.57°
β_0	60.43°

TABLE 2. JOINTS' GEOMETRICAL PARAMETERS AND MATERIAL PROPERTY

T	1.00 mm
U	10.00 mm
L_R	42.25 mm
L_P	40.00 mm
L_d	40.00 mm
W	14.50 mm
λ	2.17
γ	18.78°
E	69 GPa
σ_s	276 MPa

Figures 2 to 6 show the FEA results for the two case studies. It is suggested that the analytical model agrees better with the FEA model of the non-compact CPG (Fig. 1(c)) than that of the compact CPG (Fig. 1(b)). The output of the compact design (Fig. 3) is much smaller than the analytical model as predicted. This is mostly due to the fact that the compact CPG has worse load transmissibility. It is noted that the results in Fig. 4 show that the amplification ratio is not a constant value with slight fluctuation.

However, the compact CPG has better characteristics in its compact configuration (Fig. 1(b)), and smaller parasitic translation and parasitic rotation (Figs. 5 and 6).

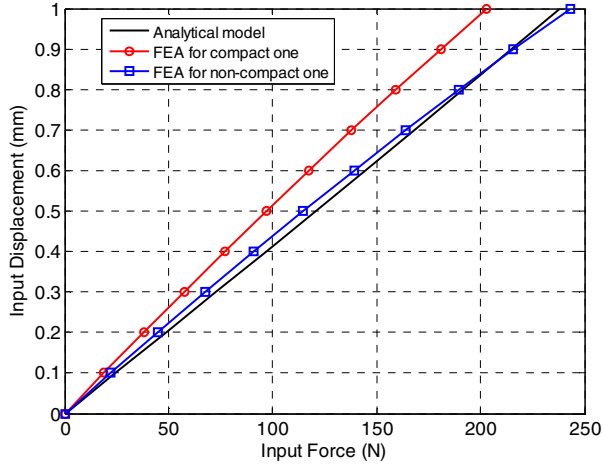


Figure 2. Relationship between input force and input displacement

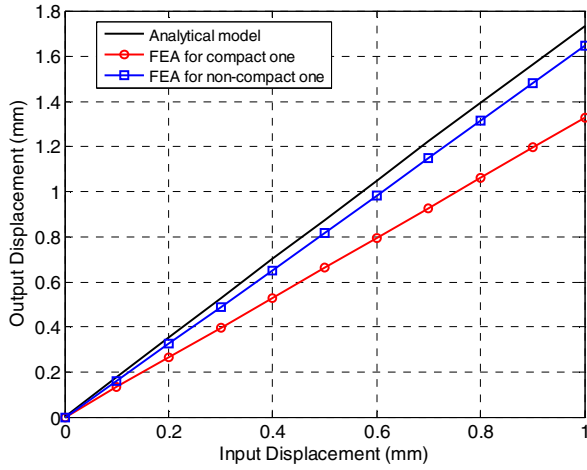


Figure 3. Relationship between input displacement and output displacement

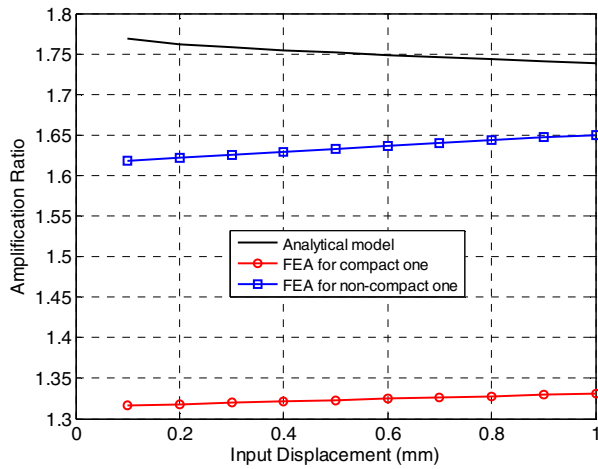


Figure 4. Relationship between input displacement and amplification ratio

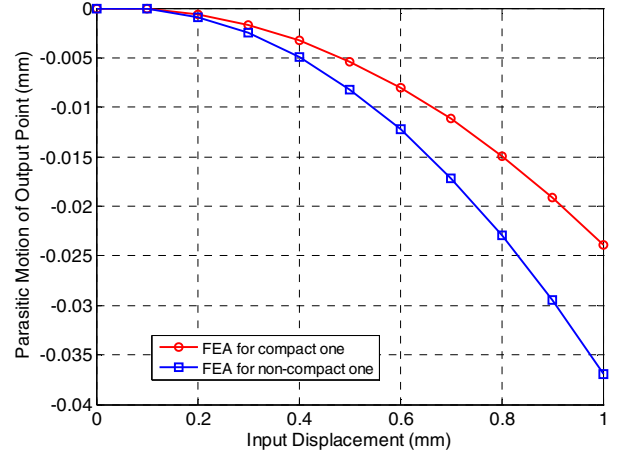


Figure 5. Relationship between input displacement and parasitic translation of output point

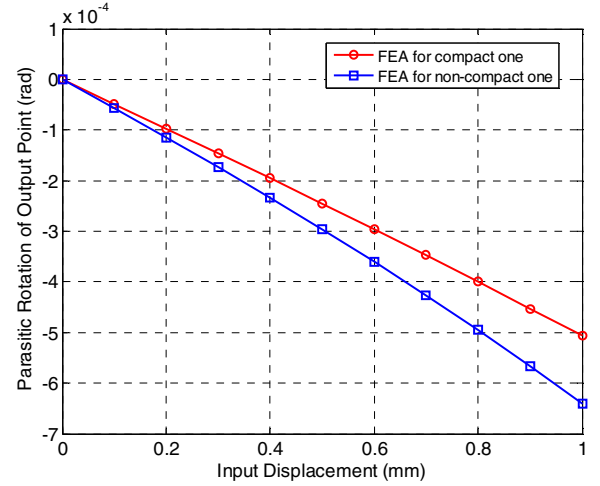


Figure 6. Relationship between input displacement and parasitic rotation of output jaw

V. FURTHER CONSIDERATIONS

The performance characteristics of the proposed CPG based on the 2-PRRP mechanism not only change with the structure reconfiguration (joint positions) but also are influenced by the geometrical parameters of the PRBM of the PRRP mechanism, compliant joint type, and beam length and thickness. Therefore, more optimization work on selecting the influence factors can be conducted to obtain better performance characteristics. However, the structure reconfiguration is a paramount method to design a CPG.

Alloy can be selected to fabricate the CPG with AL6061-T6 and AL7075-T6 being recommended due to their low internal stresses, good strength and phase stability, and relatively low cost.

The CPG can be fabricated monolithically from a piece of blank plate using the well-known planar manufacturing methods such as CNC multi-axis milling machining, wire electrical discharge machining (wire EDM), and water jet.

In order to control the CPG's two jaws to handle a sub-millimeter object, a PZT actuator can be adopted to produce the input force due to its merits including large force, high stiffness, fast response, compact size and up to nano-positioning precision. The compression force on the object grasped by two jaws may be measured by two strain gauges bonded to the two jaws to avoid crushing the sub-millimeter objects. In addition, the strain gauge can be used for the sophisticated closed-loop control. A visual assembled prototype for the compact design (Fig. 1(b)) incorporating the PZT actuator and strain gauges is shown in Fig. 7. A similar 3-D printed prototype of the non-compact design (Fig. 1(c)) is also presented in this paper as shown in Fig. 8.

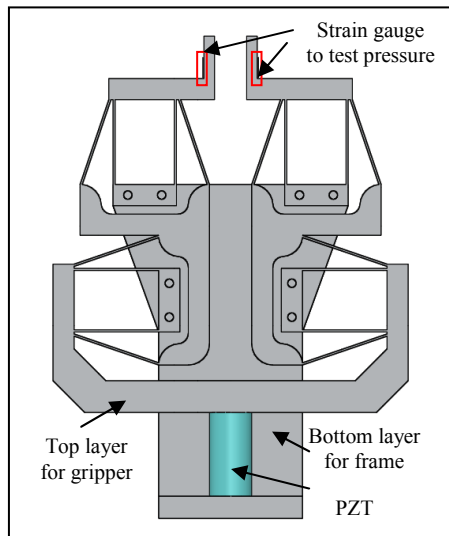


Figure 7. A visual assembled prototype for compact design



Figure 8. A similar 3-D printed prototype for non-compact design

VI. CONCLUSIONS

The position space concept has been used in this paper to design new CPGs based on a 2-PRRP mechanism. The structure of compliant mechanisms can be reconfigured based on specific requirements through changing the positions of each compliant module thereof within their position spaces. As a result, two typical CPGs (compact one and non-compact one) have been designed and analyzed using the analytical model and/or nonlinear FEA model.

It has been observed that the analytical model agrees better with the FEA model of the non-compact CPG than that of the compact CPG.

This work provides a solid starting point for further physical testing and control of CPGs for manipulating sub-millimeter objects.

ACKNOWLEDGEMENT

The authors would like to thank Royal Irish Academy Charlemont Grants 2015 and EPSRC (EP/K018345/1) to provide financial support to this research.

REFERENCES

- [1] SFI report: http://www.sfi.ie/assets/media/files/downloads/Funding/Funding%20Calls/investigators_programme/IvP%202014%20full%20description%20of%20themes.pdf
- [2] EPSRC grant details: <http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/K018345/1>
- [3] L.L. Howell, *Compliant Mechanisms*, New York: John Wiley & Sons, 2001.
- [4] L.L. Howell, S.P. Magleby, and B.M. Olsen, *Handbook of Compliant Mechanisms*, John Wiley & Sons, New York: John Wiley & Sons, 2013.
- [5] S. Kota, J. Joo, Z. Li, S.M. Rodgers, and K. Sniegowski, "Design of compliant mechanisms: applications to MEMS," *Analog Integr. Circ. Sig. Process.*, vol. 29 (1-2), pp. 7–15, 2001.
- [6] M.N.M. Zubir, B. Shirinzadeh, and Y. Tian, "Development of novel hybrid flexure-Based microgrippers for precision micro-object manipulation," *Rev. Sci. Instrum.*, vol. 80, pp. 065106, 2009.
- [7] A. Nikoobin, and H.M. Niaki, "Deriving and analyzing the effective parameters in microgrippers performance," *Scientia Iranica B*, vol. 19(6), pp. 1554-1563, 2012.
- [8] S.K. Nah, and Z.W. Zhong, "A microgripper using piezoelectric actuation for micro-object manipulation," *Sensor Actuat. A-Phys.*, vol. 133, pp. 218–223, 2007.
- [9] M. Goldfarb, and N. Celanovic, "A Flexure-based gripper for small-scale manipulation," *Robotica*, vol. 17, pp. 181–187, 1999.
- [10] J.D. Beroz, S. Awtar, M. Bedewy, T. Sameh, and A.J. Hart, "Compliant microgripper with parallel straight-line jaw trajectory for nanostructure manipulation," *Proceedings of 26th American Society of Precision Engineering Annual Meeting*, Denver, USA, 2011.
- [11] A.N. Reddy, N. Maheshwari, D.K. Sahu, and G.K. Ananthasuresh, "Miniature compliant grippers with vision-based force sensing," *IEEE Trans. Robot.*, vol. 26(5), pp. 867–877, 2010.
- [12] G. Hao, and X. Kong, "Design and modelling of a self-adaptive compliant parallel gripper for high-precision manipulation," *Proceedings of the ASME 2012 International Design Engineering Technical*

- Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2012), August 12-15, Chicago, IL, USA, 2012.
- [13] H. Li, and G. Hao, "A Constraint and position identification (CPI) approach for the synthesis of decoupled spatial translational compliant parallel manipulators," *Mech. Mach. Theory*, vol. 90, pp. 59–83, 2015.
 - [14] G. Hao, H. Li, R. Kavanagh, "Design of Decoupled, Compact, and Monolithic Spatial Translational Compliant Parallel Manipulators Based on the Position Space Concept", *Proc. of the IMechE Part C: J. of Mechanical Engineering Science*. (In press)
 - [15] H. Li, and G. Hao, "Compliant mechanism reconfiguration based on position space concept for reducing parasitic motion," in *ASME 2015 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, August 2–5, Boston, Massachusetts, USA, 2015, DETC2015-46434. (In press)
 - [16] G. Hao, Q. Meng, and Y. Li, "Design of large-range XY compliant parallel manipulators based on parasitic motion compensation," *Proceedings of the ASME 2013 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, Portland, USA, 2013.

Analysis and Characterisation of a Kinematically Decoupled Compliant XY Stage

Xavier Herpe, Ross Walker, Xianwen Kong, Matthew Dunnigan

Department of Mechanical Engineering, School of Engineering and Physical Sciences, Heriot-Watt University
Edinburgh, EH14 4AS, UK

Email: xh28@hw.ac.uk, r.walker@hw.ac.uk, x.kong@hw.ac.uk, m.w.dunnigan@hw.ac.uk

Abstract— Due to today's product miniaturisation trend and a constant need for fast changeover to meet the rapid production change, there is a need to develop high yield affordable systems for miniaturised product assembly. Compliant micro-positioning stages offer low-cost high precision and high repeatability but limited workspace and nonlinear behaviour. This paper presents a kinematically decoupled XY stage with a travel range of $\pm 2\text{mm}$, less than 2.36% coupling and a ratio between axial (X/Y) and radial (Z) stiffness of 2.3. The stage is designed to be integrated within the fine positioning system of a hybrid miniaturised product assembly system.

Index Terms—Compliant mechanism, kinematic decoupling, PRBM, stiffness, vibration.

I. INTRODUCTION

Due to their advantages such as compactness, cost reduction and enhanced performances, compliant XY stages are a promising alternative to conventional linear stages. They have a wide range of applications, such as: fibre alignment; semi-conductor positioning; ultra-precision micromachining centres; scanners for Atomic Force Microscope (AFM); and micro-assembly [1-5]. Their inclusion in micro-motion applications has allowed for accuracy and repeatability values in the nanometre scale [2]. Compliant stages have been reported to have no backlash; no friction; no noise emission and no need for lubrication [6-8]. However, they also have several disadvantages such as non-linear behaviour, limited working area, and off-axis deviation which induce errors if neglected [6, 7]. Serial compliant motion stages often combine two 1-DOF compliant prismatic joints, and have the advantage of being easier to design and fabricate and are naturally decoupled [8], but have lower accuracy and higher inertia than parallel motion stages.

Clearly there is a motivation to design a low-cost kinematically decoupled parallel 2-DOF micro-motion stage with a large workspace. Unlike micro-motion applications, miniaturised product assembly doesn't require nanometre scale accuracy but requires a larger range of motion. It is

desirable to have a high stiffness ratio and minimal cross-coupling. Parallelogram structures with flexure hinges are used in the design of XY stages to generate translational motion. The four main types of parallelogram structure are presented in Fig. 1. The design criteria are to keep the system's structural frequency high to improve the system's response time [9], the ratio between off-axis stiffness and axial stiffness as high as possible [10], and the cross-axis deflection as low as possible [7]. In this context, the biggest challenge is to achieve large stroke with kinematic decoupling. For instance, a 4-PP (Prismatic joint) XY micro-motion stage with complete decoupling has been developed in [11]. The experimental results show that the stage has a first natural frequency of 720.52Hz and has less than 5% coupling but only has a potential $105\mu\text{m} \times 105\mu\text{m}$ displacement range. A stage with a working range of $5 \times 5 \text{ mm}^2$ and a coupling error of less than $39.15\mu\text{m}$ was designed in [12]. A more complex fully decoupled XY micro-positioning stage with parasitic translation compensation was designed in [13]. The first natural frequency is 100Hz for a workspace of $10 \times 10 \text{ mm}^2$. Larger displacement is achieved in [3] by designing a 1-DOF micro-positioning stage with an 11.033mm stroke. This stage has a multitude of double compound parallelogram structures serially connected.

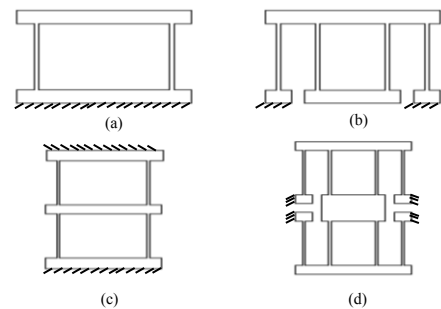


Fig. 1. Common parallelogram structures. a) Basic parallelogram, b) Double parallelogram, c) Compound basic parallelogram and d) Compound double.

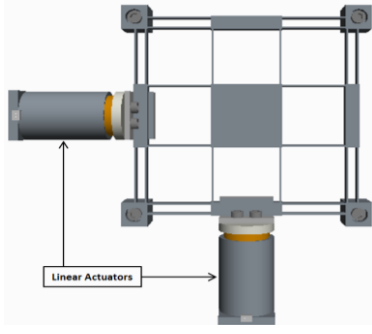


Fig. 2. CAD model of the proposed XY compliant stage, with mounted linear actuators.

The first natural frequency is 60Hz, mainly because of the type of parallelogram structures used. Based on the same structure, [14] designed an XY stage which can achieve a $10.5 \times 10.5 \text{ mm}^2$ workspace with sub-micrometre accuracy and has resonant frequencies of 48.3Hz, 48.7Hz and 100.Hz along the X, Y and Z directions respectively. Finally, a 4-PP XY stage designed in [15] can achieve a workspace of $20 \times 20 \text{ mm}^2$ with a coupling of less than 1.56%. It is important to note that large workspace often implies large footprint, which is not acceptable in confined environment.

To avoid cross-axis deflexion and low stiffness, the XY stage designed in this paper is exclusively composed of compound basic parallelogram structures (Fig. 1 c)) with leaf-spring type flexures. Since for compliant mechanisms a very small elongation of the beam is tolerated, a small translation is possible without cross-axis error as long as the elongation of the beam is much smaller than the length of the beam [16]. Although the range of motion is limited, this design allows for a high structural rigidity and decoupled motion. The accuracy, repeatability and working range rely heavily on the selected actuators. The prototype of this stage will later be driven by voice-coil actuators because they are easier to control and implement than electromagnetic actuators [17, 18] and have a centimetre range stroke, no backlash, and can be controlled by force or position [3, 19]. A CAD model of the XY stage with mounted actuators is shown in Fig. 2.

This paper presents the analysis and characterisation of an XY-stage with a simple structure. The force-displacement characteristics of the stage are defined and the impact of dimensional parameters on the stiffness and travel range is evaluated. The linear models are derived in Section II, the Finite Element Analysis (FEA) of the stage is presented in Section III and experimental results are presented in Section IV. Finally, Section V concludes this paper.

II. LINEAR ANALYTICAL MODEL

As only basic compound parallelograms are used for the XY stage and the lengths of the beams are all equal, each leaf-spring flexure can be considered as a spring with a stiffness K linking the stage to the base. The springs are considered as being in a parallel configuration. Hence, the overall stiffness of the stage is twelve times the single beam stiffness K . This configuration allows for deriving the theoretical force-displacement relationship using a single

beam deflection model. For comparison, the force-displacement relationship of the stage will be determined using the Pseudo-Rigid Body Model (PRBM) and Euler-Bernoulli beam deflection theory.

A. Pseudo-Rigid Body Model

The PRBM was first introduced by Howell and Midha in 1994 [20], and allows flexible elements to be modelled as rigid bodies connected together by torsional springs undergoing large non-linear deflexion. This model does not take into account shear stress as a result of flexures and axial deformation [17, 21]. Therefore, it will only be considered accurate for determining the force displacement relationship of the stage for small displacements [22, 23]. Based on the work from [24], a fixed-guided beam is represented as three rigid links connected with torsional springs. The PRBM representation of a single beam is presented in Fig. 3. From [24], the displacement of the end point of the beam is given by:

$$\delta = \gamma l \sin \varphi \quad (1)$$

where l is the total length of the beam, γ is the characteristic radius factor from [24] and φ is the angle between the rigid link and the origin. From Euler-Bernoulli beam deflexion theory, the slope and moment at the midpoint of a fixed-guided beam is zero. The torque is therefore derived from the input force and beam's length at this midpoint:

$$T = \frac{\gamma l}{2} F_a \cos \varphi = K_T \varphi \quad (2)$$

where T is the torque applied at one of the pivots centre point, K_T is the torsional spring stiffness of this pivot and F_a is the external force applied at the end of the beam. From [24], the torsional spring stiffness is given by:

$$K_T = 2\gamma K_\varphi \frac{EI}{l} \quad (3)$$

where K_φ is the pseudo-rigid-body stiffness coefficient given by [24], E is the Young's modulus of the material, and I is the second moment of area of the beam given by $I = bh^3/12$ where b is the width of the beam and h is the thickness of the beam. As the force is applied vertically, $\gamma = 0.8517$ and $K_\varphi = 2.67617$ [24].

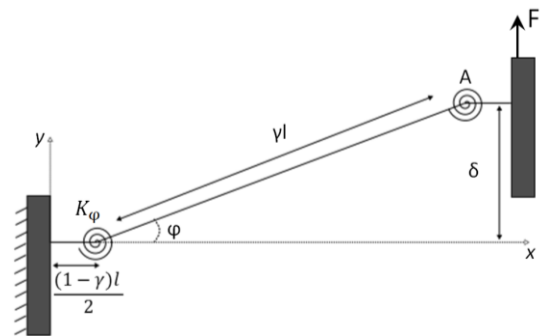


Fig. 3. Pseudo-Rigid Body Model of a single beam

Rearranging Eqs. (1) to (3), the relationship between force and displacement is:

$$F_a = \frac{4K_\phi EI \sin^{-1}\left(\frac{\delta}{\gamma l}\right)}{l^2 \sqrt{\frac{\gamma^2 l^2 - \delta^2}{\gamma^2 l^2}}} \quad (4)$$

The resulting linear force/displacement obtained in Eq. (4) differs from the resulting equation derived in [16].

B. Euler-Bernoulli Model

The Euler-Bernoulli representation of a beam is presented in Fig. 4. As the displacement at the end of the beam δ is assumed to be very small, the curvature of the beam is also assumed to be small, and can be described as:

$$\frac{1}{\rho} = \frac{d^2 y}{dx^2} = \frac{M}{EI} \quad (5)$$

where ρ is the radius of curvature of the beam; M is the moment applied to the beam; and $d^2 y/dx^2$ is the curvature of the beam. For a fixed-guided beam, the expression of the bending moment M applied to the beam is:

$$EIy''(x) = M = F_A(l - x) - M_A \quad (6)$$

where F_A and M_A are the force and the moment respectively applied at point A. From Eqs. (5) and (6), the slope at any point x on the beam is:

$$EIy'(x) = \int M \cdot dx = (F_A l - M_A)x - \frac{F_A}{2}x^2 + C_1 \quad (7)$$

The boundary conditions dictate that at $x = 0$, $y'(x) = 0$. Hence, the constant $C_1 = 0$ is 0. For $x = l$, the following relationship is determined:

$$\frac{F_A l^2}{2EI} - \frac{M_A l}{EI} = 0 \quad (8)$$

Therefore the force F_A can be expressed as a function of the moment M_A :

$$F_A = \frac{2M_A}{l} \quad (9)$$

The deflection at any point x on the beam is:

$$EIy(x) = \iint M \cdot dx = \frac{(F_A l - M_A)x^2}{2} - \frac{F_A x^3}{3} + C_2 \quad (10)$$

Given the boundary condition, $C_2 = 0$. Substituting Eq. (9) into Eq. (10), the maximum deflection of the beam at $x = l$ can be expressed as:

$$\delta = \frac{F_A l^3}{12EI} \quad (11)$$

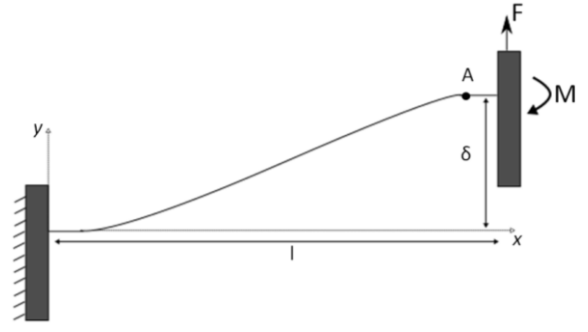


Fig. 4. Euler-Bernoulli beam deflection model of a single beam

III. FINITE ELEMENT ANALYSIS

Nonlinear FEA is carried out using ABAQUS for comparison with the analytical models, and to further study the behavior of the XY stage under large displacements. The dimensions of the beams are chosen to be 35mm length, 6mm width and 1mm thick. The impact of the length and width on the stiffness and travel range will be evaluated.

A. Static Analysis

The static analysis consists in applying a load on the stage to study the displacement and stress engendered. The material used for the analysis is Aluminium 7075-T6 because it is widely used in the literature [2, 3, 11, 17]. It has a Young's modulus (E) of 71.7GPa; a Poisson's ratio (ν) of 0.33; a density (ρ) of 2810kg/m³; and yield strength (σ_{\max}) of 503MPa. The material is described as hyperelastic and the model is based on the Neo-Hookean solid model. Force-displacement, stress, kinematic coupling and modal analyses are further deliberated on below:

1) Force-Displacement Analysis

To further study the force-displacement relationship, a force of 700N is gradually applied and the displacement along the direction of motion is recorded. The results are then compared with the PRBM and Euler-Bernoulli analytical models in Fig. 5a) and Fig. 5b). It has been determined that in the range of 0 to 70N, the displacement error varies from 8.8% to -1.1% for the PRBM model and from 4.5% to -5.9% for the Euler-Bernoulli model, but increases significantly after 70N. These results clearly show that although some linear behaviour can be observed in the range of 0 to 0.55mm, as in [16], PRBM is not suitable for large displacements, and the accuracy of simple derivations based on elastic beam theory is limited. This nonlinear behaviour is attributed to the parallelogram structure which causes a load stiffening phenomena, resulting in a constantly increasing stiffness of the beams. From Fig. 5c) and Fig. 5d), it can be seen that the stiffness of the stage increases as the beam's length becomes smaller or as the beam's width becomes larger. These parameters will be taken into account for topology optimisation of the final design.

1) Stress Analysis

A stress analysis is used to define the maximum allowable displacement, limited in function of the yield strength of the material.

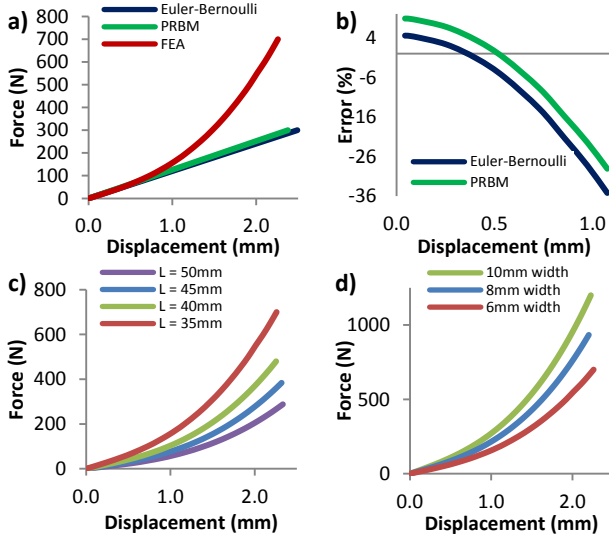


Fig. 5. a) FEA of the force-displacement relationship compared with the analytical models, b) Error from comparison with analytical models, c) FEA with variable beam's length, b) FEA with variable beam's width.

The expression for the maximum stress, occurring at the end of the beams on the cross-section's farthest edge from the neutral axis, is given by:

$$\delta_{max} = \frac{\sigma_{max} l^2}{3Eh} \quad (12)$$

The results of the stress analysis from the Euler-Bernoulli model and the nonlinear FEA model are shown in Fig. 5a). As for the force-displacement analysis, it is clearly shown that the analytical linear model deviates from the nonlinear FEA model. The Euler-Bernoulli and FEA models indicate a maximum allowable displacement of 2.865mm and 1.975mm respectively, which corresponds to an error of 45%. This is due to the linear model, in which tension loading is ignored. Observing Fig. 6b) and Fig. 6c), it can be seen that increasing the beam's length will increase the range of motion but the beam's width has almost no influence.

2) Kinematic coupling

As the XY stage is designed to be kinematically decoupled, a coupling analysis is carried out by first applying a preload force of 500N along the X direction, as presented in Fig. 7, corresponding to the theoretical maximum displacement calculated in Section II, and gradually applying a force from 0N to 500N along the Y direction. From Fig. 8a), the Y-displacement error with preloading is 11.5% at 50N and reduces to 2.44% at 500N. This error is mainly due to the increase in stiffness because the applied preload is very large. When a preload of only 70N is applied, this error is lower than 1.9%. From Fig. 8b), the cross-axis coupling, which means the parasitic displacement along the X direction for every unit displacement along the Y direction, is 0.28% at 50N and goes up to 2.36% at 500N, corresponding to a maximum coupling error of 44.8μm.

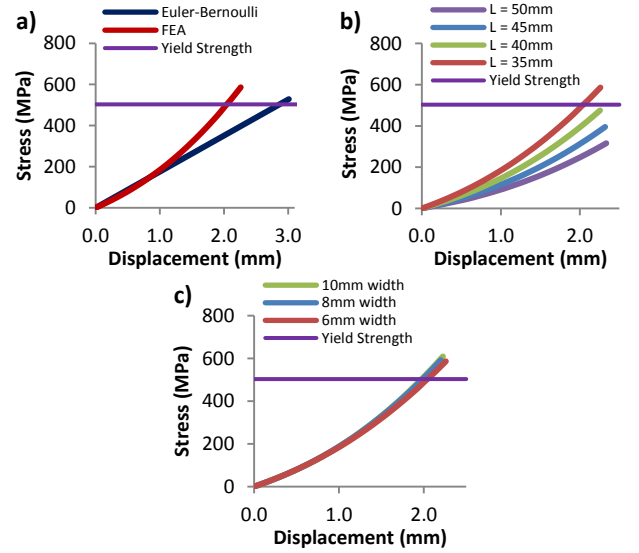


Fig. 6. Stress analysis of the XY stage a) in comparison with analytical models b) with variable beam's length, c) with variable beam's width.

B. Modal Analysis

An analysis of the frequency response of the stage is carried out with ABAQUS using the Lanczos Eigen solver. The first two modes, corresponding to vibrations along the X and the Y directions respectively, occur at 262.7Hz and the third mode, corresponding to vibrations along the Z direction, has a frequency of 600.4Hz, corresponding to a stiffness ratio of 2.3 between X/Y and Z. The frequency response of the stage is also evaluated as a function of the geometrical parameters. The results from Fig. 9 show that the width of the beam only has impact on the frequency response along the Z direction while the length of the beam clearly modifies the stiffness along all the directions.

IV. EXPERIMENTAL STUDY

As in [25], a 3D printed prototype made with Polylactic Acid (PLA) was fabricated. Although the material properties of 3D-printed polymers have many uncertainties, this was the simplest, fastest and cheapest way to fabricate a prototype and study the force-displacement relationship and frequency response of the XY stage. The prototype was printed with a Replicator® 2 from MakerBot.

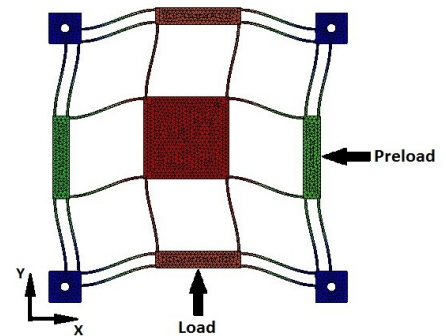


Fig. 7. Deformed XY stage with forces applied along X and Y directions

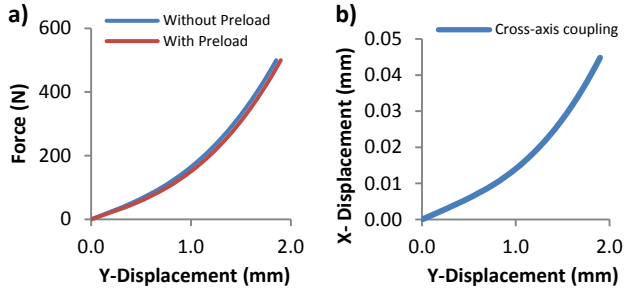


Fig. 8. a) Y-Displacement with and without preload, b) Cross-axis coupling.

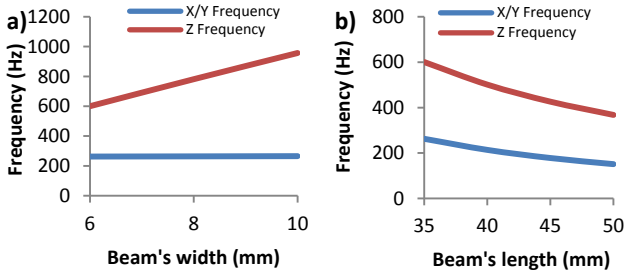


Fig. 9. Frequency response of the XY stage for different a) beam's widths, b) beam's lengths

A. Frequency Response Analysis

In order to obtain the frequency response of the XY stage, a testing rig is set up with a shaker to generate guided vibrations along the X direction of motion and measure the amplitude with an accelerometer (ICP-T356A16) placed at the centre of the stage. A second accelerometer will be added later to characterise the coupling between the center and edges of the stage. The sensitivity of the accelerometers is 100mV/G and their output signals are processed by a Dual Channel Accelerometer Amplifier (FE-376-IPF) and acquired by a Data Acquisition card from National Instruments. Labview is used to obtain the frequency domain response using the Fast Fourier Transform (FFT). The sampling rate is 10kHz. The testing rig setup with two accelerometers is presented in figure 11.

Results from the literature [25, 26] show that the Young's modulus of PLA falls between 1.28GPa and 3.5GPa. The response of the stage with these Young's moduli has been modelled. For simplification, the material is considered as a hyperelastic isotropic material for the simulation. A mass of 17g is added at the centre of the model to emulate the mass of the accelerometer and the screw used for the frequency response test. The density of the material once printed is 1150kg/m³. The stage is printed with a 100% filling, a layer thickness of 0.1mm and a 0/90° orientation. The beams have a length of 26.25mm, a width of 4.5mm and a thickness of 0.9mm.

1) Test 1: single accelerometer

The results from the test with a single accelerometer placed at the center of the stage are presented in Fig. 11a). A peak in amplitude is observed when the vibrations generated by the shaker reach 75Hz. This peak is more obvious on the Y-axis as it represents free vibrations, while vibrations along the X-axis are guided by the shaker.

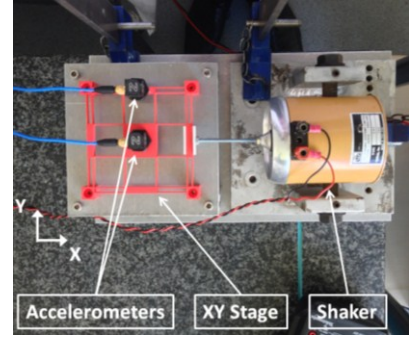


Fig. 10. Vibration testing setup

The FEA model indicates the first natural frequency occurs between 52.67Hz and 85.5Hz, when modelled with Young's modulus of 1.28GPa and 3.5GPa respectively. Hence, the FEA model is in good agreement with the data acquired from the testing rig.

2) Test 2: dual accelerometer

The results from the test with two accelerometers are presented in Fig. 11b). The peak in amplitude occurs at 70Hz. It is postulated that the 5Hz shift in occurrence of the peak amplitude is due to the addition of a second accelerometer. The output of the first and the second accelerometers are X1/Y1 and X2/Y2 respectively. The amplitude of X1 is in agreement with the value of X measured in Test 1. The large difference between X1 and X2 clearly shows that the stage has partial vibration isolation because, unless the resonant frequency has been reached, vibration from the shaker neither engenders vibration along the Y direction nor on the edge of the stage.

B. Force-Displacement Test

The force-displacement relationship has been studied by applying a load along one direction and measuring the displacement with a dial gauge of 0.0254mm resolution, as presented in Fig. 12a). From Fig. 12b), the experimental results show that the load stiffening phenomena is not obvious, which may be due to plastic behaviour of the material. Characterising 3D Printed PLA is proven here to be a complex task, due to 3D printing attributes such as the orientation of the printed fibres, the layer thickness, density and effective Young's modulus of the material once 3D-printed. In addition, the study of a single beam will give different results than the study of the full structure because of the change in the fibre orientation results in anisotropy.

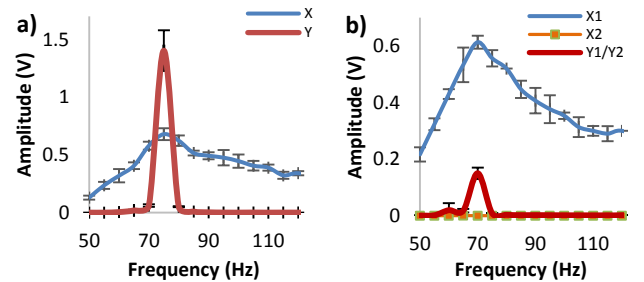


Fig. 11. Frequency response of the 3-D printed XY Stage with a) one accelerometer, b) two accelerometers.

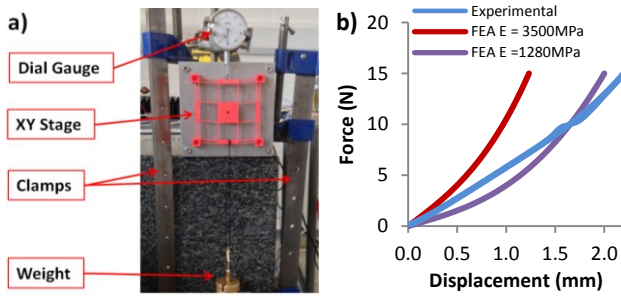


Fig. 12. a) Loading Test Rig, b) Experimental results compared to FEA results of the 3d-printed XY stage loading test.

V. CONCLUSION

The concept of a kinematically decoupled XY stage was studied and characterised. The theoretical travel range is $4 \times 4 \text{ mm}^2$ with a coupling error less than $44.8 \mu\text{m}$ and a stiffness ratio of 2.3. The impact of the beams geometrical parameters on the stiffness and the travel range of the XY stage were investigated and the results will be used as design criteria for the final design. Because the linear analytical models are inefficient when basic parallelograms are used, the finite element model will be used as a reference for additional study, such as buckling and dynamic loading. The experimental results from the 3D printed prototype cannot be generalized because of the uncertainties induced by the mechanical properties of PLA. However, the test rig used for the frequency response analysis has proven that the XY stage may have potential partial passive vibration isolation. From this study, a final model will be designed. The material will be Nylon-66 to reduce the force input requirement and to experiment the use of polymer for compliant XY stages, and it will be fabricated using abrasive-jet machining. The stage will be driven by voice-coil actuators and will be used as the fine positioning mechanism of a hybrid mini-assembly system. Further study will be carried out to investigate the impact of actuators offset and misalignment on the XY stage.

ACKNOWLEDGMENT

The authors would like to thank the support from the Engineering and Physical Sciences Research Council (EPSRC), UK under Grant No. EP/K018345/1, and Dr. D. Yurchenko for providing all the equipment necessary to the frequency measurement test.

REFERENCES

- [1] W. Dong, L. N. Sun, and Z. J. Du, "Design of a precision compliant parallel positioner driven by dual piezoelectric actuators," *Sensors and Actuators*, vol. 135, pp. 250–256, 2007.
- [2] K.-B. Choi and J. J. Lee, "Analysis and Design of Linear Parallel Compliant Stage for Ultra-precision Motion Based on 4-PP Flexural joint Mechanism," in *Proceeding of the International Conference on Smart Manufacturing Application (ICSMA)*, pp. 35–38, 2008.
- [3] L. Shaoqian, J. Yukun, L. Iok Peng, and X. Qingsong, "Design and optimization of a long-stroke compliant micropositioning stage driven by voice coil motor," in *12th International Conference on Control Automation Robotics & Vision (ICARCV)*, 2012, pp. 1716–1721.
- [4] Q. Qifeng and R. Du, "A vision based micro-assembly system for assembling components in mechanical watch movements," in *Proceeding of the International Symposium on Optomechatronic Technologies (ISOT)*, 2010, pp. 1–5.
- [5] E. D. Kunt, A. T. Naskali, I. S. M. Khalil, A. Sabanovic, and E. Yüksel, "Design and development of workstation for microparts manipulation and assembly," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 19, pp. 973–992, 2011.
- [6] J. Hesselbach, A. Raatz, and H. Kunzmann, "Performance of Pseudo-Elastic Flexure Hinges in Parallel Robots for Micro-Assembly Tasks," *CIRP Annals - Manufacturing Technology*, vol. 53, pp. 329–332, 2004.
- [7] B. P. Trease, Y. Moon, and S. Kota, "Design of Large-Displacement Compliant Joints," *Journal of Mechanical Design*, vol. 127, pp. 788–798, 2005.
- [8] P. M. Moore, M. Rakotondrabe, C. Cleve, and G. J. Wiens, "Development of a Modular Compliant Microassembly Platform with Integrated Force Measurement Capabilities," in *Proceeding of the 7th International Conference on MicroManufacturing (ICOMM)*, 2012.
- [9] S. Polit and J. Dong, "Design of high-bandwidth high-precision flexure-based nanopositioning modules," *Journal of Manufacturing Systems*, vol. 28, pp. 71–77, 7// 2009.
- [10] P. R. Ouyang, "A spatial hybrid motion compliant mechanism: Design and optimization," *Mechatronics*, vol. 21, pp. 479–489, 2011.
- [11] Y. Li, J. Huang, and H. Tang, "A Compliant Parallel XY Micromotion Stage With Complete Kinematic Decoupling," *IEEE Transactions on Automation Science and Engineering*, vol. 9, pp. 538–553, 2012.
- [12] J. Yu, Y. Xie, Z. Li, and G. Hao, "Improved Design and Characteristic Comparisons of Large-range Decoupled XY Compliant Parallel Micromanipulators," *Journal of Mechanisms and Robotics*, 2015.
- [13] G. Hao, Q. Meng, and Y. Li, "Design of Large-range XY Compliant Parallel Manipulators Based on Parasitic Motion Compensation," in *Proceedings of the ASME International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE)*, 2013.
- [14] X. Qingsong, "New Flexure Parallel-Kinematic Micropositioning System With Large Workspace," *IEEE Transactions on Robotics*, vol. 28, pp. 478–491, 2012.
- [15] G. Hao and X. Kong, "A Novel Large-Range XY Compliant Parallel Manipulator With Enhanced Out-of-Plane Stiffness," *Journal of Mechanical Design*, vol. 134, pp. 061009–061009, 2012.
- [16] X. Yang, W. Li, Y. Wang, L. Zhang, G. Ye, and X. Su, "Analysis of the Displacement of Compliant Double Parallel Fourbar Mechanism," in *Proceeding of the IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2009, pp. 2760–2763.
- [17] S. Xiao and Y. Li, "Design and analysis of a novel flexure-based XY micro-positioning stage driven by electromagnetic actuators," in *Proceeding of the International Conference on Fluid Power and Mechatronics*, 2011, pp. 953–958.
- [18] X. Shunli and L. Yangmin, "Development of a large working range flexure-based 3-DOF micro-parallel manipulator driven by electromagnetic actuators," in *Proceeding of the IEEE International Conference on Robotics and Automation* 2013, pp. 4506–4511.
- [19] J. Degang, S. Lining, L. Yanjie, Z. Yuhong, and C. Hegao, "Design and simulation of a macro-micro dual-drive high acceleration precision XY-stage for IC bonding technology," in *6th International Conference on Electronic Packaging Technology*, 2005, pp. 161–165.
- [20] L. L. Howell and A. Midha, "A Method for the Design of Compliant Mechanisms With Small-Length Flexural Pivots," *Journal of Mechanical Design*, vol. 116, pp. 280–290, 1994.
- [21] H. Tang and Y. Li, "Design, Analysis, and Test of a Novel 2-DOF Nanopositioning System Driven by DualMode," *IEEE Transactions on Robotics*, vol. 29, pp. 650–662, 2013.
- [22] R. Khan, M. M. Billah, and M. Watanabe, "Nonlinear modeling of compliant mechanism," in *Proceeding of the IEEE International Nanoelectronics Conference (INEC)*, 2013, pp. 294–297.
- [23] D. C. Handley, T.-F. Lu, Y. K. Yang, and C. Eales, "Workspace investigation of a 3 DOF compliant micro-motion stage," in *Proceeding of the ICARCV International Conference on Control, Automation, Robotics and Vision* 2004, pp. 1279–1284.
- [24] L. L. Howell, "Pseudo-Rigid-Body-Model," in *Compliant Mechanisms*, ed: John Wiley and Sons, 2001, pp. 135–217.
- [25] G. Hao and J. Yu, "A Completely Kinematically Decoupled XY Compliant Parallel Manipulator through New Topology Structure," in *Proceedings of the IFToMM Workshop on Fundamental Issues and Future Research Directions for Parallel Mechanisms and Manipulators*, 2014.

Analysis of Frictional Effects on the Dynamic Response of Gear Systems and the Implications for Diagnostics

Khaldoon F. Brethee¹, Jingwei Gao², Fengshou Gu¹, Andrew D. Ball¹

¹Centre for Efficiency and Performance Engineering, University of Huddersfield, Huddersfield, HD1 3DH.

²Basic Education College, National University of Defense Technology, Changsha, China
khaldoon.brethee@hud.ac.uk

Abstract—To develop accurate diagnostic techniques, this study examines gear dynamic responses based on a model including the frictional effect of tooth mesh process. An 8-DOF (degree-of-freedom) model is developed to include the effect of not only gear dynamics but also supporting bearings, a driving motor and a loading system. Moreover, it takes into account the nonlinearity of both the time varying stiffness and the time-varying forces due to the friction effect. The latter causes additional vibration responses in the direction of the off-line-of-action (OLOA). To show the quantitative effect of the friction, vibration responses are simulated under different friction coefficients. It shows that an increase in friction coefficient value causes a nearly linear increase in the vibration features. However, features from torsional responses and the principal responses in the line-of-action (LOA) show less changes in the vibration level, whereas the most significant increasing is in the OLOA direction. In addition, the second and third harmonics of the meshing frequency are more influenced than the first harmonic component for all motions. These vibration responses are more sensitive for indicating lubrication changes and enhancing conventional diagnostic features.

Keywords—diagnosis; sliding velocity; friction coefficient; vibration response; simulation;

I. INTRODUCTION

In order to achieve accurate diagnostics, a significant number of studies have been carried out on the modelling and simulation of gear dynamics. They have resulted in a wide variety of dynamic models available to predict the response of gear vibration in order to improve the current techniques of diagnosis and monitoring [1]. Simulation can be very valuable for gaining a better understanding of complex interaction between transmission components in a dynamic environment and hence improving machine diagnostics and prognostics. It helps to develop effective signal processing methods for characterizing complicated weak fault signatures contaminated by different noises [2]. Therefore, different dynamic models for various gearbox systems were presented in [3-8]. In which both torsional and translational vibration responses of gears were studied as a tool for aiding gearbox diagnostic inferences. Moreover, vibration relating to gear spalling

or tooth breakage [5, 9-11], tooth crack [12-15], tooth surface pitting and wear [16-19] have been used to study these faults in terms of gear fault monitoring and diagnostics. In general, these models included both translation and rotational motions to show the fault effects on the dynamic characteristics. However, most of presented models ignored the friction effect or did not consider the friction between gear tooth contacts effectively, which may give less accuracy of diagnostic results.

In the meantime, sliding friction between the tooth surfaces has been reported to be one of the main sources of power loss in geared transmissions as well as an effective source of undesired vibration and noise [20-22]. A six-degree-of-freedom dynamic model of a spur gear pair influenced by friction was proposed in [23, 24], which was used to examine gear design modifications on the gear dynamic responses. Cheng-zhong et al [25] and Howard et al [26] detailed gear dynamic model to study the friction effect on some vibration characteristics of the gears, but they did not signify the friction effects precisely.

This study develops a comprehensive model coupling with tooth friction and necessary transmission components. Then a series of simulation studies are carried out to investigate the characteristics of vibration features when a gearbox is influenced by different frictional cases. In particular, the mesh components will be examined in order to define effective and accurate vibration features for monitoring tooth surface defects and lubrication conditions.

II. MESHING MODEL

A. Gear Tooth Meshing Process

The relative contact motions between two compressed elastic bodies (gear teeth) are the origin of internal excitations of vibration in gearing. They result in contacting forces that act on both bodies with the same intensity but in opposite directions. Especially, these forces cause impacts at transitions of gear tooth meshing events within a mesh cycle. As shown in Fig. 1, the transition can be determined from the un-deformed gear

pair geometry. The line AB represents the line of action (LOA) between the tangential points to the base circle of the gears. There are four regions along AB due to the change of tooth pairs in contact. The actual zone of the line of action (LOA = CF) is represented as the line between the intersection of the addendum circle of pinion and gear with the line AB (points C and F). D and E are two points on the line AB such that $CE=DF=p_b$, where p_b is the base pitch of the gear tooth curve. Sections DP and PE are the single-tooth contact regions while sections CD and EF are the double-tooth contact regions. The main geometric relations of these regions used in this model are given by:

$$AB = (r_{bp} + r_{bg}) \tan \alpha = (r_p + r_g) \sin \alpha \quad (1)$$

$$LOA = CF = \sqrt{r_{ap}^2 - r_{bp}^2} + \sqrt{r_{ag}^2 - r_{bg}^2} - (r_p + r_g) \sin \alpha \quad (2)$$

$$AC = (r_p + r_g) \sin \alpha - \sqrt{r_{ag}^2 - r_{bg}^2} \quad (3)$$

$$AF = \sqrt{r_{ap}^2 - r_{bp}^2} \quad (4)$$

$$FB = (r_p + r_g) \sin \alpha - \sqrt{r_{ap}^2 - r_{bp}^2} \quad (5)$$

$$CE = DF = p_b = \frac{2\pi r_{bp}}{Z_1} = \frac{2\pi r_{bg}}{Z_2} \quad (6)$$

$$P_{angle} = \frac{2\pi}{Z_1} \quad (7)$$

$$\varepsilon_{ratio} = \frac{CF}{p_b} \quad (8)$$

$$\psi_{sp} = \tan^{-1} \left(\frac{AC}{r_{bp}} \right) \quad (9)$$

$$\psi_{ep} = \tan^{-1} \left(\frac{AF}{r_{bp}} \right) - \psi_{sp} \quad (10)$$

where α is the pressure angle and the ratio of the length of contact path to the base pitch is recognized as the contact ratio ε_{ratio} of a gear pair. The start angle of mesh cycle is named by ψ_{sp} while the end angle of LOA is ψ_{ep} as illustrated in Fig. 1. In addition, the time-varying moment arms $\rho_p(t)$ and $\rho_g(t)$ for the i^{th} meshing pair can be found by:

$$\rho_p(t) = AC + \text{mod}(r_{bp} \omega_p, p_b) \quad (11)$$

$$\rho_g(t) = FB + \text{mod}(r_{bg} \omega_g, p_b) \quad (12)$$

where the function $\text{mod}(x, y) = x - y \cdot \text{floor}(x/y)$ is the modulus function, if $y \neq 0$, ω_p and ω_g are the nominal speeds in (rad/s), and AC and FB are the geometric length constants. The sliding friction forces on each contact pair are denoting by F_{p1} , F_{g1} , F_{p0} and F_{g0} respectively.

These forces affect gear rotations by frictional torques about the gear centres and excite the off-line-of-action gear translations significantly as it will be explained later in form of $F_{fi}(t)$.

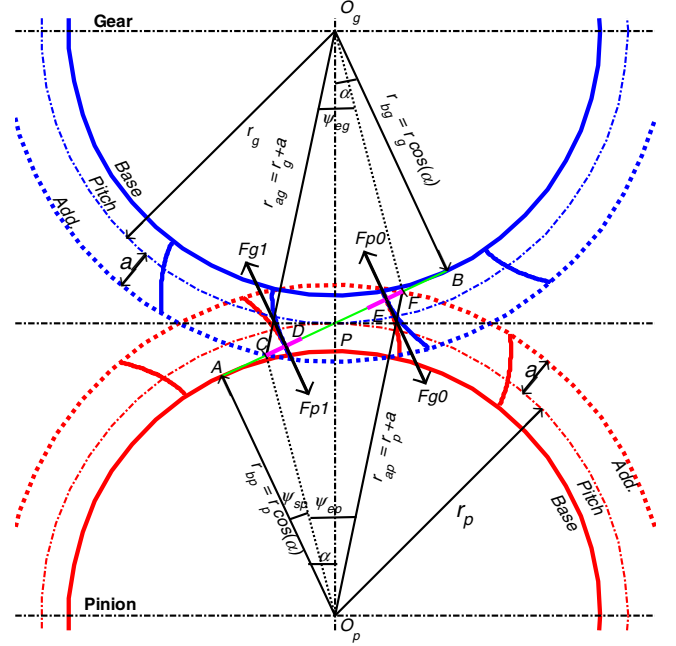


Figure 1 Meshing process of spur gear pairs

B. Varying Meshing Stiffness

The major variations in gear stiffness are caused by changes in meshing pair number. Spur gears have single-tooth and double tooth meshing appearing alternately during the process of mesh [27]. For normal spur gears with a contact ratio of more than one, the meshing pair numbers usually in the range between 1.0 and 2.0 [25, 28]. In existing literature, the tooth meshing stiffness is simplified as a rectangular wave [29] based on the equal load sharing formulation, which proposed by Vaishya and Singh in [22, 30, 31]. The existing model considered the sudden changing in the meshing stiffness value by a periodic square wave function at every stage. It makes the single-tooth meshing and the double tooth meshing appears alternately and changes suddenly during the mesh transitions. Figure 2(a) explains the various positions of gear tooth meshing events for identical spur gears within a pinion pitch duration angle P_{angle} as in (7). The dynamic model considers the pair of spur gears as two rigid disks coupled along the line of action through a time varying mesh stiffness $k(t)$ and damping $c(t)$ [28]. The mesh contact cycle starts from the angle ψ_{sp} at point C, denoting as the starting point of contact, where the addendum circle diameter of the gear intersects the active line of action (LOA). The mesh period of double pair tooth contact (M_{double}) begins when pair1 contact at point C whereas pair 0 is already in contact at point E, which is denoting as the ending point of single tooth contact. As the gears rotate, within the angle ψ_{ep} , the points of contact move along the line of action CF. When the pair 1 reaches the point D (the starting point of single tooth

contact), pair 0 disengages at point F (the finishing point of the mesh cycle) and leaves only the pair 1 in the single contact zone (M_{single}). In addition, while pair 1 reaches to point E, the next tooth pair engages at point C which starts another mesh cycle. Finally, when pair 1 rotates to point F, one meshing cycle is completed. Therefore, the meshing process leads to mesh stiffness that varying with time as illustrated in Fig. 2(b).

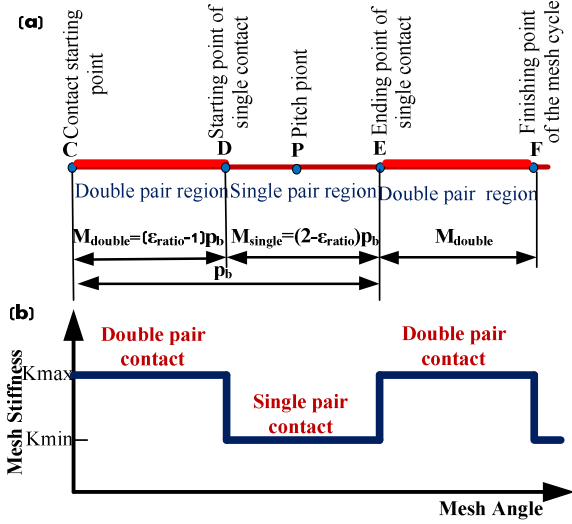


Figure 2 Mesh stiffness regions of meshing gear pair in one period

C. Varying Friction Effects between Tooth Surfaces

Friction forces and the nonlinearity excitation between tooth contact surfaces are the main sources of vibration [23]. Due to the velocity reversion at pitch point, friction can be associated with a large oscillatory component due to high forces in the sliding direction. The sliding velocity for each tooth pair in contact can be derived from meshing kinematics and oscillating torsional motion of the gear and pinion. This dependency upon the implicit non-linearity of vibrating velocity in the gear dynamic system [22]. The normal contact force and the friction force between pair of gears is calculated by Howard et al. [32], which is modelled as the combination of linear elastic and damping forces as shown in Fig. 3(a),

$$N_i = C(t)(r_{p1}\dot{\theta}_1 - r_{g1}\dot{\theta}_2 + \dot{y}_{p1} - \dot{y}_{g1}) + K_{mi}(t)(r_{p1}\theta_1 - r_{g1}\theta_2 + y_{p1} - y_{g1}) \quad (13)$$

where $i=0, 1$ denoting meshing tooth pair. The surface friction generated between the meshing tooth surfaces are:

$$F_{fi}(t) = \mu N_i \quad (14)$$

The dynamic friction formulation is modelled as a time-varying parameter; see Fig. 3(b). The friction coefficient (μ_0) formula of tooth surface is stated as constant; however it changes its sign with the direction of relative sliding velocity, i.e.

$$\mu = \mu_0 \operatorname{sgn}(V_s) = \begin{cases} \mu_0 & , V_s > 0 \\ -\mu_0 & , V_s < 0 \end{cases} \quad (15)$$

where, V_s refers to the sliding velocity at the contact point of interest. The sliding velocity is considered as the difference between surface velocities at each contact point. For i^{th} gear pair, its sliding velocity is:

$$V_{si} = \rho_{pi}(t)\omega_p - \rho_{gi}(t)\omega_g \quad (16)$$

For individual gear and pinion, $\rho(t)$ and ω are the radius of curvature of the corresponding contact point and the angular velocity of precise gear respectively. Hence, the friction moment of the pinion and gear is produced by the tooth friction forces $F_{fi}(t)$ and friction arms $\rho_i(t)$:

$$\begin{cases} T_{fp}(t) = \rho_{pi}(t) F_{fi}(t) \\ T_{fg}(t) = \rho_{gi}(t) F_{fi}(t) \end{cases} \quad (17)$$

The direction of friction torque is dependent on the instantaneous sliding velocity and the contact point location as illustrated in Fig. 3(c).

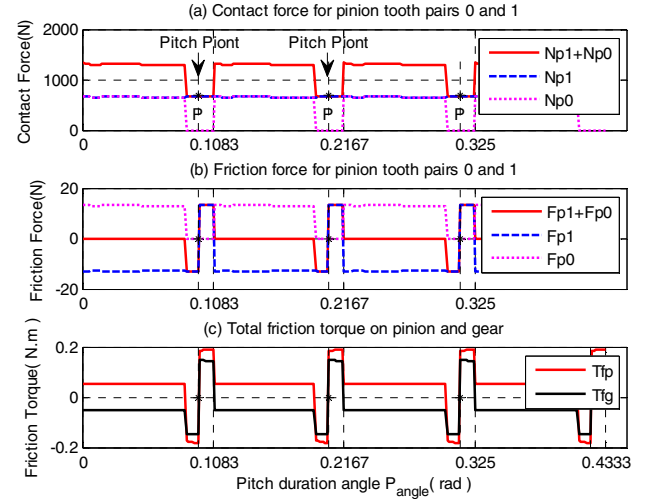


Figure 3 Variation of normal contact forces, friction forces and frictional torque with the pitch period

D. Friction Coefficient

Many parameters affect friction coefficient μ_0 because of the complex lubricating problem in gearing. Different empirical formulae were proposed to estimate the friction coefficient [33]. However, these empirical formulae for μ_0 , valid within certain ranges of key system parameters. They are not general and often represent certain lubricants, operating temperatures, speed and load ranges, and surface roughness conditions of roller specimens that might differ from those of the actual gear pair of interest [33]. In general, the theoretical friction coefficient is derived from elasto-hydrodynamic lubrication and tribology theory, however several experimental works show that, a constant friction coefficient is acceptable for dynamic analysis as indicated in [34-36]. Benedict and Kelley's empirical equation shows that, the coefficient of friction varies between 0.03 to 0.1 [37], furthermore the value of 0.1 or even values as high as 0.2 are commonly used in several gear dynamic models as explained in [36]. To get meaningful values of μ_0 , the variation from 0.0 to 0.2 have been used in this study to simulate the Coulomb

friction effect. The friction coefficient function is determined by the direction of the sliding velocity as represented in (15). The variation in sliding velocity (16) can be shown in Fig. 4(a), which gives a variant square wave shape during the mesh period. For example, a constant friction coefficient ($\mu_0=0.02$) is represented in Fig. 4(b). It can be seen that, a significant changes in μ_0 give an effective simulation to the friction coefficient during the meshing process.

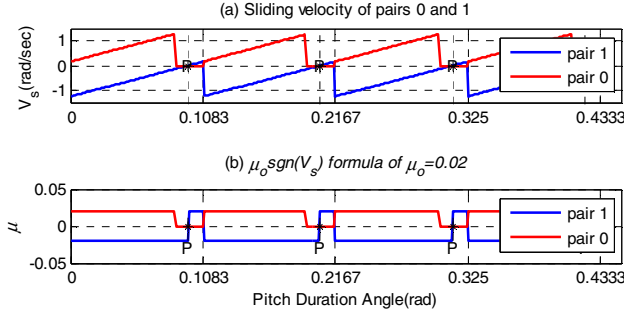


Figure 4 Sliding velocity and friction coefficient function of pair0 and pair1 during the mesh period

III. DYNAMIC MODEL AND SOLUTION METHOD

A. Gear Dynamic Model

To investigate friction influences, the model considered in this research is based on the one developed and subsequently modified by Kahraman [23] and Singh [24]. However, to represent gear transmission more accurately, the model also takes into account effects of speed-torque characteristics of motor driving systems. As shown in Fig. 5, the model is an 8-degree-of-freedom nonlinear model. The pinion and gear, denoted with subscripts 1 and 2 respectively have translational motions and rotational motions. As shown by the geometric specification in Table 1, the gear system is a speed increaser which is the same configuration as wind turbine applications. The pinion and gear are coupled by a spring having time varying mesh stiffness $K_m(t)$ and a varying mesh damping $C_m(t)$. The model includes four inertias, namely load, motor, pinion and gear. The torsional compliances of shafts and the transverse compliances of bearings combined with those of shafts are included in the model. The resilient elements of supports are described by stiffness and damping coefficients K_{x1} , K_{x2} , C_{x1} and C_{x2} for the pinion and gear respectively in the OLOA direction, besides K_{y1} , K_{y2} , C_{y1} and C_{y2} in the LOA direction. The shafts between the input motor, output loading motor and the gears are represented by torsional stiffness and torsional damping components k_1 , k_2 , c_1 and c_2 . Moreover, the model takes into account the influence of torque T_m and T_L as the driving torque and load torque respectively. The transverse vibrations of the gears are considered along LOA and off-line of action (OLOA).

The equation of motions are arranged into the state space formulation base on vibration analysis and then with MATLAB operation supported by ODE solver. The governing equations of motion for the model depicted in

Fig. 5 are written based on the following key assumptions:

- Pinion and gear are modelled as rigid disks;
- Applying input torque and applied load to the system;
- Shaft mass and inertia are lumped at the gears;
- Coulomb friction is assumed with a constant coefficient of friction μ_0 ;
- Manufacturing and assembly errors are ignored;
- Static transmission error effects are neglected;
- Backlash is not considered in this model.

Table 1 Geometric property of the meshing gears

Geometric Properties	Pinion	Gear
Number of teeth	$Z_p=58$	$Z_g=47$
Pitch radius (mm)	$r_p=40.08$	$r_g=32.48$
Mass (kg)	$m_p=0.86$	$m_g=0.68$
Rotation speed (rpm)	1485	1832.6
Pressure angle (°)	$\phi=20$	
Module (mm)	$m=1.38$	
Addendum (mm)	$a=1.4$	
Contact ratio	$\epsilon_{ratio}=1.7822$	
Motor torque (Nm)	$M_0=36$	
Applied torque (Nm)	$T_L=29.2$	

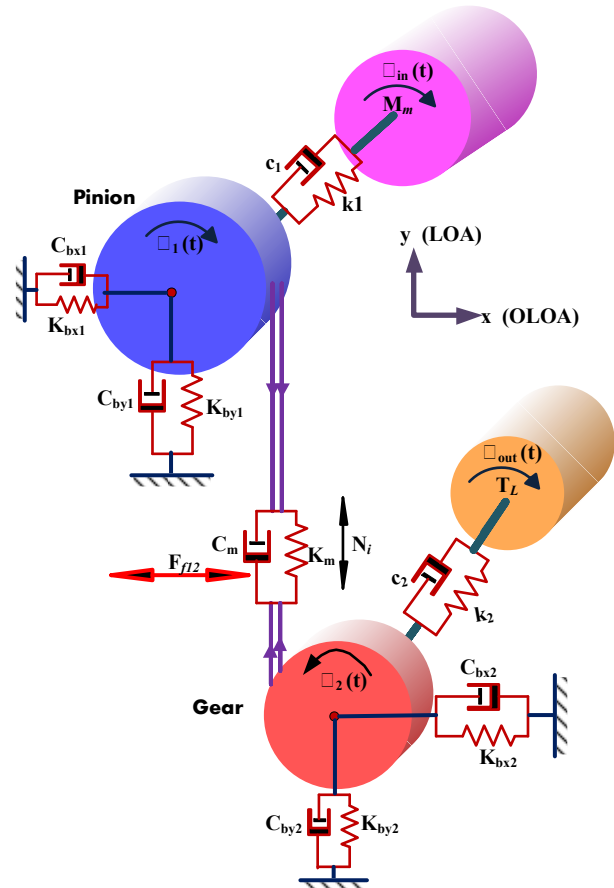


Figure 5 Schematic diagram of gear dynamic model with friction

According to the Newtonian law the equations of the motion are for the motor rotor, pinion rotation, gear rotation, Y-direction of pinion and gear translations, X direction of pinion and gear translations and load rotor respectively:

$$I_m \ddot{\theta}_{in} + c_1 (\dot{\theta}_{in} - \dot{\theta}_1) + k_1 (\theta_{in} - \theta_1) = M_m \quad (18)$$

$$I_p \ddot{\theta}_1 - c_1 (\dot{\theta}_{in} - \dot{\theta}_1) - k_1 (\theta_{in} - \theta_1) + r_p C_m (r_p \dot{\theta}_1 - r_g \dot{\theta}_2 + \dot{y}_p - \dot{y}_g) + r_p K_m (r_p \theta_1 - r_g \theta_2 + y_p - y_g) + F_{f12} \rho_p(t) = 0 \quad (19)$$

$$I_g \ddot{\theta}_2 + c_2 (\dot{\theta}_2 - \dot{\theta}_{out}) + k_1 (\theta_2 - \theta_{out}) - r_g C_m (r_p \dot{\theta}_1 - r_g \dot{\theta}_2 + \dot{y}_p - \dot{y}_g) - r_g K_m (r_p \theta_1 - r_g \theta_2 + y_p - y_g) - F_{f12} \rho_g(t) = 0 \quad (20)$$

$$I_L \ddot{\theta}_{out} - c_2 (\dot{\theta}_2 - \dot{\theta}_{out}) - k_1 (\theta_2 - \theta_{out}) = -T_L \quad (21)$$

$$m_p \ddot{y}_p + C_m (r_p \dot{\theta}_1 - r_g \dot{\theta}_2 + \dot{y}_p - \dot{y}_g) + K_m (r_p \theta_1 - r_g \theta_2 + y_p - y_g) + C_{by1} \dot{y}_p + K_{by1} y_p = 0 \quad (22)$$

$$m_g \ddot{y}_g - C_m (r_p \dot{\theta}_1 - r_g \dot{\theta}_2 + \dot{y}_p - \dot{y}_g) - K_m (r_p \theta_1 - r_g \theta_2 + y_p - y_g) + C_{by2} \dot{y}_g + K_{by2} y_g = 0 \quad (23)$$

$$m_p \ddot{x}_p + C_{bx1} \dot{x}_p + K_{bx1} x_p - F_{f12} = 0 \quad (24)$$

$$m_g \ddot{x}_g + C_{bx2} \dot{x}_g + K_{bx2} x_g + F_{f12} = 0 \quad (25)$$

$$M_m = M_m + 10(\omega_p - \dot{\theta}_1) \quad (26)$$

Equation (26) is used to adjust the motor input torque to maintain its speed as constant as possible. Especially, additional static torque is needed in order to balance the torque due to friction effects. This torque adaptation is to simulate the speed-torque characteristics for a common induction motor used widely. So that, a slight changes in the motor parameters will be predicted as it will explain later.

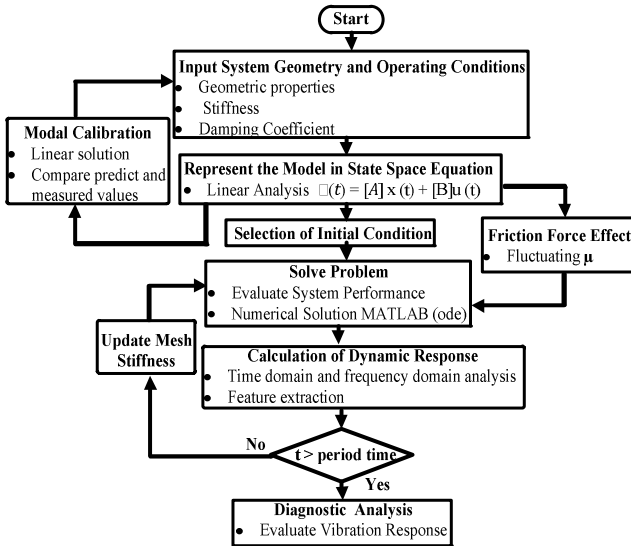


Figure 6 Simulation procedure used in this study

B. Solution Procedure

A numerical simulation study was performed to obtain the solution of the nonlinear equations. However, to ensure the correctness of parameters used and model structures, linear solutions was obtained when an average meshing stiffness value is used in the model without fiction influences, which allows the adjustment of the

model parameters so that major resonances agree with real system as close as possible. Subsequently, the non-linear effects of varying friction and mesh stiffness have been applied to the model and numerical integration method is used to solve the model. The difference of the gear vibration responses are examined between different friction coefficient values. More details of the simulation procedure used in this study are summarized in a flowchart shown in Fig. 6.

IV. MODAL CALIBRATION

A. Linear Solution

A simplified linear version of this model is developed by using the average mesh stiffness value in (19)-(23). It allows modal parameters including resonance frequencies and damping ratios to be found conveniently using the standard eigen method. By considering linear factors of the system, the vibration differential equation is expressed as:

$$[M]\{\ddot{q}\} + [C]\{\dot{q}\} + [K]\{q\} = f(t) \quad (27)$$

$$\{\dot{V}\} = [A]\{q\} \quad (28)$$

where, $[M]$ is mass matrix, $[C]$ is damping matrix, $[K]$ is stiffness matrix and q is vibration response vector consisting of displacements and velocity of the system. Using standard method for linear system analysis, the frequency response can be obtained conveniently under different parameters settings. Figure 7 shows the system responses with refined parameters. It can be seen that the 1st mode is at 128Hz which is 4 times away from the shaft frequency at about 25Hz. The third and fourth modes are close to the 2nd harmonic of $2 \times f_m = 2 \times f_r = 2 \times 1435.5 \text{ Hz}$.

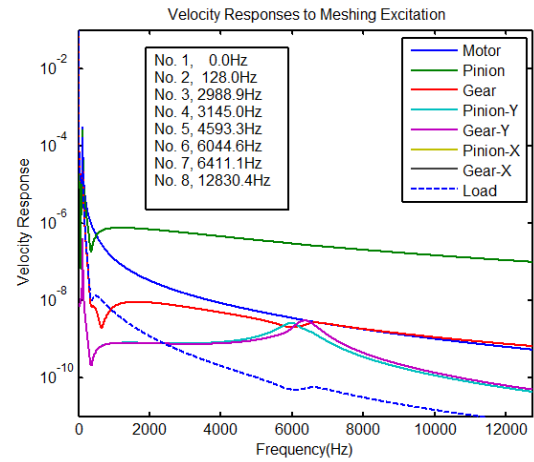


Figure 7 Frequency responses of gear system excited with impulsive inputs at the pinion and gear

To maintain the solution stability in the case of solving the nonlinear equations, these modes are applied with high damping ratios so that the frequency responses around these frequency ranges are relatively flat. Also note that there is no response in X-directions as there is no friction effect included in the linear mode. Moreover

the frequency responses are similar to that of measurements from the gearbox installed in the lab. It shows that the key parameters such as tooth stiffness values and damping ratios are used appropriately and numerical solutions can be proceeded to obtain the nonlinear responses.

B. Nonlinear Solution

The time domain behavior of the nonlinear system is obtained by integrating the set of governing differential equations numerically using an ode15s Runge–Kutta algorithm with a fixed time step size. This is suitable for solving differential algebraic stiff problems with high fluctuations and large noises in the solution. An appropriate set of initial conditions was applied to integrate the problem. The operating conditions of the system observed convergent responses corresponding to constant speed of interest. Figure 8 presents acceleration responses in the time domain and frequency domain for a case with friction included. In the time domain, all the responses including pinion and gears in rotations (θ_1 , θ_2), translations in the LOA (y_p , y_g) and OLOA (x_p , x_g) directions exhibit periodic profiles following stiffness changes, which is confirmed in the frequency domain in which the spectral peaks are observed at the gear mesh frequency $f_m = f_r Z = 1435.5 \text{ Hz}$ and its higher order harmonics. This spectral pattern is of typical for gear vibrations. However, because of the effect of resonances, the amplitudes at the higher order harmonics are higher than the fundamental one. For the same reasons, the rotational response of the pinion is higher than that of the gear, which is also seen in the frequency response characteristics.

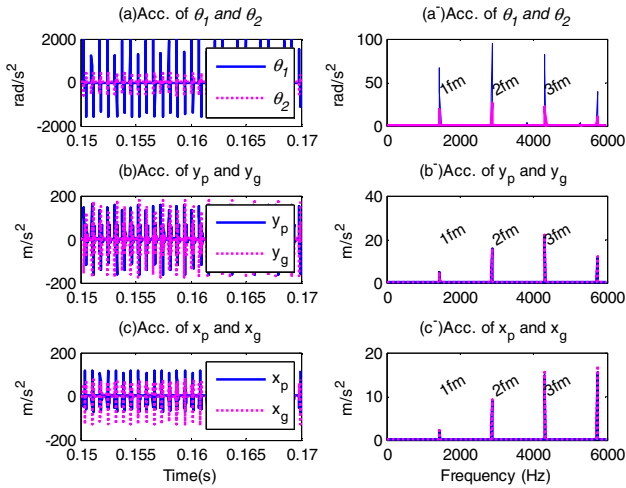


Figure 8 Vibration responses in the time domain and frequency domain

V. SIMULATION RESULTS AND DISCUSSION

A. Speed and Transmission Power

Having confirmed that the general solution of the system is close to reality, simulation studies were performed under a successive increment of friction coefficients μ_0 from 0 and 0.2 which is the range explored in previous studies. The operating conditions

were kept exactly the same for different values of coefficients. The load torque is $T_L = 29.2 \text{ Nm}$, which corresponds an input torque 36 Nm at the speed of 1485 rpm.

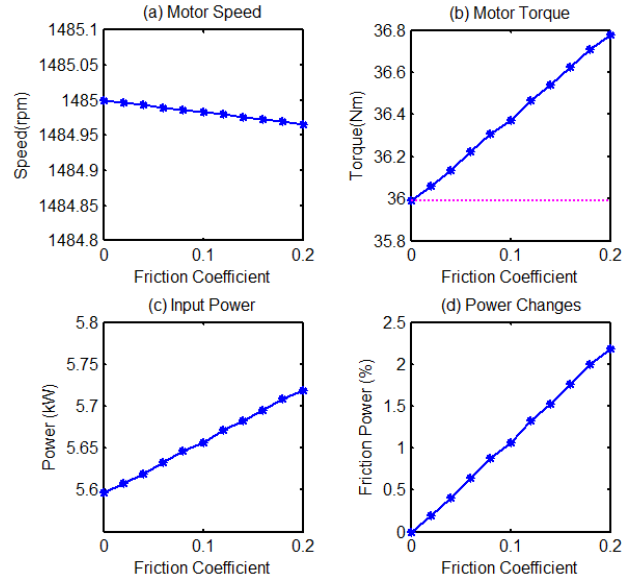


Figure 9 Effect of friction on motor operating parameters

Figure 9 shows the change of operating parameters with friction coefficient. It can be seen that there is a slight drop in the speed but a significant increase in the input torque. It means that with more friction effect, more input power is required to maintain the speed as close as to the setting point. However, because of the torque adaptation of (26) used, the speed has such a slight drop. Moreover, it is observed that there is a nearly linear increase in the motor power and the maximum change is 2.18%. It shows that it is clear that power measurement can be used for indicating lubrication degradation. These changes in operating conditions show that the model prediction is consistent with real operations and hence the vibration responses can be examined realistically.

B. Vibration Responses

Commonly, accelerations are measured for monitoring machine vibration characteristics. So the numerical solutions are converted into accelerations by differentiating the velocity responses. In addition to calculating the root mean squared (RMS) values for examining changes in overall vibration levels, spectral amplitudes at meshing frequencies are also extracted from the spectra of the acceleration responses in order to obtain a quantities assessment of frictional effect on default diagnostic features. As shown in Fig 10, RMS values for nearly all vibration signals show a monotonous increase, which is consistent with that of previous studies for noise reduction. However, because of the effect of nonlinearity, the response of the gear rotation exhibit quadratic nonlinear increase. In general, the vibration response increases with friction. Therefore, higher vibration level may indicate that the lubrication condition is poorer.

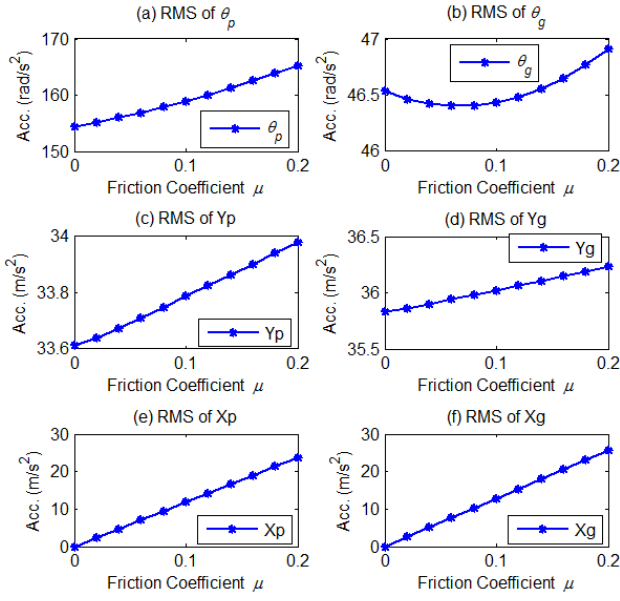


Figure 10 RMS of acceleration signals in rotation and translation transverse for pinion and gear

C. Vibration at Meshing Frequency

For more detailed and accurate friction diagnosis, the change of spectral amplitudes is usually indicating the gearbox conditions. Figure 11 presents the first three harmonic components of rotational responses for the meshing frequency. It can be seen that they behave diversely. The first and the third harmonics on the pinion show a nearly linear increase trend with friction, which can be based on the friction effect indicator. However, due to the nonlinear responses, the three components of the gear show inverse change and may not be so direct to be taken as good indicator for frictional influences.

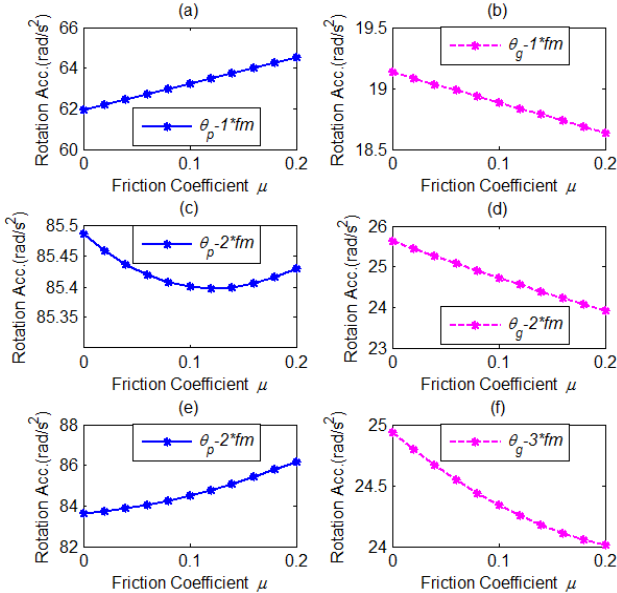


Figure 11 Rotation responses at mesh frequency with friction

In the same way the nonlinear response also cause the second and the third harmonic components of the translational responses in Y-direction to decrease with increasing in friction, as showing in Fig. 12. However,

the first harmonic increases with the friction coefficient and hence can be based on to indicate the change of friction due to lubricant degradation.

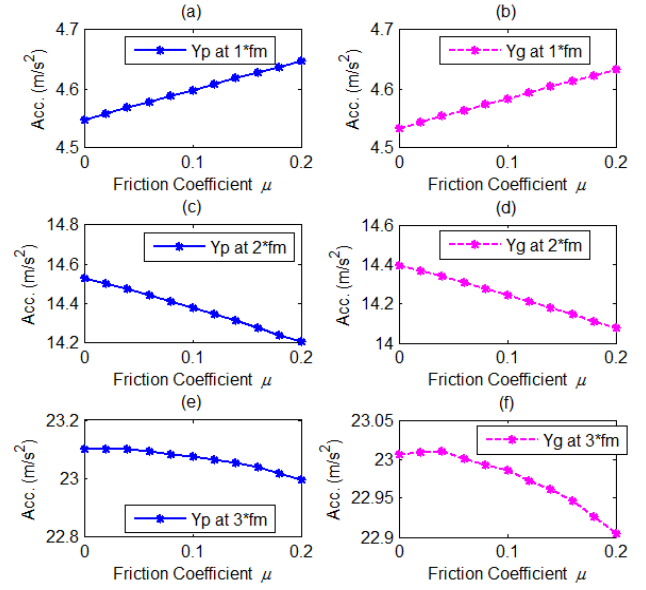


Figure 12 Spectral peaks of translation responses in Y-direction (LOA)

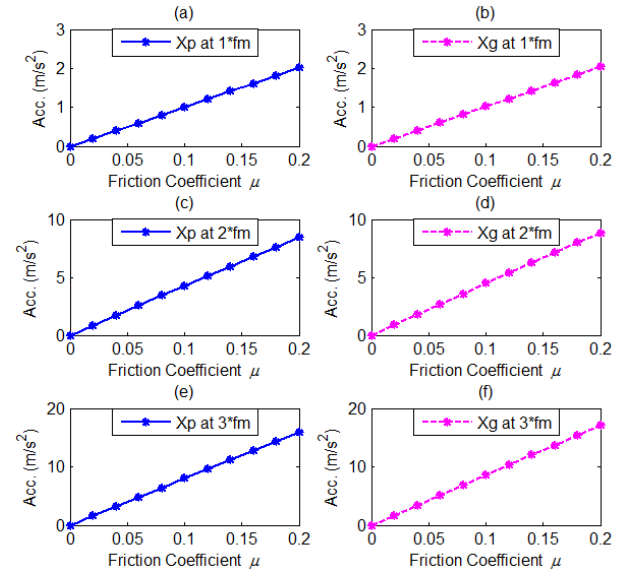


Figure 13 Spectral peaks of translation responses in X-direction (OLOA)

For the translation responses in X-direction, all harmonic components exhibit good increase trend that is proportional to the friction coefficient. Therefore, any of them can be used for lubrication condition monitoring. Moreover, the amplitude of increase is more significant, compared with the changes in the Y-direction. Therefore, the combination of the responses in two directions could result in an overall increase trend, which represents the real measurement values perceived by a sensor on the housing of a gearbox. Figure 14 is the combined responses obtained by $a_{xy} = \sqrt{a_x^2 + a_y^2}$ provided that the frequency response of housing is in linear range. As shown in the figure, the entire three component exhibit as

a monotonous increase with friction and it can be effective indicators for the friction. Moreover, as the change is tiny for the small friction coefficients, it means that vibration responses measured on the housing are relatively stable for good lubrication conditions. In other words, diagnostic features for other fault such as tooth breakages are also stable for obtaining a reliable severity diagnostic result. In the meanwhile, the diagnostic features will be further enlarged by poor lubrications, which is helpful to detect incipient tooth problems.

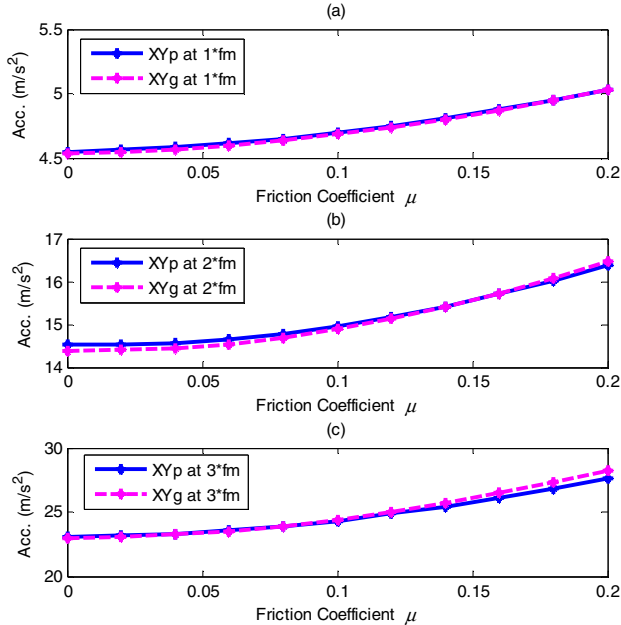


Figure 14 Spectral peaks of combined translation responses

In addition, the combined responses also show that the difference of the responses between the pinion and gear is very close, which means that the measurement at a position near either to the pinion or the gear will produce the same results for monitoring.

CONCLUSION

The dynamic model coupling with tooth friction produces consistent vibration responses to the change in friction due to lubrication degradation. It shows that there is an increase up to 2.18% in power consumption due to friction coefficient changes. However, the maximum increase of vibration responses of spectral peaks can be more than 100%. These show that it is much significant to use vibration responses to monitor the change in friction behavior. In the meantime, the power consumption may need a more accurate measurement system to resolve the small changes.

Both rotational responses and translational responses of vibration can be good indicators for lubrication conditions but the translational one is more sensitive even though the rotational responses are generally more nonlinear.

REFERENCES

[1] Randall, R., *A new method of modeling gear faults*. Journal of Mechanical Design, 1982. **104**(2): p. 259-267.

[2] Jardine, A.K., D. Lin, and D. Banjevic, *A review on machinery diagnostics and prognostics implementing condition-based maintenance*. Mechanical systems and signal processing, 2006. **20**(7): p. 1483-1510.
 [3] Begg, C.D., et al. *Dynamics modeling for mechanical fault diagnostics and prognostics*. in *Maintenance and Reliability Conf.* 1999.
 [4] Bruns, C.J., *Dynamic gearbox simulation for fault diagnostics using a torque transducer*. 2011.
 [5] Begg, C.D., C.S. Byington, and K.P. Maynard. *Dynamic simulation of mechanical fault transition*. in *Proceedings of the 54th Meeting of the Society for Machinery Failure Prevention Technology, Virginia Beach, VA.* 2000.
 [6] Bartelmus, W., *Mathematical modelling and computer simulations as an aid to gearbox diagnostics*. Mechanical Systems and Signal Processing, 2001. **15**(5): p. 855-871.
 [7] Van Khang, N., T.M. Cau, and N.P. Dien, *Modelling parametric vibration of gear-pair systems as a tool for aiding gear fault diagnosis*. technische mechanik, 2004. **24**: p. 3-4.
 [8] Diagnostics, R. and I. Gilon, *Gear Diagnostics—Fault Type Characteristics*.
 [9] Chaari, F., et al., *Effect of spalling or tooth breakage on gearmesh stiffness and dynamic response of a one-stage spur gear transmission*. European Journal of Mechanics-A/Solids, 2008. **27**(4): p. 691-705.
 [10] Jia, S. and I. Howard, *Comparison of localised spalling and crack damage from dynamic modelling of spur gear vibrations*. Mechanical Systems and Signal Processing, 2006. **20**(2): p. 332-349.
 [11] Lu, D., X. Gong, and W. Qiao. *Current-based diagnosis for gear tooth breaks in wind turbine gearboxes*. in *Energy Conversion Congress and Exposition (ECCE), 2012 IEEE*. 2012. IEEE.
 [12] Tian, Z., M.J. Zuo, and S. Wu, *Crack propagation assessment for spur gears using model-based analysis and simulation*. Journal of Intelligent Manufacturing, 2012. **23**(2): p. 239-253.
 [13] Wu, S., M.J. Zuo, and A. Parey, *Simulation of spur gear dynamics and estimation of fault growth*. Journal of Sound and Vibration, 2008. **317**(3): p. 608-624.
 [14] Chen, Z. and Y. Shao, *Dynamic simulation of spur gear with tooth root crack propagating along tooth width and crack depth*. Engineering Failure Analysis, 2011. **18**(8): p. 2149-2164.
 [15] Mohammed, O.D., M. Rantatalo, and J.-O. Aidanpää, *Dynamic modelling of a one-stage spur gear system and vibration-based tooth crack detection analysis*. Mechanical Systems and Signal Processing, 2015. **54**: p. 293-305.

- [16] Choy, F., et al., *Analysis of the Effects of Surface Pitting and Wear on the Vibrations of a Gear Transmission System*. 1994, DTIC Document.
- [17] Ding, H., *Dynamic Wear Models for Gear Systems*. 2007, The Ohio State University.
- [18] Ding, H., *A study of interactions between dynamic behavior of gear systems and surface wear*. 2007, The Ohio State University.
- [19] Flodin, A., *Wear of spur and helical gears*. Royal Institute of Technology, Stockholm, Doctoral Thesis, 2000.
- [20] Jiang, H., Y. Shao, and C.K. Mechefske, *Dynamic characteristics of helical gears under sliding friction with spalling defect*. Engineering Failure Analysis, 2014. **39**: p. 92-107.
- [21] Diab, Y., F. Ville, and P. Velex, *Investigations on power losses in high-speed gears*. Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology, 2006. **220**(3): p. 191-198.
- [22] Vaishya, M. and R. Singh, *Sliding friction-induced non-linearity and parametric effects in gear dynamics*. Journal of Sound and Vibration, 2001. **248**(4): p. 671-694.
- [23] Kahraman, A., J. Lim, and H. Ding. *A dynamic model of a spur gear pair with friction*. in *Proceedings of the 12th IFToMM World Congress*. 2007.
- [24] He, S., S. Cho, and R. Singh, *Prediction of dynamic friction forces in spur gears using alternate sliding friction formulations*. Journal of Sound and Vibration, 2008. **309**(3): p. 843-851.
- [25] Cheng-zhong, G. and C. Lie. *Effects of teeth surface friction on the vibration of gear transmission*. in *Mechanical and Electronics Engineering (ICMEE), 2010 2nd International Conference on*. 2010. IEEE.
- [26] Howard, I., S. Jia, and J. Wang, *The dynamic modelling of a spur gear in mesh including friction and a crack*. Mechanical systems and signal processing, 2001. **15**(5): p. 831-853.
- [27] Shing, T.-K., *Dynamics and control of geared servomechanisms with backlash and friction consideration*. 1994.
- [28] Kokare, D. and S. Patil, *Numerical Analysis of variation in mesh stiffness for Spur Gear Pair with Method of Phasing*. 2014.
- [29] Lin, J. and R.G. Parker, *Mesh stiffness variation instabilities in two-stage gear systems*. Journal of vibration and acoustics, 2002. **124**(1): p. 68-76.
- [30] Vaishya, M. and R. Singh, *Analysis of periodically varying gear mesh systems with Coulomb friction using Floquet theory*. Journal of Sound and Vibration, 2001. **243**(3): p. 525-545.
- [31] Vaishya, M. and R. Singh, *Strategies for modeling friction in gear dynamics*. Journal of Mechanical Design, 2003. **125**(2): p. 383-393.
- [32] Jia, S., I. Howard, and J. Wang, *The dynamic modeling of multiple pairs of spur gears in mesh, including friction and geometrical errors*. International Journal of Rotating Machinery, 2003. **9**(6): p. 437-442.
- [33] Xu, H., *Development of a generalized mechanical efficiency prediction methodology for gear pairs*. 2005, The Ohio State University.
- [34] Velex, P. and V. Cahouet, *Experimental and numerical investigations on the influence of tooth friction in spur and helical gear dynamics*. Journal of Mechanical Design, 2000. **122**(4): p. 515-522.
- [35] Rebbechi, B., F.B. Oswald, and D.P. Townsend, *Measurement of Gear Tooth Dynamic Friction*. 1996, DTIC Document.
- [36] Liu, G., *Nonlinear dynamics of multi-mesh gear systems*. 2007, The Ohio State University.
- [37] He, S., R. Gunda, and R. Singh, *Effect of sliding friction on the dynamics of spur gear pair with realistic time-varying stiffness*. Journal of Sound and Vibration, 2007. **301**(3): p. 927-949.

Characterisation of Acoustic Emissions for the Frictional Effect in Engines using Wavelets based Multi-resolution Analysis

Nasha Wei^{1,2}, Fengshou Gu^{1,3}, Tie Wang³, Guoxing Li³, Yuandong Xu³, Longjie Yang³, Andrew D. Ball¹

¹ Centre for Efficiency and Performance Engineering, University of Huddersfield, HD1 3DH, UK

² Department of Economics and Management, Taiyuan University of Science and Technology, Shanxi, P.R. China

³ Department of Vehicle Engineering, Taiyuan University of Technology, Shanxi, P.R. China

F.Gu@hud.ac.uk, Nasha.Wei@hud.ac.uk

Abstract—The friction between piston ring-cylinder liner is a major cause of energy losses in internal combustion engines. However, no experimental method is available to measure and analyze the frictional behavior. This paper focuses on the investigation of using acoustic emission (AE) to characterize the friction online. To separate the effect relating to friction sources, wavelets multi-resolution analysis is used to suppress interfering AE events due to valve impacts and combustion progress. Then a wavelet envelope indicator is developed to highlight AE contents from friction induced AE contents. The results show that the AE contents in the middle strokes correlate closely with viscous friction process as their amplitudes exhibit a continuous profile similar to piston speed. Furthermore, the AE envelope indicator proposed can distinguish the differences between two types of lubrication oils, showing superior performance of AE based online lubrication diagnosis.

Keywords- diesel engine; acoustic emission; piston-cylinder system; wavelet multi-resolution analysis; envelope analysis

I. INTRODUCTION

Diesel engines, which are primary components of the transportation system, have attracted much attention due to their wide usage and higher thermal efficiency. Previous research has shown that the piston-cylinder system, as a main working part of the engine, is a significant source of mechanical losses in internal combustion engines [1] [2]. Therefore, considering the reliability and economy of engine system, it is important to monitor the operation condition of the piston-cylinder system.

Acoustic emission (AE), as a very useful tool of non-intrusive test, has been used to monitor the engine condition. The high spatial and temporal fidelity of the AE signals acquired from engines in service make it possible to detect individual events and processes [3]. The frequency of AE signal is, usually in the range from 100 kHz to 1 MHz, higher than the frequency of vibration signal which is generally from 20Hz to 20 kHz. Consequently, the AE signal has a very high signal-to-noise ratio to avoid the influence of mechanical noise. Moreover, the AE sensors which could be placed on the surface of an engine body are less likely to suffer from the harsh environments (high pressure and high temperature) and none detrimental effects in the engine structure

compared to pressure and temperature sensors. Therefore, AE signal is effective to assess the engine conditions such as combustion quality, valve performance, wear degrees inside engines for earlier fault detection and diagnosis.

Previous research has shown that the obvious bursts of AE for engines are related to valve landing and combustion progress. References [4] and [5] investigated the suitability of acoustic emission (AE) technique for the condition monitoring of diesel engine valve faults. Sharkey et al. [6] developed an engine fault diagnosis approach for combustion process using acoustic emission sensors. These researches indicate that the obvious bursts of AE events are caused by valve impacts and combustion. Douglas [7] reported that the continuous AE possibly induced by the ring/liner interaction related to asperity contacts between the ring-pack and cylinder liner. And the AE activity was found to be proportional to piston speed and pressure, indicating that the boundary friction acting upon the oil-control ring was the most likely AE source. Mishra et al. [8] presented that the total friction is addition of contributions made through viscous shear and asperity interactions. They also reported that viscous friction contribute main losses during suction and exhaust strokes and also accounts for sizeable friction losses in power strokes. Therefore, the friction between piston ring and cylinder liner is composed of asperity friction and viscous friction that will generate AE events during four strokes. It shows that less attention has been paid to the enhancement of AE signal quality for more accurate monitoring the frictional effect of piston-ring and cylinder liner.

This paper reports a method to monitor the engine friction online based on characterization of AE signals. As AE signals have strong nonstationary contents and there are a number of distinctive AE sources in the engine, wavelets based multi-resolution analysis (WMRA) including denoising techniques is used to extract the weak AE events that are relating to friction process. The key WMRA parameters are tuned based on the behavior of viscous friction in engine operation cycle. Then the envelope of extracted signals is used to develop diagnostic parameters for monitoring engine lubrication condition.

II. ACOUSTIC EMISSION GENERATION OF PISTON ASSEMBLY

Originally, AE signals are sourced from the generation and propagation of cracking and fracture, slipping and

The project of this paper is supported by Natural and Scientific foundation of China.(NO. 51375326).

elastic plastic deformation in the point view of engineering material failures. By sharing similar mechanisms the major AE sources in an engine can be understood to be from three physical processes including friction and wear, rapid pressure oscillations and mechanical impacts.

Friction and wear processes are generated between two sliding surfaces caused by the synthetically effects of adhesion, deformation, material fracture, heat and chemical process. These processes of friction and wear generate high-frequency AE stress waves predominantly pseudo-continuous with superimposed burst emissions due to sporadic high-amplitude events such as single asperity fracture, particle interactions [9]. The excessive friction and wear on pistons, cylinder liners and piston rings are also sources of AE. Shuster et al. [10] demonstrated the usefulness of acoustic emission RMS measurements for studying the piston ring cylinder liner scuffing phenomenon based on the understanding that the AE can be induced by the friction between piston ring and cylinder liner segments. Douglas et al. [7] used AE RMS and AE energy RMS measurements to provide information pertaining to the interaction between piston rings and cylinder liners in a range of diesel engines. However, the original AE signals processing methods mentioned above were hard to extract the feature about tribology of piston and cylinder system to detect the type of the friction.

The RMS value of AE signal is well correlated with the pressure signal in the time and frequency domain [11]. Vibrations and deformations between contact surfaces are induced by alternating low and high pressures in the piston ring against the cylinder wall. The excessively high cylinder pressures also generate the high pressures between piston ring and cylinder wall, and cyclically varying sealing force exerted by the in-cylinder pressure on the piston rings was found to influence the resultant AE activities [7].

Because of the elastic plastic deformation and stain of metal material, the AE signals could be excited by the mechanical impacts between different engine components. In particular, the valve events from air intake exhaust and fuel injections can cause strong impacts. Hence, a number of acoustic emission based studies have been carried out to detect the faults of valve opening and closing impacts [4] [12] and needle impacts of injection [13] [14].

Based on these advancements in AE studies, AE from engines friction effect are much weaker and occurring throughout full engine cycle. On the other hand, AE from other two mechanisms are very strong but present only in certain crank angle positions. Based on these differences, it is likely to separate these sources to perform diagnostics for different purposes.

III. WAVELETS-BASED MULTIREOLUTION ANALYSIS

A. Wavelet Transform

Wavelet transforms are particularly effective to analyse signals that contain non-stationary phenomena [15] which is the typical feature of engine AE signals. It has been used for numerous studies in fault diagnostics

such as engines [16], bearings[17] [18], gears [19] [20] [21], and also has been applied to the feature extraction in acoustic emission study.

The continuous wavelet transform of a signal $f(t) \in L^2(R)$ ($t = 1, 2, \dots, N$) is defined as

$$CWT(a, b) = \int_R f(t) \frac{1}{\sqrt{a}} \psi^* \left(\frac{t-b}{a} \right) dt \quad (1)$$

where the mother wavelet function is $\psi_{ab}(t) = \frac{1}{\sqrt{a}} \psi \left(\frac{t-b}{a} \right)$ $a, b \in R; a \neq 0$.

in which a represents dilation index, b is translation index, and $\psi^*(t)$ represents the complex conjugate of function $\psi(t)$.

Similarly, discrete wavelet transform (DWT) is a discretization of the CWT which has a huge number of applications in science and engineering with the discretized scales and positions. The expression of DWT is defined as

$$Wf(j, k) = \int_R f(t) \frac{1}{\sqrt{2^j}} \psi^* \left(\frac{t}{2^j} - k \right) dt \quad (2)$$

The discrete mother wavelet function is

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \psi \left(\frac{t}{2^j} - k \right), \quad (3)$$

and the discrete father wavelet function is

$$\phi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \phi \left(\frac{t}{2^j} - k \right). \quad (4)$$

where parameters ' a ' and ' b ' are replaced by $a = 2^j$, $b_{j,k} = 2^j k$, $j, k \in Z$ respectively to represent the time shift indices $k = 1, 2, \dots, N/2^j$ and decomposition levels $j = 1, 2, \dots, J$. Consequently, the wavelet transform of $f(t)$ can be obtained by using (5) and (6) [22]:

$$a_{j,k} = \int f(t) \phi_{j,k}(t) dt \quad (5)$$

$$d_{j,k} = \int f(t) \psi_{j,k}(t) dt \quad (6)$$

and $a_{j,k}$ is the the approximation coefficients reflecting the low frequency content of the sigand whereas $d_{j,k}$ is detailed coefficients reflecting the high frequency content.

B. Wavelet Multi-Resolution Analysis

The wavelet multi-resolution analysis (WMRA) implements discrete wavelet transforms with filters and decomposes the signal into several sub-signals at different levels of resolution [23]. A mathematical model of WMRA and its practical application were introduced for the first time by Mallat [24]. A signal can be represented by WMRA in terms of the scaling and the wavelet functions mathematically presented as: [22] [25]

$$f(t) \approx \sum_k a_{j,k} \phi_{j,k}(t) + \sum_{j=0}^{J-1} \sum_k d_{j,k} \psi_{j,k}(t). \quad (7)$$

Consequently, $f(t)$ can be reconstructed by the approximation signal $A_j(t)$ and the detail signals $D_j(t)$ at each level j which are expressed by (8) and (9) respectively:

$$A_j(t) = \sum_k a_{j,k} \phi_{j,k}(t), \quad J, k \in Z \quad (8)$$

$$D_j(t) = \sum_k \sum_{j=0}^{J-1} d_{j,k} \psi_{j,k}(t), \quad j, k \in Z \quad (9)$$

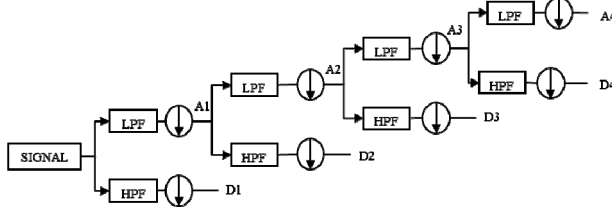


Figure 1. WMRA decomposition with 4 levels

For ease of understanding WMRA, a decomposition of a signal into four levels is illustrated in Fig. 1. The WMRA decomposes the original signal $f(t)$ with a low-pass filter (LPF) and a high-pass filter (HPF) in every decomposition level. At level j , $D_j(t)$ is the detail coefficients produced by the HPF, and $A_j(t)$ is the approximation coefficients produced by the low pass filter LPF. Therefore, the merits of WMRA are its ability to produce a good time resolution at high frequencies and good frequency resolution at low frequencies. Moreover, as the samples of wavelet coefficients at high levels are much smaller than the original signal, further analysis on them will be more efficient. This merit is very useful for processing AE signals acquired at MHz ranges and huge number samples.

However, the wavelet types and decomposing levels are usually needed to be selected appropriately in order to achieve high performances in its typical applications such as signal denoising and data compression. In this study, they are determined based on the friction behavior of engines which will be depicted in section 5.

IV. THE FACILITY FOR EXPERIMENTAL STUDIES AND METHODOLOGY

To investigate the AE based friction diagnosis, a single cylinder diesel engine was used for experimental studies. It has less AE events, compared with a multiple cylinder engine and it allows more accurate charactering the weak AE signals from frictions. Photos in Figures 1 and 2 show the basic structure of the engine, engine test bed and location of AE sensor. In addition, Table I provides the key specification.

A SR800 AE sensor from Soundwel Technology Co., Ltd is used for AE measurement, which has a flat frequency response over a range of 50 - 800 kHz. The AE sensor was located on the engine cylinder body surface by a magnetic hold-down as shown in Fig. 3. The sensor position is closer to the friction source between the ring and cylinder liner but relatively farer from other sources

including valve impacts and combustion. The output voltage of an AE sensor is usually very low. So a preamplifier is used to amplify the signal so that it can be acquired adequately by a SEAU2S two channel AE measurement system. Simultaneously, a crank encoder signal is also acquired by this data acquisition system in order to mark the top dead centre (TDC) which can be used to align the AE signal with crank angle signal and to obtain the speed of the engine for performing angular domain based analysis.

TABLE I. SPECIFICATION OF THE TEST ENGINE

Manufacturer	Anhui Quanchai Engine Co., Ltd., PR. China
Engine type	QCH1110II
Number of cylinders	one
Combustion system	Direct injection, vertical type
Bore/stroke	110/115 mm
Displacement volume	1.093 L
Compression ratio	18:1
Rated power	14.7/2400 kW/r/min
Max. torque	67/1920 Nm/r/min



Figure 2. The photo of the diesel engine test rig

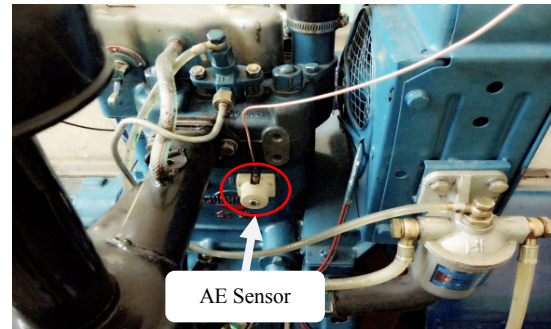


Figure 3. The installation of AE sensor on the engine body

To examine the effect of engine operation and lubrication oil conditions, two types of engine lubricant oil were tested when the engine operated under the conditions shown in in Table II. More speed setpoints were tested as the friction effect is more influenced by velocity variation than by the load. Table III shows the decisive parameter; viscosity values of the two types of oils at two different temperatures, which were measured before the engine tests. Note that the measurement instrument produced high diversity results at temperature 40°, showing that it is difficult to obtain reliable oil property measurement.

The AE signal measurements for each type of lubricating oils were repeated twice for all the operating

conditions. More than 8 MEG data points were obtained at a sampling rate of 800kHz for each AE measurement, which covers over 100 engine cycles to be averaged for obtaining reliable results.

TABLE II. ENGINE OPERATING CONDITIONS

Engine speed (rpm)	Load (Nm)	Running time (min.)
Warm-up running (1000)	0	20~30
1000	10/40	5/5
1200	10/40	5/5
1400	10/40	5/5
1600	10/40	5/5
1800	10/40	5/5

TABLE III. VISCOSITY-TEMPERATURE COMPARISON OF DIFFERENT LUBRICATING OILS

Grade	Viscosity(mPa·s)			
	Test 1		Test 2	
	40 °C	100 °C	40 °C	100 °C
10W30	35.54	3.53	76.46	9.93
20W50	174.2	20.27	165.1	18.33

V. WMRA ANALYSIS AND RESULTS

A. WMRA Analysis

The measured AE signals in the timed domain are firstly converted into the angular domain based on the TDC marker signals. Thus, it is easier to identify various AE events in association with the engine operation process. Fig. 4 shows a typical AE signal in the angular domain. It can be seen that for one working cycle of the four stroke engine there are a number of significant AE bursts. Based on the engine operation process, it is straightforward to identify the events that correspond sequentially to inlet valve opening (IVO), exhaust valve closing (EVC), inlet valve closing (IVC), fuel injection and combustion. However, it is difficult to see the exhaust valve opening due to that the valve is a bit far away to the AE sensor. Therefore, there are many other AE contents cannot be identified unquestionably. Particularly, the events assumed to be cylinder friction closing are very unconvinced because they show little characteristics relating the friction process.

To ensure that the AE signal contains friction information, the angular domain signals are decomposed onto successive wavelets levels. There are 144000 data points. The maximum level J could be 17. However, by a trial and error approach with using different types of wavelets including, Daubechies wavelets and Symlets wavelets families, it has found that Symlets of order 9 wavelet at level $J=4$ is sufficient so as to enhance the AE signal to show the friction effects in that the viscous friction exhibit higher amplitudes in the middle of the each stroke, which has been investigated in [5, 6, 23] about friction effects between piston ring and cylinder liner. In general, the continuous characteristics of the AE signals indicated by the viscous friction agree with the piston velocity profile and load variation.

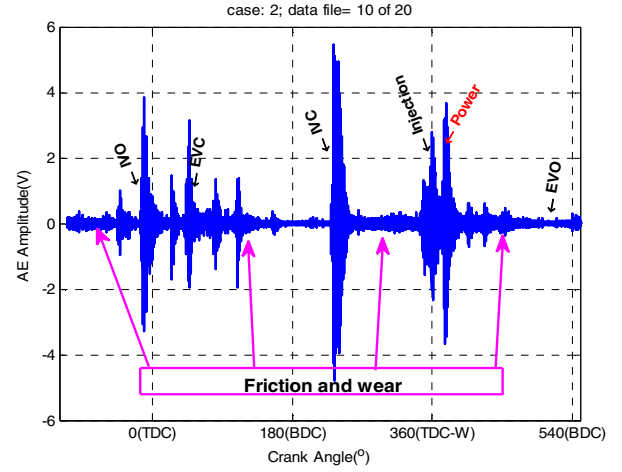
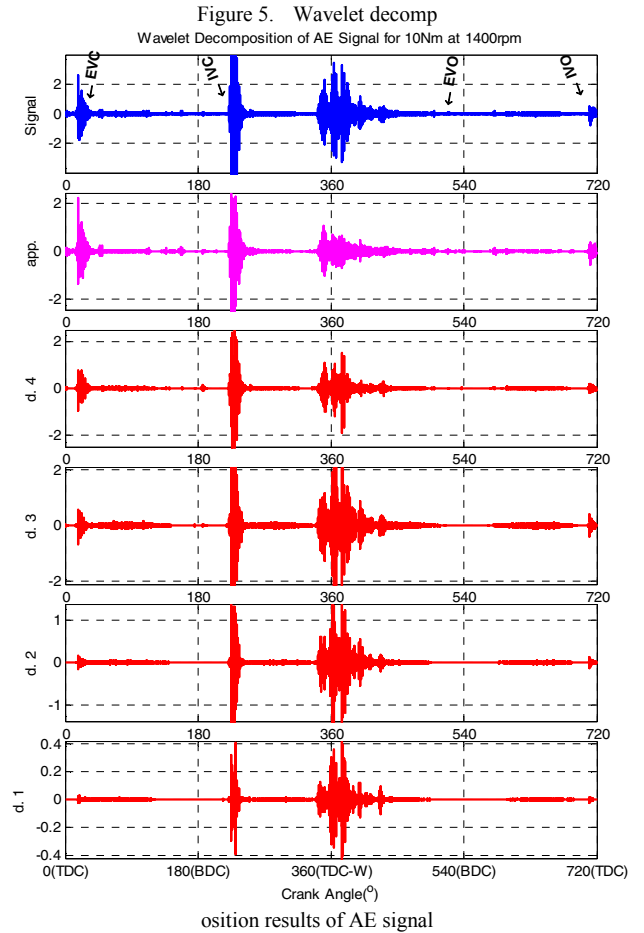


Figure 4. Raw AE signals from the engine cylinder body



As shown in Fig. 5, the decomposition with all detailed coefficients shows an effective enhancement of the friction characteristics. Particularly, d_3 and d_4 based decompositions produce higher amplitudes in the middle of engine strokes, compared with that of d_1 and d_2 , which is more consistent with that of viscosity induced friction. Therefore, the decomposed signal from d_3 and d_4 are taken as the candidates for developing quantitative diagnostic features.

B. Wavelet based Envelope

However, because of the wide band characteristics, the AE events due to valve impacts, fuel inject and combustion spread across all different levels, which may influence the accuracy of quantitative features. Therefore, a wavelet denoise approach is used to suppress these dominant events. It is realized by applying a hard threshold of 0.25 to wavelets coefficients at to exclude the wavelet coefficients due to the dominant events and then reconstruct the signal with d4 coefficient only as it less influenced by these large events. Furthermore, the envelope of reconstructed signals are calculated for each engine cycle and an average is performed over different envelope signals over different cycles, which allows the random variation of viscous friction to be aggregated more effectively, compared with that of averaging the decomposed signals directly.

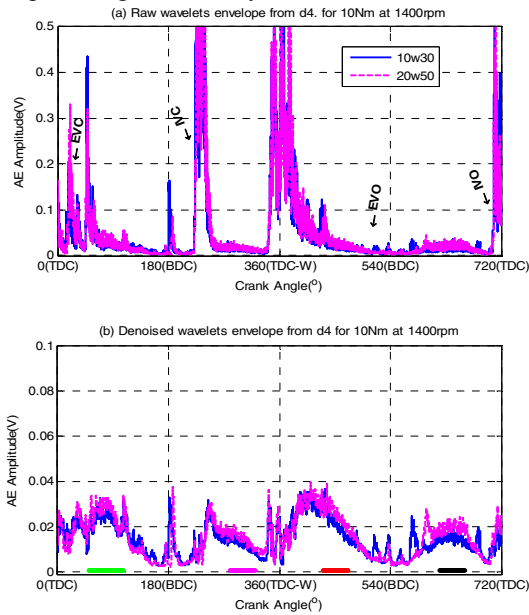


Figure 6. Envelope signals of decomposed signal with d4 coefficient

Fig. 6 presents typical envelope signals with and without denoise processing. It can be seen that the denoised envelope signal allows a clearer representation of the frictional characteristics. Especially, the AE amplitudes are very distinctive in the middle of each stroke, which allows the frictional effect to be examined more accurately. As shown in Fig. 6(b), the oil with higher viscosity produces higher AE amplitudes, which agrees with theoretical prediction in that higher viscosity causes more viscous friction. Therefore, it is possible to use AE for indicating the engine lubrication condition.

C. Results and Discussion

Based on above analysis, four envelope indicators can be obtained by averaging amplitudes with respective to the four angular ranges illustrated by the bold horizontal lines in Fig. 6(b) to represent lubrication condition in corresponding strokes. As these four amplitudes are calculated around the mid stroke where the AE envelope has high signal-to-noise ratio, they can best reflect the viscous friction characteristics under different operating conditions.

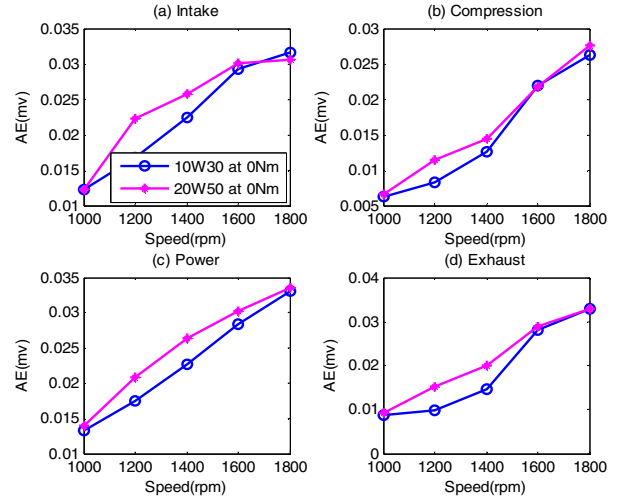


Figure 7. Average AE envelope indicator for no load operating.

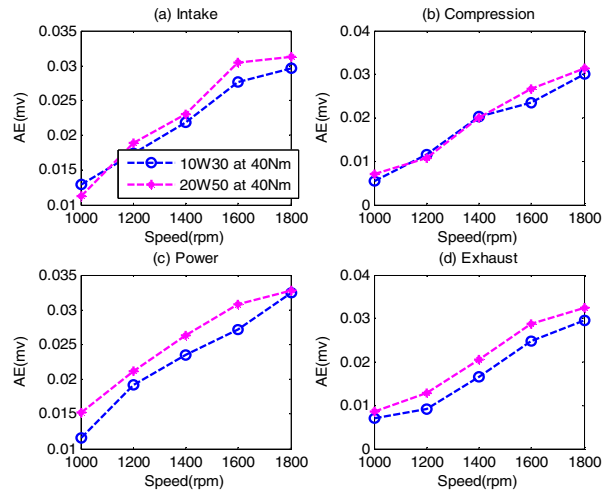


Figure 8. Average AE envelope indicator for higher load operation

Fig. 7 shows the AE envelope indicator at different speeds for the load free operation. It shows that AE indicator increases with speed, showing the strong relation with friction process. Comparing the indicators between the two types of oils, the AE indicator for 20W50 oil is higher for most speeds. Especially for the speeds of 1200 and 1400rpm, the difference in AE indicators is very clear for all strokes and it is definite to make difference between the two oils. However, the difference for other speed operations is not so clear because of high AE noise resulting from clearance induced instability.

When engine operates with the loads the clearance instability is smaller. The AE indicators from power and exhaust strokes can show clear differences between the two oils for all the speeds tested. Especially, the fraction induced AE in the exhaust stroke is much less influenced by other AE events of valve impacts and combustion progress. Therefore, the result in the exhaust stroke is reliable. On the other hand, the influence of valve impacts in the inlet and compressions strokes becomes higher under the high loads where the AE indicators cannot make good difference between the two oils. In addition, AE indicators for the high load operation is slightly

higher than that of load-free operation, showing that the fraction induced AE is less connected to the engine load.

VI. CONCLUSIONS

The study shows that wavelets-based multiresolution analysis can highlight the local nonstationary contents in engine AE signals for effective suppression. In meantime, it is also efficient for implementation further analysis as the discrete wavelets coefficient signal is much smaller and hence easier to be processed.

Based on the effective analysis, it is obtained that the fraction induced AE content can be extracted from the AE signals which contaminated by strong AE events including valve impacts, combustion progress and fuel injection excitation. The results show that the AE envelope indicator can reflect the connections between AE and engine friction under different speeds and loads. Especially, they can make good difference between two types of engine oils even they have little difference affecting engine performance. This shows that AE has higher performances in diagnosing engine lubrication conditions. Nevertheless, the research needs to be advanced more in optimizing WMRA symmetrically in order to find a better AE indicator.

REFERENCES

- [1] S. C. Tung and M. L. McMillan, "Automotive tribology overview of current advances and challenges for the future," *Tribol. Int.*, vol. 37, no. 7, pp. 517–536, 2004.
- [2] G. Smedley, "Piston ring design for reduced friction in modern internal combustion engines," PhD Thesis, Massachusetts Institute of Technology, 2004.
- [3] J. A. Steel and R. L. Reuben, "Recent developments in monitoring of engines using acoustic emission," *J. Strain Anal. Eng. Des.*, vol. 40, no. 1, pp. 45–57, Jan. 2005.
- [4] F. Elamin, Y. Fan, F. Gu, and A. Ball, "Diesel Engine Valve Clearance Detection Using Acoustic Emission," *Adv. Mech. Eng.*, vol. 2010, Jun. 2010.
- [5] T. L. Fog, L. K. Hansen, J. Larsen, H. S. Hansen, L. B. Madsen, P. Sorensen, E. R. Hansen, and P. S. Pedersen, "On condition monitoring of exhaust valves in marine diesel engines," *Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, 1999, pp. 554–563.
- [6] A. J. C. Sharkey, G. O. Chandroth, and N. E. Sharkey, "Acoustic emission, cylinder pressure and vibration: a multisensor approach to robust fault diagnosis," presented at the IJCNN 2000, *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, 2000, 2000, vol. 6, pp. 223–228 vol.6.
- [7] R. M. Douglas, J. A. Steel, and R. L. Reuben, "A study of the tribological behaviour of piston ring/cylinder liner interaction in diesel engines using acoustic emission," *Tribol. Int.*, vol. 39, no. 12, pp. 1634–1642, 2006.
- [8] P. C. Mishra, H. Rahnejat, and P. D. King, "Tribology of the ring—bore conjunction subject to a mixed regime of lubrication," *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.*, vol. 223, no. 4, pp. 987–998, Apr. 2009.
- [9] H. S. Benabdallah and D. A. Aguilar, "Acoustic Emission and Its Relationship with Friction and Wear for Sliding Contact," *Tribol. Trans.*, vol. 51, no. 6, pp. 738–747, 2008.
- [10] M. Shuster, D. Combs, K. Karrip, and D. Burke, "Piston Ring Cylinder Liner Scuffing Phenomenon Studies Using Acoustic Emission Technique," *SAE International*, Warrendale, PA, SAE Technical Paper 2000-01-1782, Jun. 2000.
- [11] M. El-Ghamry, J. A. Steel, R. L. Reuben, and T. L. Fog, "Indirect measurement of cylinder pressure from diesel engines using acoustic emission," *Mech. Syst. Signal Process.*, vol. 19, no. 4, pp. 751–765, Jul. 2005.
- [12] F. Elamin, Y. Fan, F. Gu, and A. Ball, "Detection of diesel engine valve clearance by acoustic emission," in *Proceedings of Computing and Engineering Annual Researchers' Conference 2009: CEARC'09*, G. Lucas and Z. Xu, Eds. Huddersfield: University of Huddersfield, 2009, pp. 7–13.
- [13] F. Elamin, Y. Fan, F. Gu, and A. Ball, "Detection of Diesel Engine Injector Faults Using Acoustic Emissions," presented at the COMADEM 2010: *Advances in Maintenance and Condition Diagnosis Technologies towards Sustainable Society*, Nara, Japan, Jun-2010.
- [14] A. Albarbar, F. Gu, and A. D. Ball, "Diesel engine fuel injection monitoring using acoustic measurements and independent component analysis," *Measurement*, vol. 43, no. 10, pp. 1376–1386, 2010.
- [15] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 961–1005, Sep. 1990.
- [16] J.-D. Wu and C.-H. Liu, "Investigation of engine fault diagnosis using discrete wavelet transform and neural network," *Expert Syst. Appl.*, vol. 35, no. 3, pp. 1200–1213, Oct. 2008.
- [17] S. Seker and E. Ayaz, "Feature extraction related to bearing damage in electric motors by wavelet analysis," *J. Frankl. Inst.*, vol. 340, no. 2, pp. 125–134, Mar. 2003.
- [18] R. Rubini and U. Meneghetti, "Application of the Envelope and Wavelet Transform Analyses for the Diagnosis of Incipient Faults in Ball Bearings," *Mech. Syst. Signal Process.*, vol. 15, no. 2, pp. 287–302, 2001.
- [19] C. Kar and A. R. Mohanty, "Monitoring gear vibrations through motor current signature analysis and wavelet transform," *Mech. Syst. Signal Process.*, vol. 20, no. 1, pp. 158–187, Jan. 2006.
- [20] W. J. Staszewski and G. R. Tomlinson, "Application of the wavelet transform to fault detection in a spur gear," *Mech. Syst. Signal Process.*, vol. 8, no. 3, pp. 289–307, 1994.
- [21] N. Baydar and A. Ball, "Detection of Gear Failures via Vibration and Acoustic Signals using Wavelet Transform," *Mech. Syst. Signal Process.*, vol. 17, no. 4, pp. 787–804, 2003.
- [22] T.-Y. Yang and L. Leu, "Study of transition velocities from bubbling to turbulent fluidization by statistic and wavelet multi-resolution analysis on absolute pressure fluctuations," *Chem. Eng. Sci.*, vol. 63, no. 7, pp. 1950–1970, Apr. 2008.
- [23] A. Datta, C. Mavroidis, J. Krishnasamy, and M. Hosek, "Neural Netowrk Based Fault Diagnostics of Industrial Robots using Wavelt Multi-Resolution Analysis," in *American Control Conference*, 2007. ACC '07, 2007, pp. 1858–1863.
- [24] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.
- [25] A. M. Gaouda, E. F. El-Saadany, M. M. A. Salama, V. K. Sood, and A. Y. Chikhani, "Monitoring HVDC systems using wavelet multi-resolution analysis," *IEEE Trans. Power Syst.*, vol. 16, no. 4, pp. 662–670, Nov. 2001.

Implementation of an Advanced Heating Control System at the University Academic Building

Vildan V. Abdullin, Dmitry A. Shnayder

Dept. of Automatics and Control
South Ural State University
Chelyabinsk, Russia
email: vildan@ait.susu.ac.ru

Abstract—This paper proposes an implementation of advanced heating feed-forward control system for multi-storey buildings based on indoor air temperature inverse dynamics model. The suggested model structure enables real-time assessment of the impact of unmeasurable perturbing factors on indoor air temperature, along with distinguishing between fast and slow processes occurring within the system. The paper also contains the actual results obtained by deployment of the suggested heating control system in the academic building of South Ural State University. The results obtained demonstrate the reducing of overall energy consumption by building heating system, at the same time increasing the comfort level of the building.

Keywords—building heating system; indoor air temperature; inverse dynamics model; feed-forward control

I. INTRODUCTION (HEADING 1)

In the last few decades, the mankind has been seeking to reduce consumption of all the known energy resources. In most countries and regions, including Northern Europe, Russia, Canada, New England, etc., energy utilized in the cold season for heating purposes makes up an essential portion of resources consumed. In addition to employing renewable energy sources, effective energy consumption based on a controlled process is one of the key trends in energy saving [1]–[3]. An important aspect here is the use of advanced control algorithms to secure optimal energy consumption by the heating system, while preserving a comfortable environment inside a building [4]–[5].

The quality of the heating process depends on the ability to maintain indoor air temperature stabilized at a preset comfort level. The indoor air temperature T_{ind} of a building mainly depends on its volume and building envelope type as process variables, the quantity of applied thermal energy Q_h as control signal, and the outdoor air temperature T_{out} as a key perturbing factor. Baseline control principle frequently employed in real-life situations provides control of the indoor temperature by reference to the key perturbing factor, i.e. outdoor temperature T_{out} . This approach appears viable, as it guarantees adequate quality, while implementing simple control algorithms and using data that is easy to measure. However, it does not account for minor perturbing factors such as solar radiation J_{rad} , wind V_{wind} , internal heat release Q_{int} , and the building's accumulated internal thermal energy Q_{acc} , which also have a significant impact on the indoor air temperature T_{ind} (Fig. 1), being also hard to measure or evaluate in real terms. As a result, indoor air temperature is maintained with poor precision because of

static error which occurs in the daytime under the influence of the above factors.

There is also a different approach that implements control loop with negative feedback by indoor air temperature T_{ind} , but there are certain challenges that make its practical implementation rather problematic [6]. The main challenge is that the building heating system is highly inertial. Still, a number of perturbances directly affect T_{ind} with minimum inertia. Therefore, a controller configured to adjust the heat exchange processes gently will not compensate for these perturbances, which results in significant fluctuations of indoor air temperature.

The above challenges require development of an advanced algorithm for energy-efficient heating system control process that would evaluate and compensate for the impact of perturbing factors in real time.

II. CONTROL METHOD

To take into account the unmeasured factors that affect temperature $T_{ind}(t)$, we referred to the approach based on the concept of generalized temperature perturbation $T_z(t)$ [7] characterizing the effect of the factors mentioned above on the indoor air temperature.

Generic structure of the proposed feed-forward control system is described in Fig. 2. As we see in the figure,

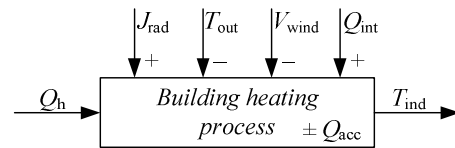


Figure 1. Factors affecting the indoor air temperature.

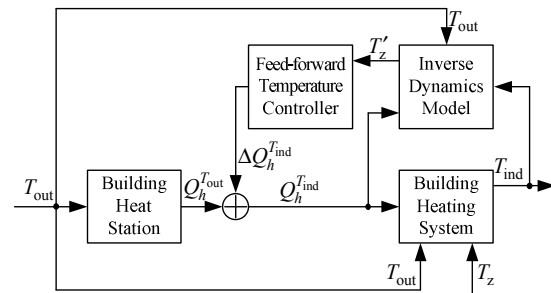


Figure 2. Block diagram of the building heating feed-forward control system based on inverse dynamics model.

baseline control of heat supply for building heating purposes follows a standard pattern with the use of automated building heat station that controls heating power $Q_h^{T_{out}}$ depending on the key perturbing factor – outdoor air temperature. The structure shown in Fig. 1 is augmented by a feed-forward control loop used to adjust $Q_h^{T_{out}}$ depending on estimated value T_z of generalized temperature perturbation T_z . Thus, the adjusted heating power value $Q_h^{T_{ind}}$ fed into the building is calculated as follows:

$$Q_h^{T_{ind}} = Q_h^{T_{out}} + \Delta Q_h^{T_{ind}}, \quad (1)$$

where $\Delta Q_h^{T_{ind}}$ stands for the adjusting value of heating power produced by feed-forward temperature controller [6].

According to Fig. 1 and energy conservation law, T_{ind} is constant provided that:

$$Q_h(t) + \sum_i [Q_{PF_i}(t)] = 0, \quad (2)$$

where $Q_h(t)$ stands for heating power applied to the heating system from the heating radiator, and $Q_{PF_i}(t)$ represents thermal energy flux reflecting the impact of i -th perturbing factor.

Relying on the concept of generalized temperature perturbation, we may put the thermal balance equation as follows:

$$T'_{ind}(t) = Q_h(t) / (q_h \cdot V) + T_{out}(t) - T_z(t), \quad (3)$$

where $T_{ind}(t)$ stands for the predicted value of indoor air temperature (the prediction horizon is determined by the fluctuation of indoor air temperature as a result of the fluctuation of indoor air temperature as a result of the perturbing factors (Fig. 1)); $T_{out}(t)$ is outdoor air temperature; q_h represents specific heat loss (per cubic meter); and V stands for external volume of the building [7].

Various perturbing factors affect the building at different rates [8]. The research showed that the best approach is to distinguish between *fast* and *slow* processes occurring within the system. Slow processes include the impacts of factors isolated from the indoor air temperature by the building's envelope, e.g. the key perturbing factor T_{out} and other perturbing factors like V_{wind} and Q_{acc} . The impact of these factors is characterized by long time response. Fast processes include the impacts of perturbing factors that have a direct effect on the indoor air temperature (J_{rad} , Q_{int}), as well as the effect of control signal Q_h . As a result, the building thermal performance inverse dynamics model can be described by the equation below:

$$T_{ind}(t) = F_{LS} \{ T_{out}(t) \} + F_{HS} \{ Q_h(t) / (q_h \cdot V) - T_z(t) \}, \quad (4)$$

where $F_{LS}\{\bullet\}$ is the dynamic operator of “slow” processes, and $F_{HS}\{\bullet\}$ is the dynamic operator of “fast” processes. A block diagram of building thermal performance dynamics

model composed in accordance with (4) is presented in Fig. 3.

According to the model shown in Fig. 3, the generalized temperature perturbation can be estimated by building an inverse dynamics model as follows:

$$T_z(t) = Q_h(t) / (q_h \cdot V) - F_{HS}^{-1} \{ F_{LS} \{ T_{out}(t) \} - T_{ind}(t) \}, \quad (5)$$

where $F_{LS}^{-1}\{\bullet\}$ is the inverse dynamics operator of “slow” processes calculated using the exponential filtration method [7], [9]. The signals Q_h , T_{out} , and T_{ind} are easy to meter in real terms. Assuming that $T_{ind}(t)$ is a statistically unbiased signal and that $T_z(t)$ is a signal with a mean of zero, the specific heat loss through the building envelope can be calculated from (5) by the equation below:

$$q_h = M_t \{ Q_h(t) \} / (V \cdot (M_t \{ T_{out}(t) \} - M_t \{ T_{ind}(t) \})), \quad (6)$$

where $M_t\{\bullet\}$ is the time-mean operator. [7]

The corresponding block diagram of the building thermal performance inverse dynamics model is presented in Fig. 4.

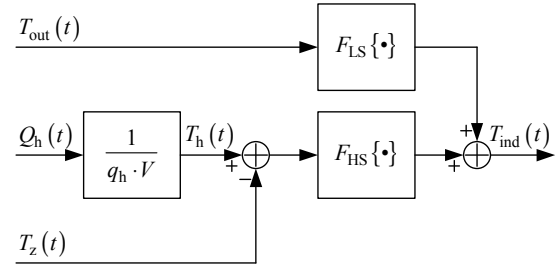


Figure 3. Block diagram of the building thermal performance dynamics model.

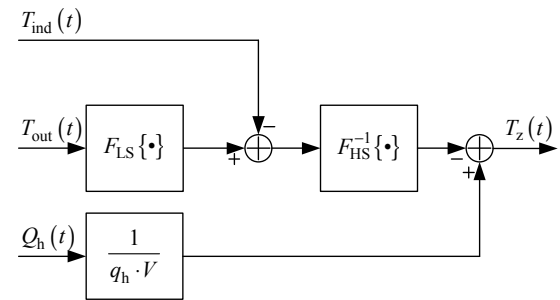


Figure 4. Block diagram of the building thermal performance inverse dynamics model.

III. IMPLEMENTATION RESULTS

Proposed advanced heating control algorithm was implemented in the 10-storey academic building of South Ural State University. Thermal energy is supplied to the building heating system by the automated building heat station. Both baseline and advanced control algorithms, as well as the building dynamics model, are implemented in the programmable logic controller of the automated

building heat station. To measure temperature in different premises of the building, we deployed a distributed network of 24 digital temperature sensors *Dallas DS18B20* in different rooms of the academic building linked together in *MicroLan (1-wire)* network [10]. In addition, we tested wireless data transfer from sensors using *RFM XDM2510HP* embedded communication modules integrated into a *WirelessHART* wireless sensor network. This technology showed its viability during the case study and it can be implemented in case of inability to use wired connection. [11]

The deployed system enables real-time evaluation of the following parameters using inverse dynamics model:

- specific heat loss (per cubic meter),
- feed-forward value of indoor air temperature,
- estimated value of generalized temperature perturbation.

This enables continuous adjustment of heat supply into the building heating system, eliminating the impacts of both slow and fast perturbing factors, which ensures better stability of indoor air temperature and reduces the building's energy consumption.

Fig. 5 demonstrates the chart of estimated value of generalized temperature perturbation along with outdoor and indoor air temperatures (measured data, April 2015).

Fig. 6 demonstrates the reference signals for heat medium temperature inside the feed pipeline when implementing baseline control method, as well as implementing proposed feed-forward control method. The building's performance is close to the design in the night time, requiring minimum adjustment of heat supply to the building; by contrast, during the day time, perturbing factors (in particular, solar radiation and internal heat release from running equipment and people inside the building) result in the growth of generalized temperature perturbation, which in turn requires countervailing the influence of the above factors by reducing heat supply temperature. Thus, the lower heat supply temperature causes reducing heat consumption.

Also, implementing the proposed advanced heating control algorithm allows increasing the comfort level inside the building. Fig. 7 shows the charts of indoor air temperature when controlled by the baseline control method and the proposed feed-forward control method. As you may see, daily fluctuations of indoor air temperature dropped from about $\pm 1^\circ\text{C}$ to about $\pm 0.5^\circ\text{C}$. In addition, the

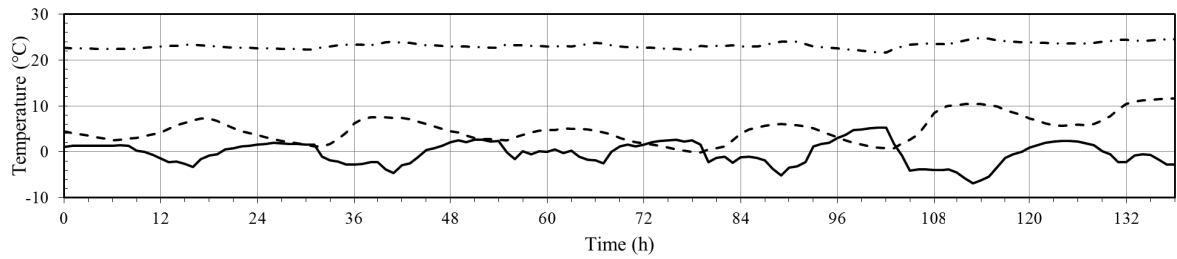


Figure 5. Daily fluctuations of generalized temperature perturbation. Dashed line stands for outdoor air temperature, dash-dotted line stands for indoor air temperature, solid line stands for estimated value of generalized temperature perturbation.

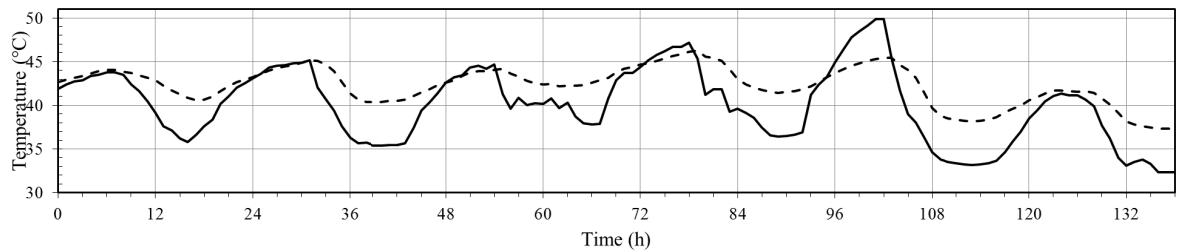


Figure 6. The reference signals for the heat medium temperature in the supply pipeline produced by PLC. Dashed line corresponds to the baseline control depending on outdoor air temperature, solid line corresponds to the proposed feed-forward control method.

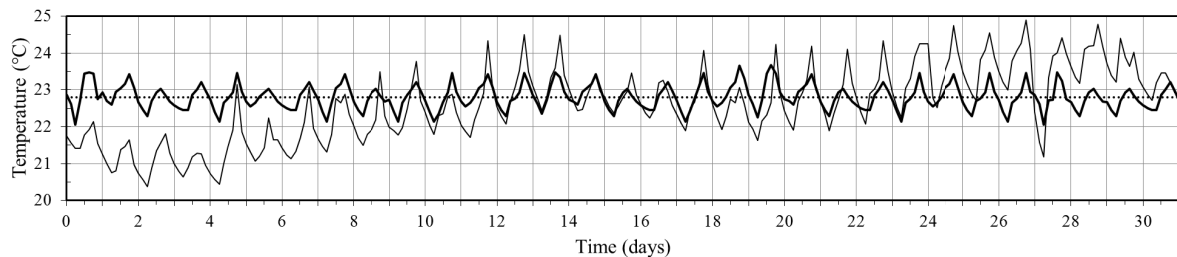


Figure 7. Indoor air temperature fluctuations. Thin solid line corresponds to the baseline control depending on outdoor air temperature (March 2015), thick solid line corresponds to the proposed feed-forward control method (April 2015), dotted line stands for temperature preset (22.8°C).

proposed method eliminates the static error that is typical for the control method based on generalized temperature perturbation.

IV. CONCLUSION

The results obtained demonstrate the applicability of the proposed advanced control method that accounts for the indoor air temperature and employs feed-forward control based on thermal performance inverse dynamics model. This approach proved its high efficiency in automatic heating control systems decreasing heat medium temperature in feed pipeline, sufficiently reducing indoor air temperature fluctuations caused by various perturbing factors, and eliminating static control error. This occasioned the reducing of overall energy consumption by building heating system, at the same time increasing the comfort level of the building.

REFERENCES

- [1] J. M. Salmerón, S. Álvarez, J. L. Molina, A. Ruiz, and F. J. Sánchez, "Tightening the energy consumptions of buildings depending on their typology and on Climate Severity Indexes," *Energy and Buildings*, vol. 58, 2013, pp. 372–377.
- [2] T. Salsbury, P. Mhaskar, and S. J. Qin, "Predictive control methods to improve energy efficiency and reduce demand in buildings," *Computers and Chemical Engineering*, vol. 51, 2013, pp. 77–85.
- [3] L. S. Kazarinov, T. A. Barbasova, O. V. Kolesnikova, and A. A. Zakharova, "Method of multilevel rationing and optimal forecasting of volumes of electric-energy consumption by an industrial enterprise. Automatic Control and Computer Sciences," vol. 48 (6), 2015, pp. 324–333.
- [4] P. H. Shaikh, N. M. Nor, P. Nallagownden, I. Elamvazuthi, and T. Ibrahim, "A review on optimized control systems for building energy and comfort management of smart sustainable buildings," *Renewable and Sustainable Energy Reviews*, vol. 34, 2014, pp. 409–429.
- [5] F. Oldewurtel, A. Parisio, C. Jones, M. Morari, D. Gyalistras, M. Gwerder, et al. "Energy efficient building climate control using stochastic model predictive control and weather predictions. Proceedings of American control conference; 2010.
- [6] V. V. Abdullin, D. A. Shnayder, A. A. Basalaev, "Building Heating Feed-forward Control Based on Indoor Air Temperature Inverse Dynamics Model", *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2014, WCECS 2014*, 22-24 October, 2014, San Francisco, USA, pp. 886-892.
- [7] V. V. Abdullin, D. A. Shnayder, L. S. Kazarinov, "Method of Building Thermal Performance Identification Based on Exponential Filtration," *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2013, WCE 2013*, 3–5 July, 2013, London, U.K., pp. 2226–2230.
- [8] E. Y. Sokolov, *Central heating and heating networks (in Russian: Теплофикация и тепловые сети)*, Moscow: MPEI, 2011.
- [9] L. S. Kazarinov, S. I. Gorelik, "Prediction of random oscillatory processes on the basis of the exponential smoothing method, *Automation and Remote Control*, 1994, 55:10, pp. 1413–1419.
- [10] D. A. Shnayder, M. V. Shishkin, "Adaptive controller for building heating systems applying artificial neural network" (in Russian: Адаптивный регулятор отопления здания на основе искусственных нейронных сетей), *Automatics and control in technical systems*, Edited book – Chelyabinsk: South Ural State University press, 2000, pp. 131–134.
- [11] D. A. Shnayder, V. V. Abdullin, "A WSN-based system for heat allocating in multiflat buildings," *2013 36th International Conference on Telecommunications and Signal Processing Proceedings, TSP 2013*, 2-4 July, 2013, Rome, Italy, Article number 6613915, pp. 181–185.

Self-Tuned Fuzzy Logic Control of a pH Neutralization Process

Parikshit Kishor Singh¹, Surekha Bhanot², Hare Krishna Mohanta³, Vinit Bansal⁴

Dept. of Electrical & Electronics & Engineering^{1,2}, Dept. of Chemical Engineering³, Application Engineering Dept.⁴

BITS Pilani, Pilani campus^{1,2,3}, National Instruments, India⁴

Pilani, Rajasthan, India^{1,2,3}, Bengaluru, Karnataka, India⁴

pkksingh.bitspilani@gmail.com¹, surekha0057@gmail.com², harekrishna.bits@gmail.com³, vinitbansal2010@gmail.com⁴

Abstract—On-line implementation of self-tuning mechanism based adaptive fuzzy logic control of a pH neutralization process which takes care of steady state error and time taken to reach steady state under varying operating conditions has been presented in this paper. The pH neutralization system is Armfield pH Sensor Accessory (PCT42) in conjunction with Process Vessel Accessory (PCT41) and Multifunction Process Control Teaching System (PCT40). The proposed adaptive scheme updates the normalized universe of discourse of output fuzzy membership functions with varying scaling factors based on error and change of error values. The speed of response of the adaptive controller is taken care by use of coarse control technique whereas amount of deviation under steady state is accounted with the help of fine control technique. The performance of adaptive scheme is tested for pH control at equivalence point. LabVIEW software is used for online communication, control and display.

Keywords—pH neutralization; nonlinear process; fuzzy logic control; self-tuning control; adaptive control; on-line control

I. INTRODUCTION

Control of pH has vital significance in our daily life. Modern process industries such as food processing units, biopharmaceutical manufacturing plants, iron & steel industry and thermal power plants run various operations where pH monitoring and controlling are critical. More importantly, vast and rapid globalization necessitated enormous focus on pH control of industrial effluents using wastewater treatment so that ecological and environmental balance could be maintained in our Mother Nature. pH control is often taken as benchmark problem for nonlinear control because of highly nonlinear nature of neutralization reaction. Time varying nature of pH neutralization process essentially makes pH as a moving target whose precise control is almost impossible to achieve. Therefore, although pH control is conventional problem but it still fascinates many young researchers.

Early works on pH control were based on design of adaptive techniques using dynamic process models developed using first-principle approaches such as laws of conservation, physical and chemical laws, reaction invariants, and strong acid equivalent [1], [2], [3], [4], [5], [6]. Rapid advances in modern process control led to development of nonlinear model based predictive control schemes. Many popular model predictive control strategies incorporated nonlinear process models including those based on Wiener, Hammerstein, Volterra series,

Laguerre polynomial techniques, and neural networks [7], [8], [9], [10], [11], [12], [13], [14]. In addition, popular conventional adaptive techniques such as gain-scheduling, model reference adaptive control and self-tuning regulators are also realized using advanced identification and control techniques [15], [16], [17].

Dynamic pH model ability to replicate the actual nonlinear behavior of neutralization processes limited the accuracy and ability of the controller. Thus few researchers used concept of model-free intelligent control [18], [19], [20]. The fuzzy logic control is based on intelligent methodology of human thinking and decision making mechanisms. Self-organizing based adaptive control has been implemented using fuzzy logic [21], [22]. This paper discusses on-line application of fuzzy logic based self-tuned pH control of strong acid-strong base neutralization process stream.

II. DESCRIPTION OF NEUTRALIZATION PLANT

The Armfield pH neutralization system is shown in Figure 1. The pH probe PCT42 calibration against buffer pH solutions of 4, 7 and 9.2 results in a linear relationship between sensor voltage and equivalent pH. The pH neutralization process takes place in PCT41 with perfect mixing and constant maximum volume (V). The PCT40 has two peristaltic pumps A and B which regulate flow of hydrochloric acid (HCl) and sodium hydroxide (NaOH) having concentrations C_a and C_b respectively. Although speed of pumps A and B, S_a and S_b respectively, can vary from 0 to 100% of maximum speed, both pumps do not start below a minimum value of 18%. During pump operation within useful range of 18 to 100%, pump flowrate varies almost linearly with speed. A brief and important specifications of pH neutralization process is given in Table I.

Using standard universal synchronous bus interface the PCT40 communicates with LabVIEW software installed on a personal computer having Windows XP Professional 2003 operating system. The PCT40 interface device driver contains a dynamic link library (DLL) file which stores various input-output analog and digital control signal values. The analog signals between 0 V or 0% to 5 V or 100% are stored in 12-bit signed-magnitude representation as 000000000000 to 011111111111 in binary or 0 to 2047 in decimal whereas the digital signals either 0 V or 5 V are stored in 1-bit representation as 0 or 1 in binary, respectively. LabVIEW software accesses the DLL file for following functionality: read analog input,



Figure 1. Armfield pH neutralization system (PCT42 plus PCT41 plus PCT40)

TABLE I. NEUTRALIZATION PROCESS SPECIFICATIONS

Quantity	Specification
PCT41 volume (V)	2000 mL
Concentration of HCl (C_a)	0.0174 mol/L
Concentration of NaOH (C_b)	0.0138 mol/L
Speed range of pumps A and B	18 to 100%
Equivalent flowrate of pump A (F_a)	0.2021 to 5.1139 mL/s
Equivalent flowrate of pump B (F_b)	0.2989 to 5.8749 mL/s
Voltage range of pH sensor	0 to 5 V
Equivalent pH reading	0.1868 to 13.2438
Sampling period	1 s

pH probe value from channel 11 (Ch11); write analog outputs, pump A and B speed values to digital-to-analog converters DAC0 and DAC1 respectively; write digital output, stirrer signal value to digital output line 7 (DO7).

III. DESIGN OF ADAPTIVE FUZZY LOGIC CONTROL

The fuzzy logic control (FLC) of pH neutralization process is based on Mamdani fuzzy inference system (FIS). The input variables used for FLC are $e^*(k) = e(k)/K_1 = (pH_{SP}(k) - pH(k))/K_1$ and $ce^*(k) = ce(k)/K_2 = (e(k) - e(k-1))/K_2$, where pH_{SP} is setpoint, pH is controlled variable, e is error, ce is change of error, k is sampling instant, and K_1 , K_2 are scaling factors. The output variable used for FLC is $co^*(k)$ where co^* is normalized change of output. The normalized membership functions for the input variables (e^* , ce^*) and output variable (co^*) are shown in Figure 2 and Figure 3 respectively. The rule base for the input and output variables are shown in Table II.

For a nonlinear process, the FLC needs to be optimized for better performance. However, often operating conditions of the nonlinear process changes which means repeated application of optimization procedure. Self-tuning mechanism allows adaptive FLC to alter its input and/or output scaling factors as per changes

in the operating conditions. In this paper, we have designed a self-tuned FLC, shown in Figure 4, whose output scaling factor K_3 are as per entries in Table III. The input variables e and ce are divided into following seven identical regions: $e_1, ce_1 \in [-6, -1]$; $e_2, ce_2 \in [-1, -0.5]$; $e_3, ce_3 \in [-0.5, -0.1]$; $e_4, ce_4 \in [-0.1, 0.1]$; $e_5, ce_5 \in (0.1, 0.5]$; $e_6, ce_6 \in (0.5, 1]$; $e_7, ce_7 \in (1, 6]$. As evident from Table III entries, K_3 has been assigned a larger value if pH is away from pH_{SP} and K_3 has been assigned a smaller values as pH approaches pH_{SP} . Using large K_3 when pH is at distance from pH_{SP} will ensure reduced settling time and using small K_3 when pH is near pH_{SP} will ensure steady-state response within settling band. Therefore Table III entries validates use of coarse control and fine control techniques.

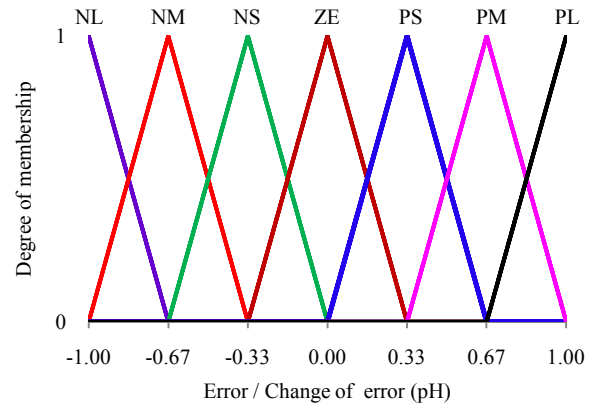


Figure 2. Fuzzy inputs membership functions

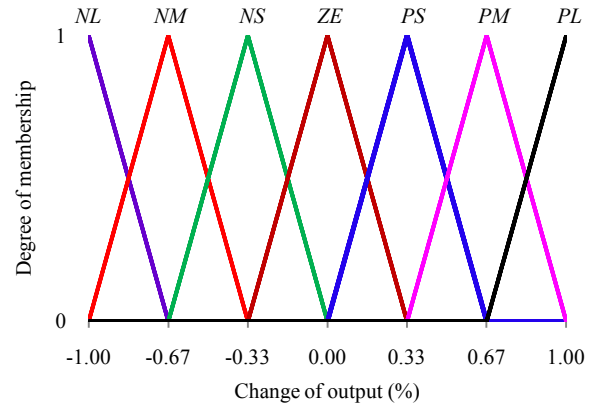


Figure 3. Fuzzy output membership functions

TABLE II. FUZZY RULE BASE

e	ce						
	NL	NM	NS	ZE	PS	PM	PL
NL	NL	NL	NL	NL	NM	NS	ZE
NM	NL	NL	NL	NM	NS	ZE	PS
NS	NL	NL	NM	NS	ZE	PS	PM
ZE	NL	NM	NS	ZE	PS	PM	PL
PS	NM	NS	ZE	PS	PM	PL	PL
PM	NS	ZE	PS	PM	PL	PL	PL
PL	ZE	PS	PM	PL	PL	PL	PL

TABLE III. DETERMINATION OF K_3

Range for e	Range for ce						
	ce_1	ce_2	ce_3	ce_4	ce_5	ce_6	ce_7
e_1	8	8	8	8	6	4	2
e_2	8	8	8	6	4	2	4
e_3	8	8	6	4	2	4	6
e_4	8	6	4	2	4	6	8
e_5	6	4	2	4	6	8	8
e_6	4	2	4	6	8	8	8
e_7	2	4	6	8	8	8	8

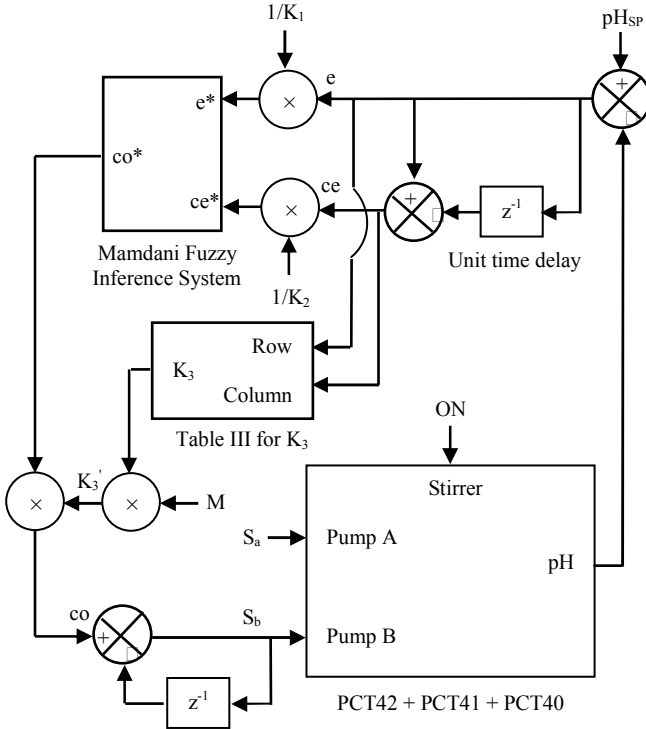


Figure 4. Self-tuned fuzzy logic based pH control

IV. RESULTS AND DISCUSSIONS

Figure 5 shows pseudocode of the proposed control algorithm which consists of two parts for sequential execution: first process initialization and second adaptive FLC. Process initialization ensures that initial pH value falls within user specified lower bound (pH_{LB}) and upper bound (pH_{UB}). In order to evaluate performance of adaptive intelligent controller, pH_{SP} is changed from 6 to 7 with [pH_{LB} (= 5.9), pH_{UB} (= 6)], and from 8 to 7 with [pH_{LB} (= 5.9), pH_{UB} (= 6)]. Table IV summarizes ISE performance of the proposed controller for various values of scaling factors K_1 , K_2 and M . In particular, graphical performance comparison in terms of pH response, pumps speed and magnified output scaling factor K_3' has been shown in Figure 6, Figure 7 and Figure 8 respectively for Cases 1 to 4. Following observations can be made from the obtained results.

(a) For $M = 1$, pH response is slower than the same for $M = 2, 3, 4$. Since mostly e and ce attains values within range

```

Start %% Begin LabVIEW implementation
for n = 1 to No. of Iterations %% Begin initialization
  write in DLL to start the stirrer
  read from DLL to obtain pH sensor voltage
  estimate pH
  while initial pH is not within range [ $pH_{LB}$   $pH_{UB}$ ]
    if  $pH < pH_{LB}$ 
      write in DLL to set  $S_a = 35$  and  $S_b = 40$ 
    end
    if  $pH > pH_{UB}$ 
      write in DLL to set  $S_a = 35$  and  $S_b = 35$ 
    end
  end
end %% End initialization
initialize  $S_a = 35$ ,  $S_b = 38$ , ISE = 0, final  $pH_{SP}$ 
for k = 1 to Sampling Duration %% Begin control
  estimate  $pH(k)$ ,  $e(k)$ , and  $ce(k)$ 
  obtain  $e^*(k) = e(k)/K_1$ ,  $ce^*(k) = ce(k)/K_2$ 
  obtain  $co^*(k)$  using FIS based on Table II
  obtain  $K_3(k)$  using Table III
  multiply  $K_3(k)$  by factor  $M$  to obtain  $K_3'(k)$ 
  scale  $co^*(k)$  by factor  $K_3'(k)$  to obtain  $co(k)$ 
  obtain  $S_b(k) = S_b(k-1) + co(k)$ 
  write in DLL to update  $S_a$  and  $S_b$ 
  update ISE(k) = ISE(k-1) + ( $e(k)$ )2
  update sampling time  $k = k + 1$ 
end %% End control
end %% End LabVIEW implementation

```

Figure 5. Pseudocode of self-tuned FLC

given by e_3 , e_4 , e_5 and ce_3 , ce_4 , ce_5 respectively, magnification of K_3 results in faster pH response.

(b) For $M = 1$, magnitude of first overshoot is largest as compared to the same for $M = 2, 3, 4$. A magnified K_3 provides better neutralization of acidic process stream using basic manipulated stream.

(c) For $M = 1, 2, 3$, pH response remains within 7 ± 0.2 pH settling band in a better way as compared to the same for $M = 4$. A magnified K_3 sometimes drives the pH response outside of the settling band.

(d) For $M = 4$, pH response indicates self-controlling property of the adaptive FLC. Figure 6 shows that during sampling instants 85 to 130 seconds, pH response is showing tendency to go unbound. Subsequently, the self-tuned fuzzy controller adjusts the value of K_3 so that pH response comes within the desired settling band.

Similarly, for Cases 25 to 28, Figure 9, Figure 10 and Figure 11 indicates graphical performance comparisons in terms of pH response, pumps speed and magnified output scaling factor K_3' respectively. From the system response, it is observed that a magnified K_3 results in less undershoot, faster response and reduced settling time. Table IV shows that, under similar operating conditions and parameters setting, ISE is largest for $M = 1$ and ISE improves for $M = 2, 3, 4$.

V. CONCLUSION

Self-tuned fuzzy logic control has been implemented on a laboratory scale pH neutralization systems from

TABLE IV. PERFORMANCE SUMMARY OF SELF-TUNED FLC

Case	Initial pH _{SP}	Final pH _{SP}	K ₁	K ₂	M	ISE
1	6	7	10	0.5	1	23.75
2	6	7	10	0.5	2	18.91
3	6	7	10	0.5	3	14.72
4	6	7	10	0.5	4	17.42
5	6	7	10	1	1	58.67
6	6	7	10	1	2	30.98
7	6	7	10	1	3	35.07
8	6	7	10	1	4	13.79
9	6	7	20	0.5	1	24.58
10	6	7	20	0.5	2	21.74
11	6	7	20	0.5	3	18.90
12	6	7	20	0.5	4	18.22
13	6	7	20	1	1	26.94
14	6	7	20	1	2	14.03
15	6	7	20	1	3	16.55
16	6	7	20	1	4	14.88
17	6	7	30	0.5	1	34.95
18	6	7	30	0.5	2	31.65
19	6	7	30	0.5	3	28.82
20	6	7	30	0.5	4	30.14
21	6	7	30	1	1	35.99
22	6	7	30	1	2	21.63
23	6	7	30	1	3	19.15
24	6	7	30	1	4	15.40
25	8	7	10	0.5	1	20.55
26	8	7	10	0.5	2	15.34
27	8	7	10	0.5	3	13.73
28	8	7	10	0.5	4	12.69
29	8	7	20	0.5	1	29.26
30	8	7	20	0.5	2	22.86
31	8	7	20	0.5	3	21.56
32	8	7	20	0.5	4	20.12

Armfield. LabVIEW software has been used for on-line signal processing. Adaptive fuzzy controller performance has been evaluated for step change in pH setpoint from 6 to 7 and 8 to 7. In almost all test results, pH response finally settles within 7 ± 0.2 pH band. In some cases, when pH response occasionally overshoots and undershoots the above band, controller adjusts its output universe of discourse and again brings the pH response back within the desired band. The self-tuned fuzzy controller gives better performance in terms of higher speed of response, less deviation from setpoint and better settling within band for magnified fuzzy output scaling factor.

REFERENCES

- [1] T.J. McAvoy, E. Hsu, and S. Lowenthal, "Dynamics of pH in controlled stirred tank reactor," *Ind. Eng. Chem. Process Des. Develop.*, vol. 11, pp. 68-70, 1972.
- [2] T.J. McAvoy, "Time optimal and Ziegler-Nichols control. Experimental and theoretical results," *Ind. Eng. Chem. Process Des. Develop.*, vol. 11, pp. 71-78, 1972.
- [3] T.K. Gustafsson and K.V. Waller, "Dynamic modeling and reaction invariant control of pH," *Chemical Engineering Science*, vol. 38, pp. 389-398, 1983.
- [4] T.K. Gustafsson, "An experimental study of a class of algorithms for adaptive pH control," *Chemical Engineering Science*, vol. 40, pp. 827-837, 1983.
- [5] R.A. Wright and C. Kravaris, "Nonlinear control of pH processes using strong acid equivalent," *Ind. Eng. Chem. Res.*, vol. 30, pp. 1561-1572, 1991.
- [6] R.A. Wright, M. Soroush, and C. Kravaris, "Strong acid equivalent control of pH processes: An experimental study," *Ind. Eng. Chem. Res.*, vol. 30, pp. 2437-2444, 1991.
- [7] Y.-K. Yeo and T.-I. Kwon, "A neural PID controller for the pH neutralization process," *Ind. Eng. Chem. Res.*, vol. 38, pp. 978-987, 1999.
- [8] B.M. Åkesson, H.T. Toivonen, J.B. Waller, and R.H. Nyström, "Neural network approximation of a nonlinear model predictive controller applied to a pH neutralization process," *Computers & Chemical Engineering*, vol. 29, pp. 323-335, 2005.
- [9] S.J. Norquay, A. Palazoglu, and J.A. Romagnoli, "Application of wiener model predictive control (WMPC) to a pH neutralization experiment," *IEEE Transactions on Control System Technology*, vol. 7, pp. 437-445, 1999.
- [10] S. Oblak and I. Škrjanc, "Continuous-time Wiener-model predictive control of a pH process based on a PWL approximation," *Chemical Engineering Science*, vol. 65, pp. 1720-1728, 2010.
- [11] H.C. Park, S.W. Sung, and J. Lee, "Modeling of Hammerstein-Wiener processes with special input test signals," *Ind. Eng. Chem. Res.*, 45, 1029-1038, 2006.
- [12] K.P. Fruzzetti, A. Palazoglu, and K.A. McDonal, "Nonlinear model predictive control using Hammerstein models," *Journal of Process Control*, vol. 7, pp. 31-41, 1997.
- [13] S. Mahmoodi, J. Poshtan, M.R. Jahed-Motlagh, and A. Montazeri, "Nonlinear model predictive control of a pH neutralization process based on Wiener-Laguerre model," *Chemical Engineering Journal*, vol. 146, pp. 328-337, 2009.
- [14] R. Diaz-Mendoza and H. Budman, "Structured singular valued based robust nonlinear model predictive controller using Volterra series models," *Journal of Process Control*, vol. 20, pp. 653-663, 2010.
- [15] R.H. Nyström, B.M. Åkesson, and H.T. Toivonen, "Gain-scheduling controllers based on velocity-form linear parameter-varying models applied to an example process," *Ind. Eng. Chem. Res.*, vol. 41, pp. 220-229, 2002.
- [16] M.C. Palancar, J.M. Aragón, J.A. Miguéns, and J.S. Torrecilla, "Application of a model reference adaptive control system to ph control. effects of lag and delay time," *Ind. Eng. Chem. Res.*, vol. 35, pp. 4100-4110, 1996.
- [17] M. Albaz, H. Hapoğlu, G. Özkan, and S. Altuntas, "Application of self-tuning PID control to a reactor of limestone slurry titrated with sulfuric acid," *Chemical Engineering Journal*, vol. 116, pp. 19-24, 2006.
- [18] Z. Zheng and N. Wang, "Model-Free Control based on Neural Networks," in *Proc. International Conference on Machine Learning and Cybernetics, 2002, Beijing*, pp. 2180-2183.
- [19] S. Syafie, F. Tadeo, and E. Martinez, "Macro-actions in model-free intelligent control with application to pH control," in *Proc. 44th IEEE International Conference on Decision and Control, and the European Control Conference, 2005, Spain*, pp. 2710-2715.
- [20] S. Syafie, F. Tadeo, and E. Martinez, "Model-free learning control of neutralization processes using reinforcement learning," *Engineering Applications of Artificial Intelligence*, vol. 20, pp. 767-782, 2007.
- [21] E.H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," *IEEE Trans. on Computers*, vol. 26, pp. 1182-1191, 1977.
- [22] T.J. Procyk and E.H. Mamdani, "A linguistic self-organizing process controller," *Automatica*, vol. 15, pp. 15-30, 1979.

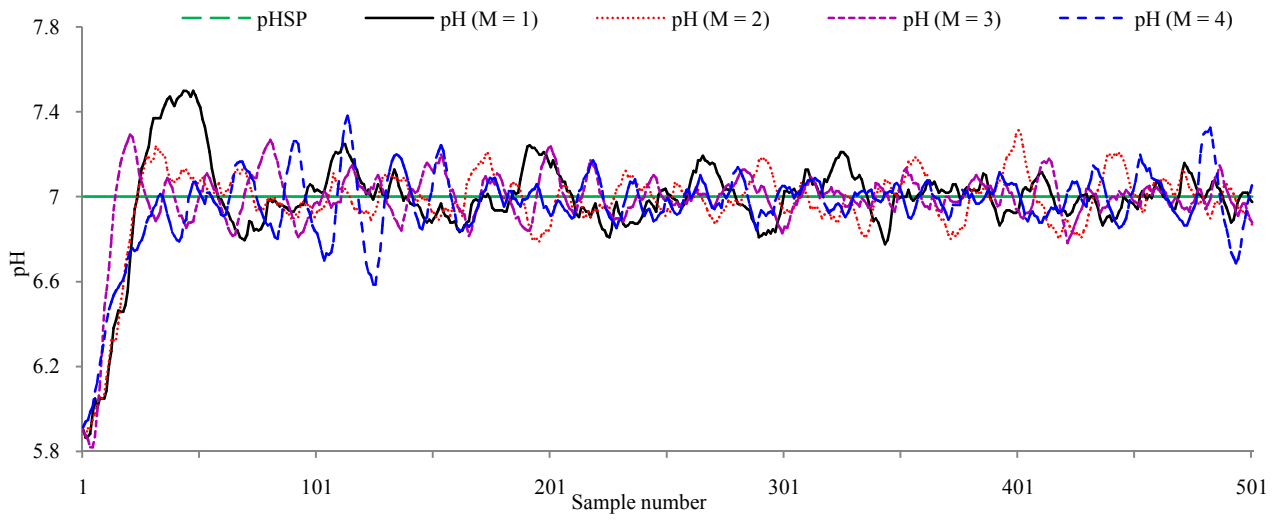


Figure 6. pH response for $K_1 = 10$, $K_2 = 0.5$, $\text{pH}_{\text{SP}} (\text{initial}) = 6$, $\text{pH}_{\text{SP}} (\text{final}) = 7$

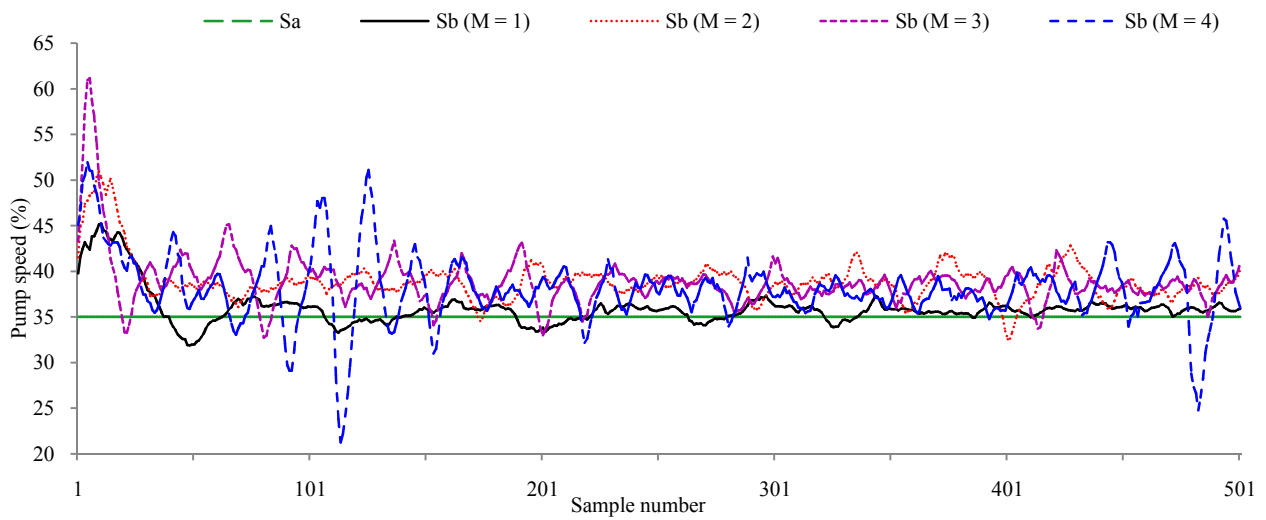


Figure 7. Pump B response for $K_1 = 10$, $K_2 = 0.5$, $\text{pH}_{\text{SP}} (\text{initial}) = 6$, $\text{pH}_{\text{SP}} (\text{final}) = 7$

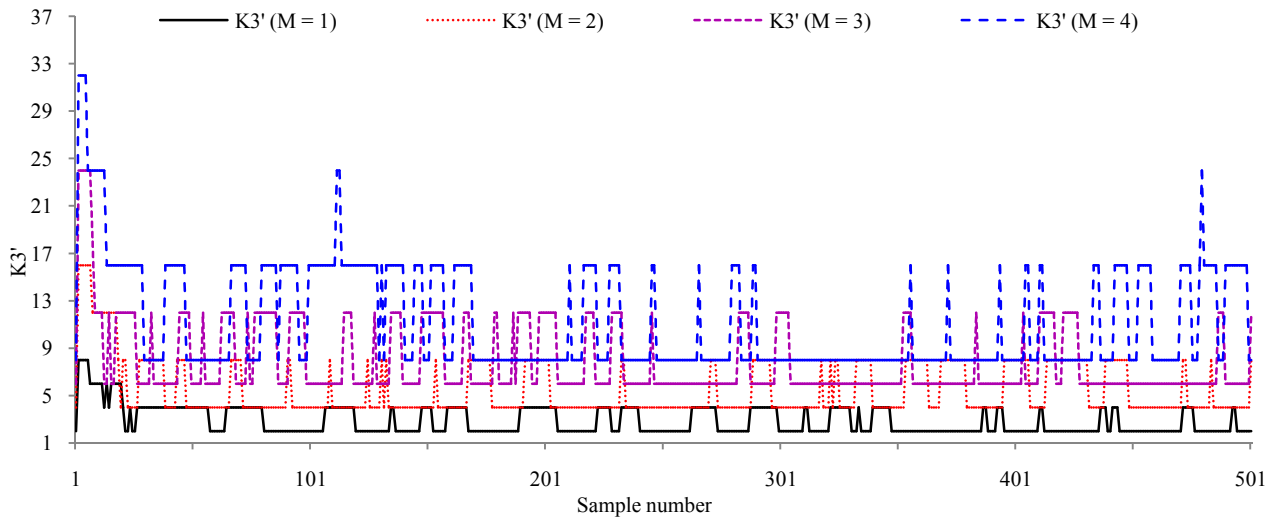


Figure 8. K_3' for $K_1 = 10$, $K_2 = 0.5$, $\text{pH}_{\text{SP}} (\text{initial}) = 6$, $\text{pH}_{\text{SP}} (\text{final}) = 7$

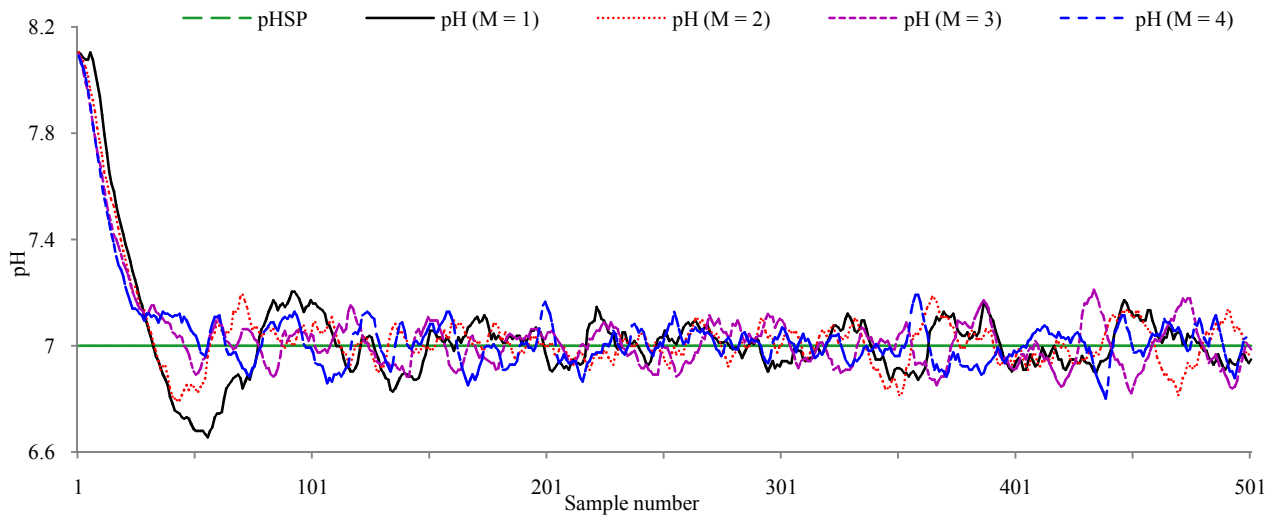


Figure 9. pH response for $K_1 = 10$, $K_2 = 0.5$, $\text{pH}_{\text{SP}}(\text{initial}) = 8$, $\text{pH}_{\text{SP}}(\text{final}) = 7$

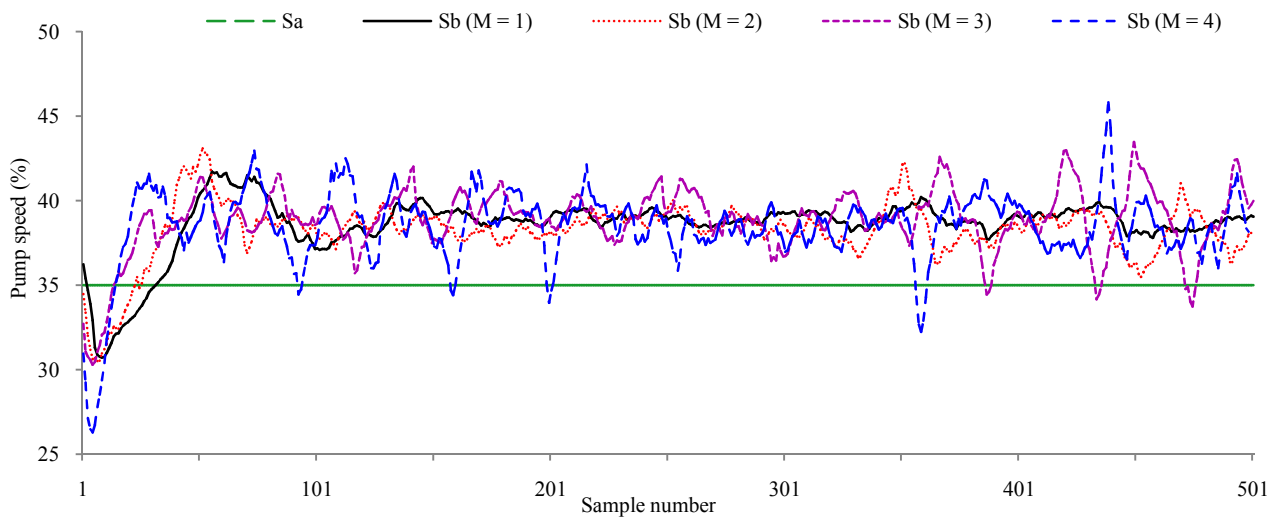


Figure 10. Pump B response for $K_1 = 10$, $K_2 = 0.5$, $\text{pH}_{\text{SP}}(\text{initial}) = 8$, $\text{pH}_{\text{SP}}(\text{final}) = 7$

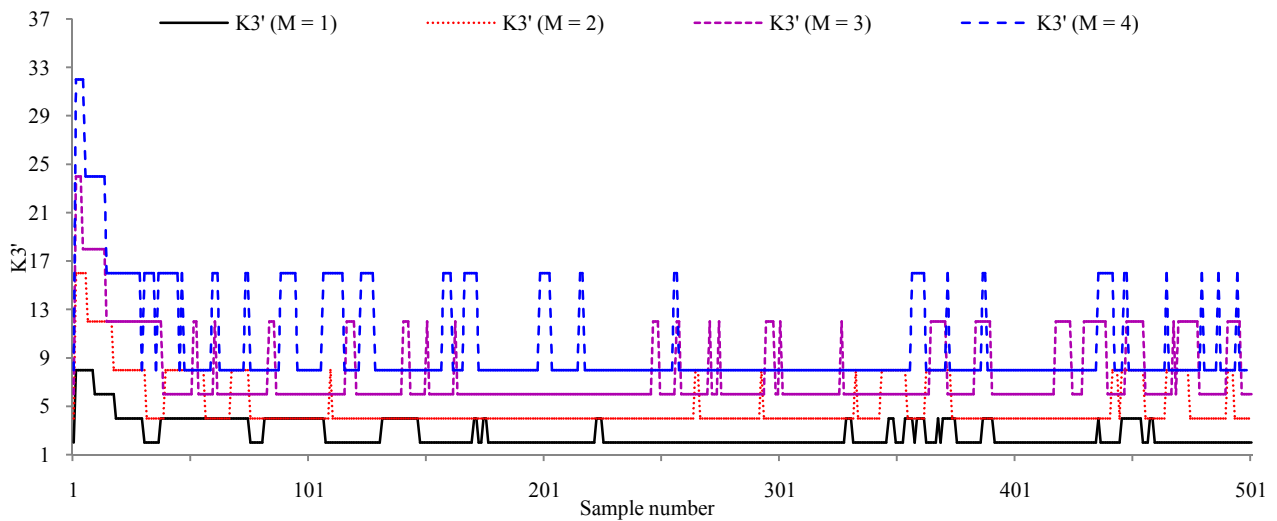


Figure 11. K_3' for $K_1 = 10$, $K_2 = 0.5$, $\text{pH}_{\text{SP}}(\text{initial}) = 8$, $\text{pH}_{\text{SP}}(\text{final}) = 7$

Decentralised PI Controller Design and Tuning Approaches

Lan-Xiang Zhu¹, Feng Yu¹,

¹ School of Electronic Information, Changchun Architecture & Civil Engineering College, Changchun, China.

Ding-Wen Yu²,

² School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao, China.

D. L. Yu³

³ Control Group, School of Engineering, Liverpool John Moores University, Liverpool, U.K.

Corresponding author: d.yu@ljmu.ac.uk

Abstract—This paper compares two approaches to tune PI controllers in a multivariable process. A modified Direct Synthesis method extends the technique to a multivariable system. Results show that the modified Direct Synthesis technique relies on model accuracy and suitability of the desired transfer function and process model. The effectiveness of this technique may break down when these conditions are not met. Genetic algorithms provide a method based on the evolution of species via natural selection. The GA method allows the use of a specific structure, such as PI controller, and then gets the best possible results for that structure. The concept has been adapted into an optimisation technique that can be used to find the best controller parameters to meet user defined transient specifications for a complex process. Thus, the robustness of the method allows it to outperform the former technique. Simulation results are shown to illustrate the comparisons between these two methods and the effectiveness of GAs.

Keywords—Decentralized control; multivariable system; PI controller tuning; genetic algorithm.

I. INTRODUCTION

Although research on modern control theory has been thriving for many years and made significant theoretical progress, it is difficult to apply many of these theories in real practical problems. For example, these approaches may produce non-standard forms of controllers that make them difficult to implement and understand. Even when a suggested multivariable solution is theoretically good, it may not be easily implemented in engineering practice. Certain existing controller structures have been standardised and are still being used widely today. The Proportional and Integral (PI) controller is included in this category. Even though the task of tuning a PI controller is quite tedious and time consuming, the controller itself is easy to obtain and apply. There are a number of empirical tuning techniques available such as the Zeigler-Nichols (Z-N) [11] methods. These empirical methods do not require a process model and the tuning of the controller parameters are obtained experimentally. This method is time-consuming especially when there are several loops to consider. Also, the final controller parameters may not be realisable. Therefore, this method is not practical for many multivariable systems. However, Franke, Kruger and Knoop [3] developed an approach for a multivariable process using the DSM by substituting all the s-terms that defines the controller

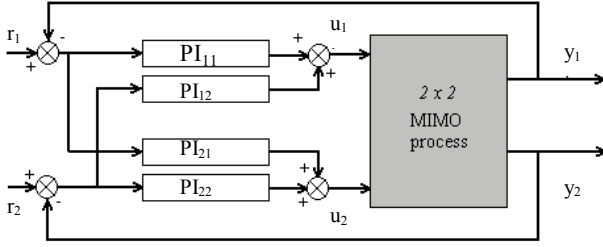
parameters. This allows the benefits of DSM to be applied to MIMO systems for a specified controller structure. The feedback controller is designed by specifying the desired transient closed-loop response. This revised method also allows the controller structure to be fixed in a PI form. Fig.1 illustrates a multivariable process with decentralised controllers for each loop in the system, known as the multiloop control structure [7]. However, the s-term and the desired transient transfer function values are limited within a certain acceptable range. This technique also depends on the accuracy of the system model.

Genetic Algorithms (GAs) provide an alternative method. They are global search methods that are based on natural population genetics and have been used as an optimising tool in control systems. Holland [4] and De Jong [2] and others have demonstrated the excellent achievement of GAs. The powerful capabilities of genetic algorithms can be utilised to locate near optimum values of controller tuning parameters to meet an operator defined performance specification [8, 9]. The required process performance will be specified in terms of output closed-loop transient responses. Controller parameters will be evaluated, by simulation, to meet the corresponding closed-loop system performance using an objective function, which can be user defined. Due to the robustness of this technique, the performance of GAs does not rely on the characteristic of the plant under control [10]. Thus, GAs are applicable to a wide range of practical plant. A decentralised PI controller structure is used for this paper. This controller structure is a desirable controller structure for a MIMO process because of its design simplicity [7]; which is a good advantage in transferring the theoretical procedures to real life controller implementations.

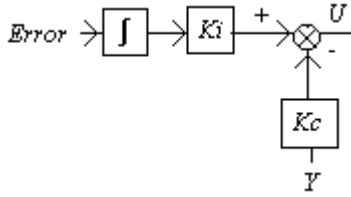
II. MODIFIED DIRECT SYNTHESIS METHOD

In the general Direct Synthesis Technique, controller expressions can be obtained from the desired system output specification and if the process model is available. This design however, may not result in a standard P/PI/PID form and it is difficult to apply to a MIMO system and the final controller parameters obtained may not be realisable at all. Franke, Kruger and Knoop [3] uses a different approach of this DSM but still uses the system closed loop transfer function and defines the form of the controller structure

required. The controller parameters are then obtained by substituting the s-term; in the controller parameters transfer functions, with a numerical value within a certain range.



(a) 2-input 2-output MIMO process with PI controllers



b) PI Controller configuration for each loop

Fig. 1 Multivariable PI control

Where:

K_c =Proportional controller parameter

K_i = Integral Action controller parameter

$U(s)$ =Controller output

The desired closed-loop transfer function can be obtained using the SISO or MIMO closed loop system as in Fig.1 (a & b) [3].

$$Y(s) = G(s)U(s) \quad (1)$$

$$U(s) = \frac{1}{s} K_i R(s) - \frac{1}{s} K_i Y(s) - K_c Y(s) \quad (2)$$

Substitute $U(s)$ into equation (1):

$$Y(s) = G(s) \left(\frac{1}{s} K_i R(s) - \frac{1}{s} K_i Y(s) - K_c Y(s) \right) \quad (3)$$

$$\frac{Y(s)}{R(s)} = F(s) \quad (4)$$

Where:

$F(s)$ = Specified system closed-loop transfer function,

$G(s)$ = Multivariable process transfer function,

Note that these equations involve matrices. Thus:

$$F(s) = \frac{Y(s)}{R(s)} = \left(I_p + G(s) \left(\frac{K_i}{s} + K_c \right) \right)^{-1} G(s) \frac{1}{s} K_i \quad (5)$$

Where:

p = number of inputs/outputs ; I = identity matrix

Note that in equation (5), the desired transfer function $F(s)$ is standardised, according to the user specifications. Rearranging this equation will lead to the controller in s-domain expression as in equation (6).

$$[K_c \ K_i] = -[G^{-1}(s) \ G^{-1}(ks)] \left[\frac{I_p}{s} (I_p - F^{-1}(s)) \quad \frac{I_p}{ks} (I_p - F^{-1}(ks)) \right]^{-1} \quad (6)$$

Where:

$k = 1, 2, \dots, n.$

$n = 2$ =PI Controller Structure

$n = 3$ =PID Controller Structure

In order to achieve a PI controller structure, the specifications is constrained for a particular process transfer functions as shown in equation (7).

$$F(s) = \text{diag}[F_1(s), \dots, F_p(s)] = \begin{bmatrix} \prod_{j=1}^{\hat{d}_1} \frac{a_{1ij}}{(s+a_{1ij})} & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & \prod_{j=1}^{\hat{d}_p} \frac{a_{pij}}{(s+a_{pij})} \end{bmatrix} \quad (7)$$

\hat{d}_i is defined to be the difference of order of s term between the denominator and the numerator of system transfer function. Also, to acquire the controller parameters, Franke Kruger and Knoop [3] substitute all the s-terms in the equation into a numerical value represented by α as shown in equation (8):

Where: $s = k\alpha$, $k = 1, 2$

$$[K_c \ K_i] = -[G^{-1}(\alpha) \ G^{-1}(2\alpha)] \left[\frac{I_r}{\alpha} (I_r - F^{-1}(\alpha)) \quad \frac{I_r}{2\alpha} (I_r - F^{-1}(2\alpha)) \right]^{-1} \quad (8)$$

The selections of α are limited to a certain range of values according to the type of output required and also the type of process model involved. The range value for α and 'a' can be determined by comparing the standard normalised form with the user-defined specification forms of transfer function, F . For example, let $d_i = 2$, and consider the following:

$$F(s) = G(s)$$

$$\frac{a^2}{s^2 + 2\xi as + a^2} = \frac{\omega_n^2}{s^2 + 2\xi \omega_n s + \omega_n^2} \quad (9)$$

Using the Settling time percentage and damping ratio equation:

$$T_{st\%} = \frac{K}{\xi a} \quad (10)$$

$$\text{Damping ratio} = \xi = \sqrt{\frac{\ln(\%overshoot)^2}{\pi^2 + \ln(\%overshoot)^2}} \quad (11)$$

For critical damped system ($\xi = 1$) :

$$T_{st\%} = \frac{K}{a} \quad (12)$$

For underdamped system:

$$T_{st\%} = \frac{K}{\xi a} \quad (13)$$

Where:

$T_{st\%}$ =System settling time, st%= percentage of settling time, K =constant value; ω_n = natural frequency.

According to Franke, Knoop [3], the range of acceptable α value are:

$$\frac{4}{100T_s}, \dots, \frac{4}{10T_s} \leq \alpha \leq \frac{4}{T_s} \quad (14)$$

$$\therefore \alpha = \frac{4}{T_{4\%s}} = \frac{4}{1 \text{ sec}} = 4$$

Therefore, from Eq. 13, the value of 'a' determines the settling time. For critically damped solution a suitable value of the constant value in Eq 10 can be found by solving Eq 15 for a step input by noting the percentage of the final output.

$$\therefore f(t) = R(1 - (1 + at)e^{-at}) \quad (15)$$

Table 1 Normalised Transient Output Response for Critical Damped System

T	$f(t)$	% Final Value Reached (Approximate)
0	0	0%
1/a	(0.2642)R	26%
2/a	(0.5940)R	59%
3/a	(0.8009)R	80%
4/a	(0.9084)R	90%
5/a	(0.9596)R	96%
6/a	(0.9826)R	98%
7/a	(0.9927)R	99%

Table 1 illustrates the transient response to the step input and provides an approximate relationship between the response and the value of 'a'. Hence, settling time of 4% is:

$$T_{1\%s} = \frac{5}{a}, \quad (16)$$

$$\therefore a = \frac{5}{T_{4\%s}}$$

Table 2 shows the solution for the underdamped solution. Let ξ equals to 0.5912 ($\approx 10\%$ overshoot).

Thus:

$$f(t) = R \left(1 - \left(\frac{1}{1 - \xi^2 + j\xi\sqrt{1 - \xi^2}} \right) e^{-a\xi t} \cos \left(at\sqrt{1 - \xi^2} \right) \right) \quad (17)$$

Table 2: Normalised Transient Output Response for Underdamped System

T	$f(t)$	% Final Value Reached (Approximate)
0	0	0%
1/a	(0.7454)R	75%
2/a	(0.7879)R	79%
3/a	(1.1274)R	113%
4/a	(1.0936)R	110%
5/a	(1.0327)R	103%
6/a	(0.9964)R	100%
7/a	(0.9872)R	99%

Eq 6 represents the expression to determine the PI controller terms. Proportional and Proportional, Integral & Derivative Controller expressions can also be developed, if necessary, in a similar manner. Although this method allows the system to obtain controller parameters in terms of the standard PID controller, it may not achieve the desired specifications due to loop interactions that are a usual occurrence in many systems. A GA will improve on the objective functions values as it tries to optimise the system performance according to the specifications. All these equations are correct if the process used has the same number of inputs and outputs.

Example 1.0

The process used in this example is a 2-input, 2-output multivariable process. Consider the process state space Eq 7 that describe the process model [6]:

$$A = \text{diag}[-0.9321 - 0.9341 - 0.217 - 0.2159 - 11.59 - 8.057]$$

$$C = \begin{bmatrix} 0.68 & -1.6427 & 0.1252 & 0.2234 & 1.4194 & 0 \\ -0.0409 & 0.1558 & 0.0217 & 0.0646 & -1.5583 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1.9826 & 1.338 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Four Proportional & Integral Action (PI) controllers were implemented for this process according the following user specifications.

i) Critical damped system with 4% of system settling time of 1 second for both first and second outputs. Therefore:

$$T_{4\%} = 1s = \frac{5}{a} \quad \therefore a = 5$$

Thus according to the equations in (11):

$$a_{1st \text{ loop}} = 5, \quad a_{2nd \text{ loop}} = 5,$$

Using the Eq 7, the desired transfer function is as follow:

$$F_w(s) = \begin{bmatrix} \frac{5^2}{s^2 + 2(5s) + 5^2} & 0 \\ 0 & \frac{5^2}{s^2 + 2(5s) + 5^2} \end{bmatrix}$$

Consider equation (8), s-terms are then substituted with numerical value. The controller parameters are:

$$K_c = \begin{bmatrix} 1.4094 & 1.8358 \\ -13.6659 & 5.5431 \end{bmatrix} \quad K_i = \begin{bmatrix} 5.091 & 11.6515 \\ -31.723 & 17.4037 \end{bmatrix}$$

Applying a unit step input to each of the system inputs using MATLAB/SIMULINK [5] package. The system transients are as shown in Fig 6.

ii) Underdamped system with settling time of 4 seconds and 10% overshoot for both outputs. Therefore:

$$\xi = \frac{\sqrt{\ln(\%overshoot)^2}}{\sqrt{\pi^2 + \ln(\%overshoot)^2}} = 0.5912$$

$$\omega_n = 1.6916$$

From Table 2, the 1% settling time is considered, thus:

$$T_{4\%s} = \frac{7}{\xi\omega_n}, \therefore a = 2.960, 1s \approx 0.57 * a$$

$$\therefore F(s) = \begin{bmatrix} \frac{3.0625}{s^2 + 1.179s + 3.0625} & 0 \\ 0 & \frac{3.0625}{s^2 + 1.179s + 3.0625} \end{bmatrix}$$

Thus, the controller parameters value:

$$K_c = \begin{bmatrix} 0.0951 & -0.5018 \\ -3.3037 & 0.2859 \end{bmatrix} \quad K_i = \begin{bmatrix} 1.7961 & 4.9595 \\ -8.4561 & 5.1785 \end{bmatrix}$$

Applying a unit step input to each of the system input pattern, the transient outputs in Fig 8.

III. GENETIC ALGORITHMS OPTIMIZATION

Genetic Algorithms (GAs) implement an optimisation technique based on a simulation of the natural law of the evolution of species via natural selection, in order to have the fittest individual to survive. The searching process of GAs is similar to the natural evolution of biological creatures in which successive generations are created and raised and they themselves continue the cycle. In this algorithm, the fittest among a group of artificial species, which are represented in a form of string structure, survive and form a new generation with those that are produced through structured but random information. In every new generation, a new offspring or set of strings is created using information of the gene of the fittest old generation. This

allows GA to exploit the historical information of a 'species' in order to have gradual improved characteristics or behaviour.

The GA can be mainly classified into three parts. First is the structure of strings, that includes a coding and decoding method, second is the fitness function that defines the specific performance requirement, and finally, the genetic operators which involve reproduction, crossover and mutation.

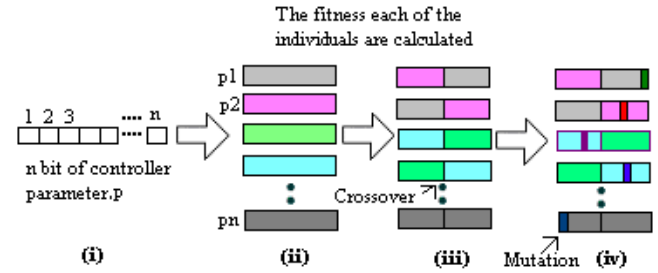


Fig.2 GA Optimisation technique

Where:

- i) Coding of controller parameters in a binary string
- ii) Create number of individuals with different and random values
- iii) New strings are produced via crossover in pairs
- iv) Once the crossover is implemented, mutation is applied to each new string according to a mutation rate

The control input of a PI controller can be represented as below:

$$u = K_I \int e(t) dt + K_c (-y(t)) \quad (18)$$

Where K_c is the proportional gain, K_I is the integral gain and e is the error of system output. In order to represent the controller parameter in a GA, binary vectors are used. The initial population may be generated randomly. Alternatively, a member of the population can be defined, for example, by using a Z-N tuned controller, DSM or the revised Direct Synthesis Method.

A. Objective Functions

In a natural system, species will evolve adapting themselves to the environment. In GA terms, the controller parameters, i.e. the species, will try to adapt to the objective functions, the environment. The objective functions that may be used in tuning the PI controller parameters in a system are in the form of a specified performance of the system. The objective function is the difference between the system actual output and the desired system output, as shown in Fig.3, when a unit step function applied to i th input, as equate in Eq 19.

$$J_{ij} = \int_0^t (y_{ij}(t) - f_{ij}(t))^2 dt \quad j=1,2,..m \quad (19)$$

Each objective function are weighted and added. Therefore, the final objective function is:

$$J = \sum_{i=1}^n w_{ij} J_{ij}$$

Where:

$y(t)$ = system output $f(t)$ = specified system output
 n = number of outputs
 m = number of different set points applied to system inputs
 w = weighting constant

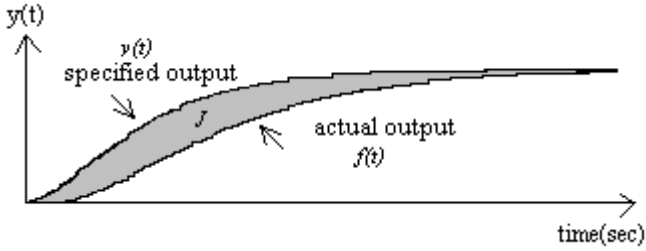


Fig.3 Error Signal

The aim is to minimize the value of J in order to obtain a good fitness value. However, since the GA is an optimisation tool, the fitness function will be calculated as:

$$\text{Fitness}, F = \frac{1}{J} \quad (20)$$

After a certain number of generations, fitness of the dominant string will be higher than the previous generation. Eventually, an optimal string will be obtained that contains the highest fitness value. This technique has the advantage obtaining acceptable system performance according to user specifications and the final controller parameters are in the standard form of the PI controller configuration.

Example 2.0

i) Using the same specifications in Example 1.0, the chromosomes are subjected to single point crossover at every generation with generation gap, crossover and mutation probabilities of 0.9, 0.5 and 0.03 respectively, with maximum generation of 100. The weighting constants are set as $w_{11}=w_{22}=1$, $w_{12}=w_{21}=0.1$. The initial individuals of the population are randomly selected. The selection for creating new offspring attained with stochastic universal sampling [1]. The reinsertions of the chromosomes back to the population are based on their rank fitness calculation. Also, at each generation, the best parent chromosome will replace the worst performed offspring chromosome. Each controller parameters are encoded using 20 bits resulting in total of 160 bits in a single string. This is to ensure the accuracy of conversion of the chromosomes into controller parameter values. The system simulations are implemented using the Runge-Kutta fifth order numerical integration with a constant step size of 0.1 second.

Graph in Fig.4 demonstrates that the objective function value converged less than 50 generation. The controller parameters obtained are:

$$K_c = \begin{bmatrix} 8.5584 & 3.6151 \\ -16.217 & 9.8889 \end{bmatrix} \quad K_i = \begin{bmatrix} 22.3517 & 14.6603 \\ -38.700 & 28.5651 \end{bmatrix}$$

ii) For the underdamped system problem similar GAs operators are applied. The controller parameters are:

$$K_c = \begin{bmatrix} 1.3858 & 0.5678 \\ -1.9231 & 1.8071 \end{bmatrix} \quad K_i = \begin{bmatrix} 6.0122 & 4.9719 \\ -1.8742 & 9.7241 \end{bmatrix}$$

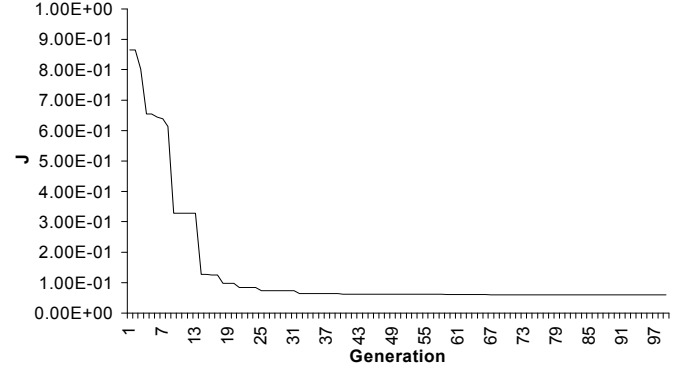


Fig.4 Best objective values for critical damped System

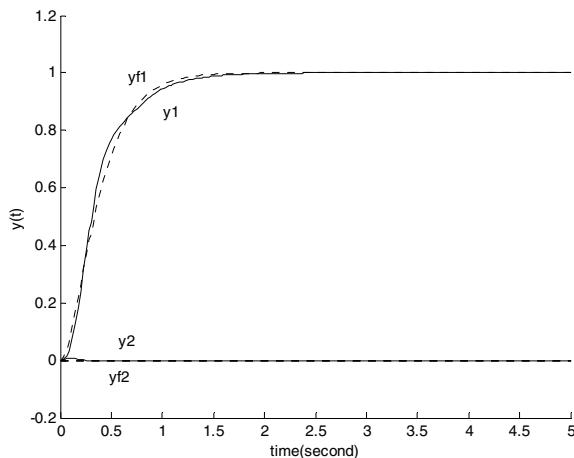
IV. DISCUSSION AND CONCLUSIONS

The outcomes of the system using PI controller to obtain the desired result in Example 1 using the mDSM, shows that the method achieved acceptable performance as illustrated in Fig 6 (a) & (b). However, when the system was required to attain a certain percentage of overshoot, the second output could not achieve the desired performance. Especially for the second loop output, the transient shoots up to almost 20% overshoot.

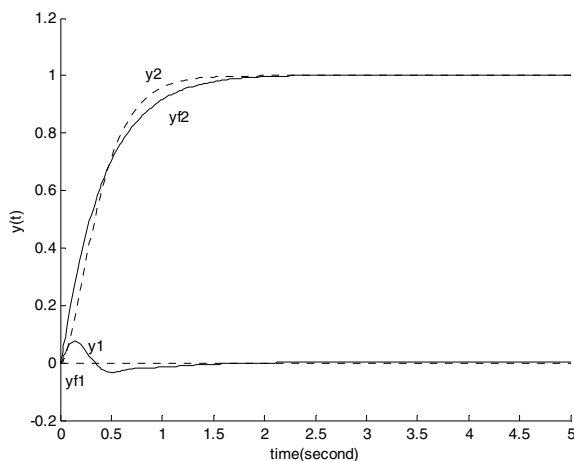
The GA was implemented to the same system and uses the same applications. The first exercise showed that GA managed to achieve the desired specifications and the controller parameters are in PI form. The best objective functions of each population converged in less than 50 generations. Also, the GA managed to produce a slightly better outcome of the second output than the system that uses the mDSM. While the second requirement in Example 2, GA performed better than the modified Direct Synthesis. The best objective function values for each generation converged as the 100 generations were reached.

In this paper, the comparisons were made between the mDSM and the GA method in tuning a decentralised PI for a multivariable system. The DSM works really well in a SISO system, the structure becomes complex in a MIMO system and the controller derived may not be realisable. The mDSM ensures a realisable and specific controller structure by substituting the s-terms in the analysis with a numerical value. The specifications transfer functions has to be chosen carefully to ensure the required controller form is achievable. GAs provides a flexible approach to controller design, for specification, controller structure and choice of cost function to achieve the objective. The differences between the desired output and the actual system outputs are

used within the desired objective function to enable the GA to search for the best possible controller parameters.

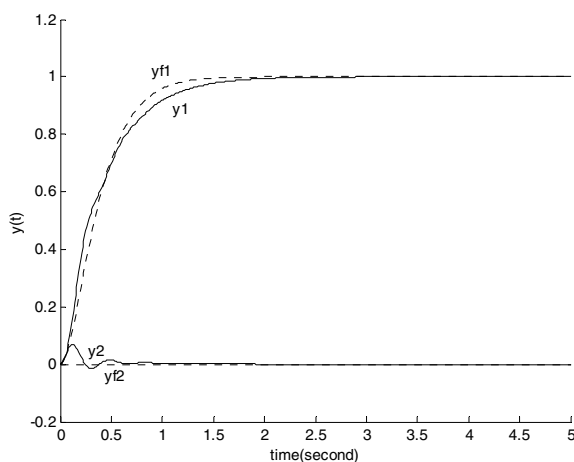


(a) Unit step signal as the first input

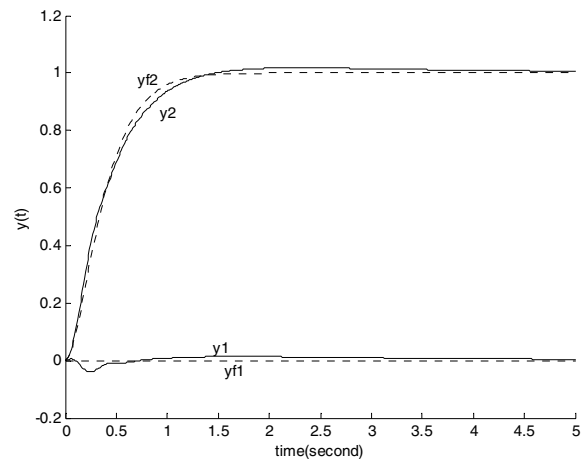


(b) Unit step signal as the second input

Fig. 6 Transient response for critical damped output



(a) Unit step signal as the first input



(b) Unit step input as the second input

Fig. 7 Transient response for critical damped output using GAs

VI. REFERENCES

- [1] J.E. Baker, Reducing Bias and Inefficiency in the Selection Algorithm, *Proceedings of the 2nd International Conference on Genetic Algorithms*, New Jersey, 1987.
- [2] K.A. DE JONG, An Analysis of the Behaviour of the Class of Genetic Adaptive Systems, *PhD thesis*, University of Michigan, 1975.
- [3] D. FRANKE, K. KRUGER, M. KNOOP, Systemdynamik und Reglerentwurf, R.Oldenbourg Verlag Munchen Wien, 1992.
- [4] **J.H. HOLLAND**, Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975.
- [5] MATLAB/SIMULINK Version 6.0.0.88, *The Language of Technical Computing, Release 12*, 2000.
- [6] R.V. Patel, N. MUNRO, Multivariable System Theory and Design, *International Series on Systems and Control*, vol(4), Pergamon Press, 1982.
- [7] D.E. Seborg, T.F. Edgar and D.A. Mellichamp, *Process Dynamics and Control*, 2nd ed. John Wiley and Sons, New York, 1989.
- [8] C. Vlachos, D. Williams and J.B. Gomm, Solutions to the Shell Standard Control Problem Using Genetic Algorithms, *Proceedings of the UKACC International Conference on Control*, UK, 1998.
- [9] C. Vlachos, D. Williams and J.B. Gomm, Genetic Approach to Decentralised PI Controller Tuning for Multivariable Processes, *IEE Proc.-Control Theory Appl.*, 1999.
- [10] P. Wang and D.P. Kwok, Auto-Tuning of Classical PID Controllers Using An Advanced Genetic Algorithm, *IEEE International Conference on Neural Network*, 1992.
- [11] J.G. Zeigler and N.B. Nichols, Optimum Settings for Automatic Controllers, *Trans ASME*, 1944.

LIDAR-based Wind Speed Modelling and Control System Design

Mengling Wang, Hong Yue, Jie Bao, William. E. Leithead

Department of Electronic and Electrical Engineering
University of Strathclyde, Glasgow, United Kingdom

E-mails: mengling.wang@strath.ac.uk; jie.bao@strath.ac.uk; hong.yue@strath.ac.uk; w.leithead@strath.ac.uk

Abstract—The main objective of this work is to explore the feasibility of using Light Detection And Ranging (LIDAR) measurement and develop feedforward control strategy to improve wind turbine operation. Firstly the Pseudo LIDAR measurement data is produced using software package GH Bladed across a distance from the turbine to the wind measurement points. Next the transfer function representing the evolution of wind speed is developed. Based on this wind evolution model, a model-inverse feedforward control strategy is employed for the pitch control at above-rated wind conditions, in which LIDAR measured wind speed is fed into the feedforward. Finally the baseline feedback controller is augmented by the developed feedforward control. This control system is developed based on a Supergen 5MW wind turbine model linearised at the operating point, but tested with the nonlinear model of the same system. The system performances with and without the feedforward control channel are compared. Simulation results suggest that with LIDAR information, the added feedforward control has the potential to reduce blade and tower loads in comparison to a baseline feedback control alone.

Keywords- wind turbine control; Light Detection And Ranging (LIDAR); disturbance rejection; feedforward control; wind speed evolution

I. INTRODUCTION

Advanced control is one of many options that can contribute to improved performance and decreased cost of wind energy production. High performance and reliable controllers could increase efficiency of power generation and reduce cost of operation and maintenance [1, 2]. In recent years, motivated by higher expectation of wind turbine performance, increased attention has been paid to new measurement technologies, among which LIDAR (Light Detection And Ranging) is able to provide the measurement of the wind upstream of the wind turbine and preview disturbance information. In the past decade, a number of wind turbine control strategies have been proposed, in which wind speed measurement are either provided or potentially provided by LIDAR.

During operation of wind measurement, a LIDAR emits a laser beam to the target wind field, and this laser beam is then backscattered by the small aerosols and particles in the wind field and then received by the LIDAR detector. The wind speed can therefore be calculated by employing the Doppler frequency shift between the two beams and the wavelength of the laser beams. With the help of preview wind measurement, feedforward control strategy is introduced into wind

turbine operations to reduce wind turbine structural loads [3, 4]. In some recent work, a feedforward channel is added to the baseline feedback control system. In this case, the feedforward controller can be designed independently of the feedback controller and will not affect the closed-loop stability. In [3], real LIDAR wind measurements information is used in wind turbine control systems instead of using an effective wind speed, where the results show reduction of tower and blade fatigue loads at high turbulent wind speeds. In [5], two feedforward controllers were designed to combine with two baseline feedback controllers, one applying model-inverse feedforward control for collective pitch control, and the other applying a shaped compensator for individual pitch control. Both of them enabled wind speed measurements that could be potentially provided by LIDAR as inputs to the feedforward controllers. An adaptive feedforward controller was proposed based on filtered-x recursive least algorithm [6].

Model predictive control (MPC) has proved to be an effective tool for multivariable constrained control systems, such as wind turbines. Henriksen *et al.* present the nonlinear MPC algorithm using future wind speeds in the prediction horizon [7]. In [8], an approach is proposed to deal with optimization problem of MPC. The nonlinear wind turbine model is linearised at different operating points, which are determined by the effective wind speed on the rotor disc. LIDAR wind speed measurement is used as a scheduling parameter.

While most of the research work concentrates on testing the load reduction performance by introducing LIDAR wind speed measurements, the energy capture performance of LIDAR-based control in below rated conditions was also investigated [9]. However, their results suggest that LIDAR-based control has limited improvements on energy capture. Therefore, applying LIDAR measurements in above rated pitch control could be more beneficial. The feasibility of applying LIDAR into wind turbine control systems needs further investigation. This motivates the work in the present paper. In this work, a wind evolution model is initially developed using the pseudo-LIDAR measurement data produced by Bladed, based on which a feedforward controller is designed and integrated to a baseline feedback controller.

The rest of the paper is organised as follows. In Section II, the details about the feedforward controller design are introduced. The wind speed evolution model is developed in Section III. Simulation studies are

conducted using an industrial-scale wind turbine model and the results are discussed in Section IV. The conclusions are summarized in Section V.

II. FEEDFORWARD CONTROLLER DESIGN

A. Feedback Baseline Controller

A standard baseline wind turbine controller normally consists of two parts. One is the torque controller which accounts for below rated operation, and the other is the pitch controller which accounts for above rated operation. In below rated conditions, torque demand is employed to ensure the tracking of the maximum power coefficient so that the maximum energy capture is achieved. In above rated conditions, pitch demand is employed to assure the generated power being maintained not to exceed its rated value, see [10] and [11]. The conventional feedback pitch control diagram is shown in Fig. 1, which is taken as the baseline controller.

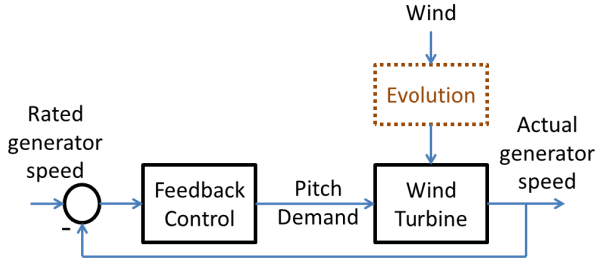


Figure 1. Block diagram of baseline feedback wind turbine control

B. Feedforward Controller

LIDAR is able to provide preview information of wind disturbances at various distances in front of wind turbines. This feature can be used in feedforward control to improve disturbance rejection. This research augments the feedback pitch controller with a feedforward control term (see Fig. 2) to alleviate turbine loads in above rated wind speed conditions. Fig. 2 shows a model-inverse-based strategy for designing feedforward controllers. The linear model-inverse feedforward controller is used to cancel the effect from the turbulence in wind speed on the wind turbine generator speed.

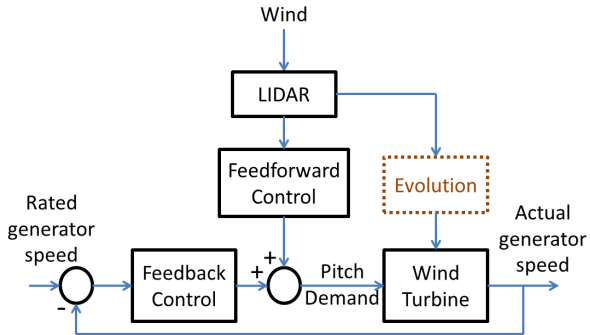


Figure 2. Combined feedback and feedforward control

Based on Fig. 1 and Fig. 2, a feedforward control scheme is developed and shown in Fig. 3. The primary control goal of the whole control system is to maintain the actual greater speed ω_{g_actual} at rated generator speed value ω_{g_rated} in the presence of varying wind v at above

rated conditions by adjusting the total pitch angle command β_c . v_T is the turbine wind speed, which indicates the wind speed approaching the turbine blades. v evolves to v_T on its way to the turbine and its variation disturbs the wind turbine system. The block P_E represents this evolution. The measurement of wind speed by a LIDAR sensor is v_L (line of sight wind speed). P_L is the LIDAR system transferring v to v_L . FB is the feedback controller and FF is the feedforward controller. The linear wind turbine model includes subsystems $P_{\omega_e \beta_c}$ and $P_{\omega_e v_t}$. $P_{\omega_e \beta_c}$ maps collective blade pitch error β to generator speed error ω_e ($\omega_{g_actual} - \omega_{g_rated}$) and $P_{\omega_e v_t}$ maps v_T to ω_e . The output of feedforward controller β_{FF} is added to the feedback pitch angle β_{FB} of collective pitch feedback controller.

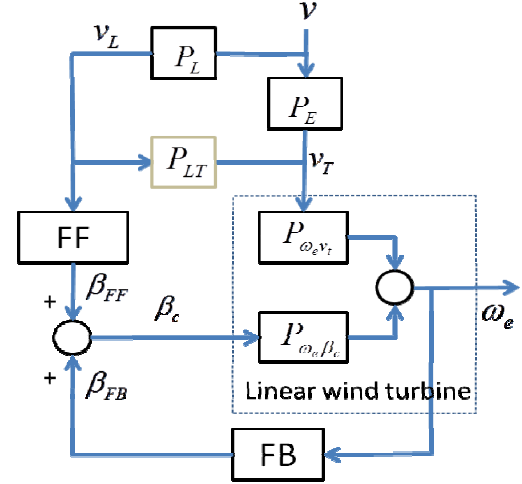


Figure 3. Feedforward control scheme

Following the control strategy in Fig. 3, we have

$$\omega_e = (v \cdot P_L \cdot FF + \omega_e \cdot FB) \cdot P_{\omega_e \beta_c} + v \cdot P_E \cdot P_{\omega_e v_t} \quad (1)$$

Since it is expected that the tracking error of generator speed should be zero, i.e., $\omega_e = 0$ [6],

$$v \cdot P_L \cdot P_{\omega_e \beta_c} \cdot FF = -v \cdot P_E \cdot P_{\omega_e v_t} \quad (2)$$

The feedforward controller is solved as

$$FF = -P_{\omega_e \beta_c}^{-1} \cdot P_{\omega_e v_t} \cdot P_E \cdot P_L^{-1} \quad (3)$$

where $P_{\omega_e \beta_c}^{-1}$ and $P_{\omega_e v_t}$ can be obtained from turbine modelling, but wind evolution P_E and LIDAR system P_L are very complex and difficult to model. In this research, the transfer function between v_L and v_T , which is $P_E \cdot P_L^{-1}$ is approximated by a transfer function

$$P_{LT}(s) = \frac{S_{LT}(s)}{S_{LL}(s)} \quad (4)$$

where S_{LT} is the cross spectrum between the LIDAR measurements and the turbine wind speed, S_{LL} is the auto spectrum of the LIDAR measurements across the distance from the measurement point to wind turbine blades.

The feedforward controller is then written as

$$FF = -P_{\omega_e \beta_c}^{-1} \cdot P_{\omega_e v_t} \cdot P_{LT} \quad (5)$$

It is remarkable that the non-minimum phase zeros contained in $P_{\omega_e\beta_c}$ would become poles that cause the system to be unstable after inverting. Therefore, a stable approximation should be used instead of the exact inverse of $P_{\omega_e\beta_c}$. Related work will be introduced later.

In the next section, the evolution of LIDAR measurements across the distance from the measurement point to wind turbine blades is developed. The cross spectrum between turbine wind speed and LIDAR measurements and the auto spectrum of LIDAR measurements are calculated.

III. WIND SPEED EVOLUTION MODELLING

A. LIDAR Wind Speed Simulation

According to the feedforward control loop in Fig. 3, LIDAR measured wind speed is fed into the feedforward controller. In this work, the simulated LIDAR measurements are used in modelling. Taylor's frozen turbulence hypothesis is employed, which assumes that the turbulent wind field is unaffected when approaching the turbine and moving with average wind speed.

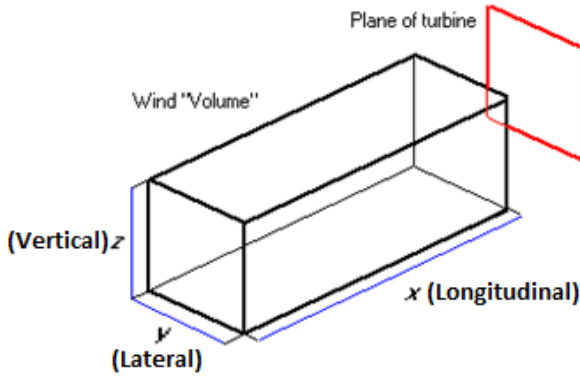


Figure 4. 3D wind volume simulated by Bladed

The continuous LIDAR shoots a continuous beam of light to the atmosphere and again all particles in the atmosphere, along the light of signal of the beam will reflect some of this light. These systems use the same frequency shift in the reflection, to determine the velocity of the particles. Each wind measurement is a vector with components at different directions. Here, only the component of the wind vector in laser beam direction (line of sight) wind speed is detected. It is constructed by averaging over the circular trajectory [12].

TABLE I. CHARACTERISES OF 3D TURBULENCE

Turbulence length scales		Component of turbulence		
		Longitudinal	Lateral	Vertical
Along x	m	244.671	58.8681	21.1883
Along y	m	79.6605	76.6657	13.7971
Along z	m	59.2508	28.5117	20.5243
Turbulence intensities:	%	16.0108	12.5465	8.92472

Bladed uses a 3-dimensional turbulent wind field with defined spectral and spatial covariance characteristics to represent real atmospheric turbulence. This option will give the most realistic predictions of loads and performance in normal conditions. In Bladed, wind speed is displayed as a vector of 3 components: Longitudinal component $x(t)$, Lateral component $y(t)$ and Vertical component $z(t)$, see Fig. 4. The parameters set up for the wind field simulation are listed in Table I. In this work, rectangular scan circle is used instead of round circle.

According to the scan principle of LIDAR instrument and Bladed display of wind field in Fig. 4, $v_T(t_T, x_T)$ and $v_{Li}(t_{Li}, x_{Li})(i = 1, \dots, 6)$ are points sampled from Bladed. Wind field in the x direction to represent the turbine wind speed and LIDAR measurements. $v_T(x_T = 0m)$ is assumed as the turbine wind speed, which is the mean wind fluctuation over the turbine plane. $v_{Li}(t_{Li}, x_{Li})(i = 1, \dots, 6)$ are assumed as LIDAR measurements with preview distances of 30m, 60m, 90m, 120m, 160m and 190m respectively ($x_{L1} = 30m, x_{L2} = 60m, x_{L3} = 90m, x_{L4} = 120m, x_{L5} = 160m, x_{L6} = 190m$). Each wind speed is the mean wind fluctuation over an area lying in the y-z plane. The distribution of all wind speed points is shown in Fig. 5.

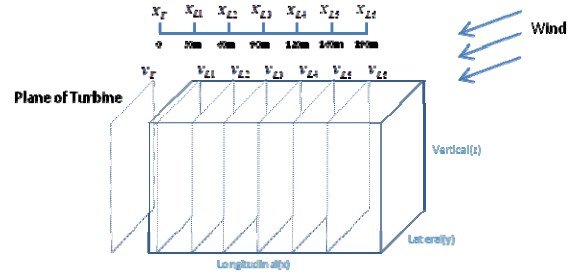


Figure 5. Distribution of the turbine wind speed and LIDAR measurements

In Fig. 5, the wind speed evolves in the direction of distance x . There is no cross correlation between $v_T(t_T, x_T)$ and $v_{Li}(t_{Li}, x_{Li})(i = 1, \dots, 6)$. Therefore, the strategy of sampling points in the wind field is needed to be modified.

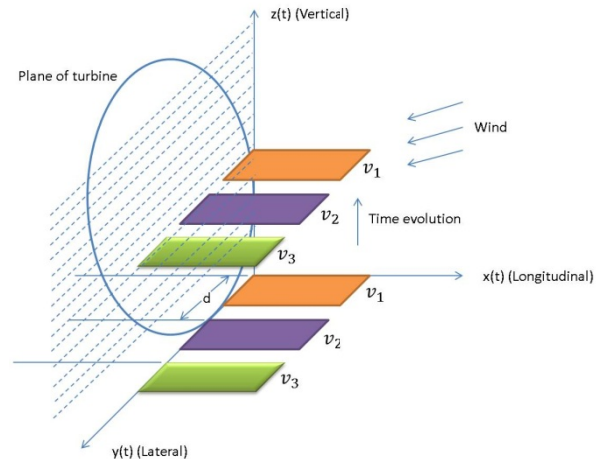


Figure 6. Schematics of wind speed sampling in Bladed

As the Bladed wind field is frozen and isotropic, the variation of the wind speed fluctuation in the x direction at a point can be equally represented by the component in the y direction as well as the component in the z direction. Hence, as depicted in Fig. 6, the cross correlation between the mean wind speed fluctuations over two areas displaced by a distance d in the x direction can be estimated from the areas lying in x - y planes separated by a distance of d in the y direction and the fluctuations in the z direction. Time evolution can be represented by moving the plates through the wind field in the z direction. Hence, the distribution of turbine wind speed and LIDAR measurements can be modified from Fig. 5 to Fig. 7. $v_T(t_T, x_T)$ and $v_{Li}(t_{Li}, x_{Li})(i = 1, \dots, 6)$ represent mean wind fluctuations in x - y plane, with time evolution in the z direction.

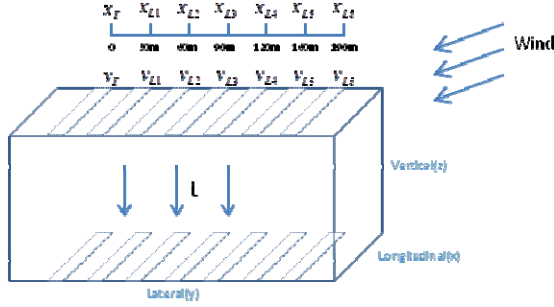


Figure 7. Modified distribution of the turbine wind speed and LIDAR measurements

B. Cross Spectrum of Wind Speed

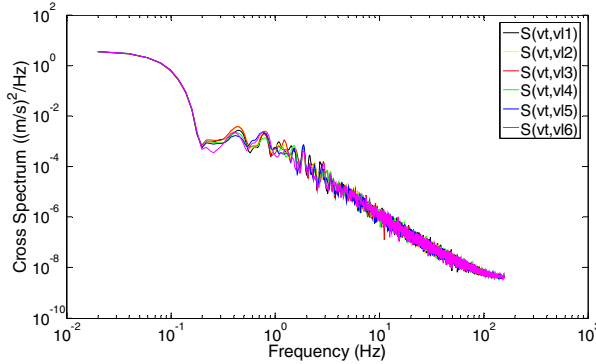


Figure 8. Cross spectrum of the turbine wind speed and LIDAR measurements

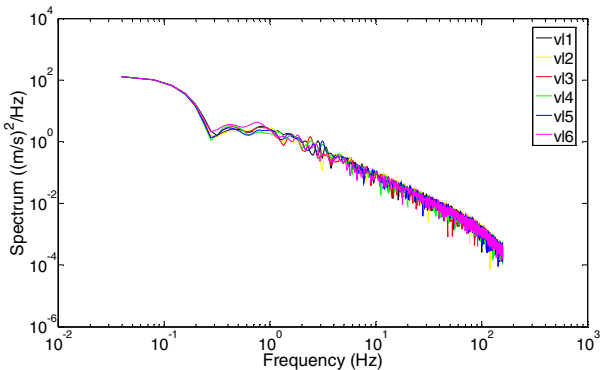


Figure 9. Auto spectrum of LIDAR measurements

Fig. 8 shows the cross spectrum of the turbine wind speed v_T and LIDAR measurements. The auto spectrum of $v_{Li}(t_{Li}, x_{Li})(i = 1, \dots, 6)$ can be seen in Fig. 9.

C. Transfer Function

Following the results of cross spectrum S_{LT} and auto spectrum S_{LL} , the transfer function P_{LT} in equation (4) is approximated by a first order low-pass filter using system identification method, see [13] for details.

$$P_{LT}(s) = -\frac{8.102 \times 10^{-7}}{s + 0.02831} \quad (6)$$

The transfer function model P_{LT} has been validated against cross spectrum $S(v_t, v_{L1})$, see Fig. 10. It can be seen that the shapes of the simulated model output and the measured model output match reasonably well. Therefore, the transfer function is acceptable for the controller design.

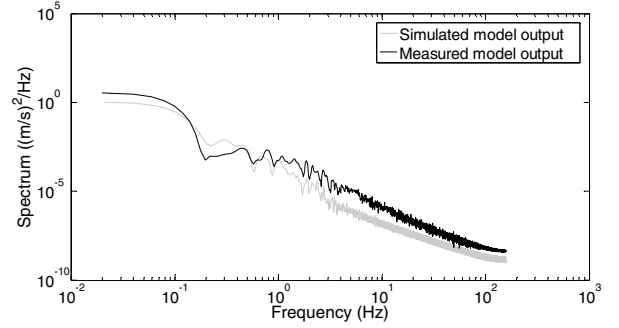


Figure 10. Comparison of model output from (5) and 'measured' model output

IV. SIMULATION STUDY

In this work, the simulation study is implemented using the Supergen 5MW exemplar wind turbine model developed in the University of Strathclyde. This is a non-linear model that is constructed in Simulink. It contains 3 main parts, the pitch mechanism, the aero-rotor and the drive train model. The main turbine parameters are listed in Table II. More details can be found in [9, 14]. A matched 5 MW Supergen feedback controller is used here as the baseline controller.

TABLE II. WIND TURBINE PARAMETERS [14]

Turbine parameters	
Rotor radius [m]	63
Effective blade length [m]	45
Hub height [m]	90
Maximum generator speed in generation mode [rad/s]	120
Cut in wind speed [m/s]	4
Cut out wind speed [m/s]	25
Nominal generator torque [Nm]	46372.7
Air density [kg/m^3]	1.225
Gearbox ratio	97

Considering one input and output in each case, the nonlinear wind turbine model can be linearised at the

operating point, and a linearised state-space model is then produced involving 11 state variables. This state-space model can be further written as a continuous transfer function model. The simulation in this work is conducted at 16m/s mean wind speed and the wind speed fluctuation are modeled by a set of small steps added to the mean wind speed. The transfer functions, $P_{\omega_e \beta_c}$ and $P_{\omega_e v_t}$ are obtained by discretization of the two continuous transfer function models, respectively, with a sampling rate of 0.0125s.

The transfer function between the generator speed error to the pitch demand is developed

$$P_{\omega_e \beta_c}(z) = \frac{B(z)}{W_1(z)} \quad (7)$$

in which

$$B(z) = -1.1095(z^2 - 1.9998z + 0.902) \\ (z^2 - 2.0502z + 1.0546) \\ (z^2 - 2z + 1.0005)(z^2 - 1.9996z + 1) \\ (z + 1.1532)(z - 0.2523)$$

$$W_1(z) = (z^2 + 0.3156z + 0.624) \\ (z^2 - 1.9796z + 0.9922) \\ (z^2 - 1.9528z + 0.9570) \\ (z^2 - 2.0022z + 0.9937) \\ (z^2 - 1.9918z + 1.0067) \\ (z - 0.9942)$$

The transfer function between the generator speed error to the approaching blade wind speed is written as

$$P_{\omega_e v_t}(z) = \frac{V(z)}{W_2(z)} \quad (8)$$

with

$$V(z) = (2.5921 \times 10^{-5})(z^2 - 1.9948z + 1.0004) \\ (z^2 - 1.9992z + 0.9997) \\ (z^2 - 1.9992z + 0.9997) \\ (z + 9.1961)(z + 1.0895)(z - 0.2666) \\ (z + 0.1549)$$

$$W_2(z) = (z^2 + 0.3156z + 0.624) \\ (z^2 - 1.9796z + 0.9921) \\ (z^2 - 1.9528z + 0.9562) \\ (z^2 - 2.0022z + 1.0028) \\ (z^2 - 1.9918z + 0.9923) \\ (z - 1.0002)$$

The feedforward controller is obtained from (5) that gives

$$FF(z) = -\frac{W_1(z)V(z)P_{LT}}{B(z)W_2(z)} \quad (9)$$

It can be seen that the developed feedforward controller is of a high order which is inconvenient for tuning. This model is therefore firstly reduced by non-minimum phase zeros ignore (NPZ-Ignore) technique to remove non-minimum phase zeros, and further reduced to 3rd-order controller as shown in (10) via approximation fitting [15]. In order to fine tune the reduced-order controller, a tuning factor, k_{FF} , is introduced in the transfer function of FF . This tuning function can also address modelling uncertainty to some extent. In this work, the designed value for k_{FF} is 2.336×10^{-5} to start with. The fine-tuned setting is $k_{FF} = 2 \times 10^{-4}$,

$$FF(z) = k_{FF} \times \frac{(z+9.1961)(z-0.2666)(z+0.1549)}{z^3} \quad (10)$$

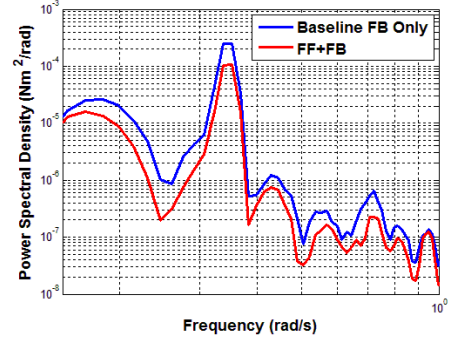


Figure 11. Comparison of the pitch angle before and after the addition of the feedforward controller

As shown in Fig. 11, with the feedforward controller, a decrease in the pitch angle demand power spectral density (PSD) is achieved. The decrease not only saves the driving energy but also helps to expand lifetime of pitch actuators.

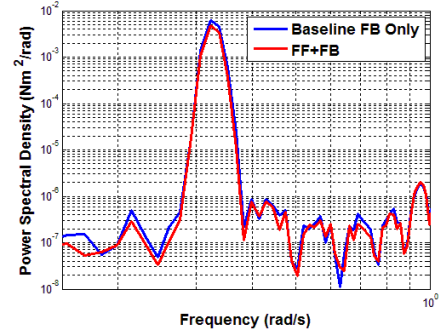


Figure 12. Comparison of the tower acceleration before and after the addition of the feedforward controller

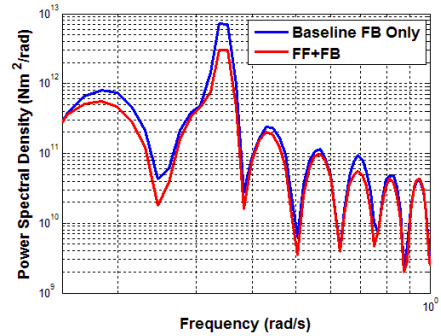


Figure 13. Comparison of the out-of-plane rotor torque before and after the addition of the feedforward controller

Compared with the baseline feedback control alone, reductions of the tower fore-aft acceleration and out-of-plane rotor torque PSD can be seen in Fig. 12 and 13 for the proposed controller. With these improvements, the oscillation of the tower and the load on the rotor are reduced and thereby the lifetime of the tower and rotor components could be expanded. Moreover, the loads that propagate from tower and rotor to drive train can also be alleviated.

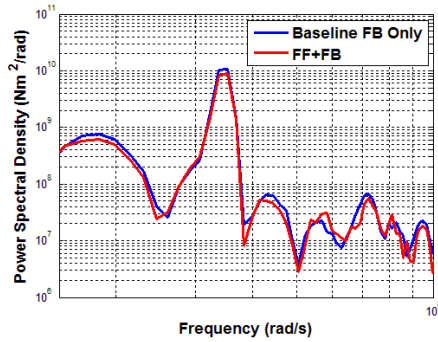


Figure 14. Comparison of the generator power before and after the addition of the feedforward controller

The comparison is also made for the generated power, as shown in Fig. 14. It can be seen that with or without the feedforward control channel, there is no clear difference between the power generated. This indicates that by introducing the feedforward controller for disturbance rejection, the power generation performance can still be maintained.

V. CONCLUSIONS

In this paper, the control strategy for designing LIDAR-based feedforward controllers has been presented. A model-inversed feedforward pitch controller is combined with the baseline feedback pitch controller of the Supergen 5MW wind turbine model.

The LIDAR measurements and the turbine wind speed are simulated by Bladed. The cross spectrum between them and the auto spectrum of LIDAR wind speed measurements are studied. The results are used to develop the transfer function representing the evolution of LIDAR measurements across the distance from the measurement point to the wind turbine blades. At last, the feedforward controller is designed based on the Supergen linear wind turbine model and the transfer function.

The performance of the feedforward controller is evaluated at wind speed 16m/s. The Simulink simulation study shows that the feedforward/feedback controller has achieved improved performance on reducing the fluctuations of the pitch angle demand, tower acceleration and the out-of-plane torque of the rotor without degrading the energy capture seriously. It should be noted that in above-rated operating conditions, the main function of the baseline feedback pitch control is to maintain the generated power at rated power level. For large scale wind turbines, load reduction is another crucial control target, which can be handled by introducing extra feedback loops and/or feedforward channels. In our work, we consider a feedforward controller mainly to take advantages of the more comprehensive LIDAR measurement information, which should help to reduce the effects of disturbance brought by the wind speed uncertainty. In fact, the feedforward controller design can be regarded as independent of the feedback controller design, which means the feedback control performance won't be deteriorated by the feedforward channel.

In our future work, the performance of feedforward controller and other LIDAR-based control strategies will be evaluated by calculating Damage Equivalent Loads (DEL), which can give more apparent comparisons between different control options in load reduction of large-scale wind turbine systems.

REFERENCES

- [1] T. Burton, D. Sharpe, N. Jenkins, and E. Bossanyi, *Wind energy handbook*: John Wiley & Sons, 2001.
- [2] F. D. Bianchi, H. De Battista, and R. J. Mantz, *Wind turbine control systems: principles, modelling and gain scheduling design*: Springer Science & Business Media, 2006.
- [3] D. Schlipf, D. J. Schlipf, and M. Kühn, "Nonlinear model predictive control of wind turbines using LIDAR," *Wind Energy*, vol. 16, pp. 1107-1129, 2013.
- [4] D. Schlipf, P. Fleming, F. Haizmann, A. Scholbrock, M. Hofsaß, A. Wright, et al., "Field testing of feedforward collective pitch control on the CART2 using a nacelle-based lidar scanner," in the *Proceedings of The Science of Making Torque from Wind 2012*, Oldenburg, Germany
- [5] L. Y. Pao, F. Dunne, A. D. Wright, B. Jonkman, N. Kelley and E. Simley, "Adding feedforward blade pitch control for load mitigation in wind turbines non-causal series expansion, preview control, and optimized FIR filter methods," Technical Report, University of Colorado, Boulder, CO, USA 2011.
- [6] L. C. Henriksen, N. K. Poulsen, and M. H. Hansen, "Nonlinear model predictive control of a simplified wind turbine," in *18th World Congress of the International Federation of Automatic Control*, 2011, pp. 551-556.
- [7] M. Mirzaei, M. Soltani, N. K. Poulsen, and H. H. Niemann, "An MPC approach to individual pitch control of wind turbines using uncertain LIDAR measurements," in *2013 European Control Conference (ECC)*, Zurich, Switzerland, 2013, pp. 490-495.
- [8] N. Wang, "LIDAR-assisted feedforward and feedback control design for wind turbine tower load mitigation and power capture enhancement," PhD Thesis, Colorado School of Mines, 2013.
- [9] W. E. Leithead and B. Connor, "Control of variable speed wind turbines: dynamic models," *International Journal of Control*, vol. 73, pp. 1173-1188, 2000.
- [10] A. P. Chatzopoulos, "Full envelope wind turbine controller design for power regulation and tower load reduction," PhD Thesis, University of Strathclyde, 2011.
- [11] N. Wang, K. E. Johnson, and A. D. Wright, "FX-RLS-based feedforward control for LIDAR-enabled wind turbine load mitigation," *Control Systems Technology, IEEE Transactions on*, vol. 20, pp. 1212-1222, 2012.
- [12] D. Schlipf, "LIDAR assisted collective pitch control," Technical Report, University of Stuttgart, 2011.
- [13] M. Wang, "Feedforward wind turbine controller design using LIDAR," Master Thesis, University of Strathclyde, April, 2015.
- [14] A. Stock, "Guide to the Supergen controllers," Technical Report, University of Strathclyde, 2014
- [15] J. Butterworth, L. Y. Pao, and D. Y. Abramovitch, "The effect of nonminimum-phase zero locations on the performance of feedforward model-inverse control techniques in discrete-time systems," *American Control Conference*, pp. 2696-2702, 2008. IEEE, 2008.

Design of Configurable DC Motor Power-Hardware-In-the-Loop Emulator for Electronic-Control-Unit Testing

Chalupa, J., Grepl, R., Sova, V.

Mechatronics laboratory (www.mechlab.cz)

Brno University of Technology, Faculty of Mechanical Engineering

Brno, Czech Republic

chalupa@fme.vutbr.cz

Abstract—Using Hardware-In-the-Loop (HIL) or extended Power-Hardware-In-the-Loop (PHIL) tools is crucial in rapid prototyping of electronic devices used in automotive or aerospace applications. Flexible tool for Electronic-Control-Units (ECU) testing enables user to test control algorithms during development or ECU quality check in manufacturing process. This paper is concentrated on design of cost-effective PHIL device, which will replace real DC motor component in the ECU testing chain. To obtain high quality substitute of the real DC motor it is necessary to create Real-Time model, which simulates characteristic behaviour of the real DC motor, like a motor/generator mode, friction modelling, current ripple caused by commutator, etc. Function is demonstrated on a fuel pump module where real DC motor is replaced with the PHIL Real-Time simulator.

Keywords- *Power-Hardware-In-the-Loop; DC motor, Electronic-Control-Unit; real-time simulation; friction; commutator current ripple.*

I. INTRODUCTION

Design of a complex control system can be accelerated and verified in many ways. Model-Based-Design (MBD) approach [1], where a system is completely described from mathematical, electrical and mechanical point of view, every part of the model may affect another. This offline simulation and verification method can take into consideration nearly all effect, which usually cannot be considered in Real-Time (RT) simulation due to high computational demands [2]. Another approach is HIL testing, where system under test is coupled to the RT testing hardware only with signal interface. The RT HIL system contains a model of simulated plant including actuators, sensors, digital interfaces etc. [3]. Plant model is mostly simplified to achieve hard-RT system requirements [2]. Adding power electronics and real sensors to a standard RT HIL simulator leads to RT PHIL topology [3], which allows energy exchange between simulator hardware and system tested hardware.

The RT PHIL systems are useful for critical state simulations. For example partial or fatal failure, overheat, fast-dynamic transients, resonance etc. Those states are hard (or nearly impossible) to achieve with a real component or may lead to permanent damage. All these

states can be triggered and controlled from a software interface.

DC motors are widely used in many applications [4] due to their simplicity and reliability. ECU for DC motor drives or actuators may become very complex [6], depending on target application and customer requests. The drives are usually requested without any additional sensing elements, to meet automotive and aerospace requirements like robustness, durability and cost-efficiency. Feedback is derived from the system known behaviour, for example velocity feedback from the DC motor can be obtained via back-EMF voltage measurement or current ripple analysis [7][8], so that there is no need for additional sensors like an encoder or a dynamometer, etc. To obtain correct information from measurements some Digital Signal Processing (DSP) algorithms have to be implemented in the ECUs.

Nowadays automotive or aerospace ECUs for the electric drives control are usually small and highly integrated devices, which contains simple sensors, control circuitry and power circuitry in one package [9]. These packages cannot be disassembled in order to gain access to signal traces, therefore standard HIL testing approach is not applicable [3]. Usage of real electro-mechanical components, with moving parts, may be unwanted due to mechanical degradation of the component or due to critical states emulation inability [10]. For these reasons the PHIL emulator appears to be a suitable solution.

Objective of this paper is to describe the design and verification of a cost-effective and configurable PHIL DC Motor Emulating Unit (MEU) based on dSPACE Autobox RT platform, which is programmable from MATLAB/Simulink environment and controlled by ControlDesk GUI.

MEU is a nonlinear RT system that emulates the behaviour of a brushed DC motor including various types of friction, current ripple, mechanic load, etc. The emulator can be used during development process of a complex ECU or in the final manufacturing process to test the ECU functions and quality.

II. MECHANIC AND ELECTRONIC ECU TESTING APPROACHES

A. Classic mechanical approach

A typical mechanical test-bench can be constructed in many ways. The ECU under test controls real DC motor which is mechanically coupled with mechanic load.

The mechanical load may be variable or constant, depending on user needs. Controlling the actual value of speed or torque is not simple in some cases. The best controllability is achieved by using a motor to motor/generator topology (Fig.1.), where two machines are bound together by a shaft. The second machine acts like a variable load/source [10].

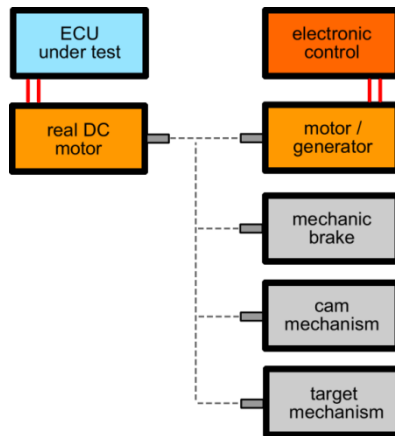


Figure 1. Schematic of a mechanical ECUs testing topology

Other types of mechanical loads (Fig.1.) are less controllable and usually can simulate only particular phenomenon. Disadvantage of the mechanical testing topology is that parameters of the used parts change in the course of time or degrade due to frequent usage. The characteristic parameters are given by mechanical construction and usually invariable, so that usage of the particular test-bench is very limited for other testing applications. From that point of view, PHIL testing approach is very flexible solution.

B. HIL testing approach

Typical HIL system consists of signal peripheral hardware and RT computational core, which runs the plant, actuator and sensor simulation models. Interface with other devices is done via signals only [3]. The peripherals collect signals like PWM, voltage levels, digital communication, etc. from the output of the ECU and present them as inputs to the RT models of actuators or power electronics [6][10]. Output of the actuator model affects states of the virtual plant model. The simulation may also include sensors models, which can generate signals according to the state of the virtual plant. These signals can be sent to the output peripherals of the HIL simulator, like a virtual feedback for the real ECU inputs. Various virtual sensors can be implemented in the RT model: quadrature encoder, potentiometer, dynamometer, digital interface sensors etc.

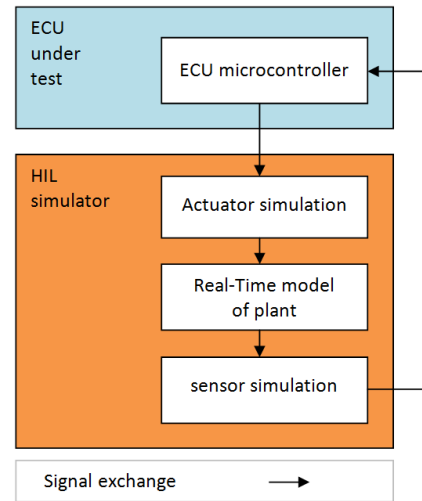


Figure 2. Schematic of a common HIL simulator unit

The HIL simulator can act as an autonomous device or as a slave device connected to a superior system (mostly PC), which can directly affect states of a virtual plant and states of other elements in simulation. User can easily setup parameters, simulate failures, apply variable load etc.

C. PHIL testing approach

Extending the common HIL simulator by real sensors and by power electronics gives us a complex system, which can be used as a controlled nonlinear electrical load/source [11] with predefined or user determined behaviour. The MEU enables user to easily substitute the real DC motor without any radical hardware changes [3].

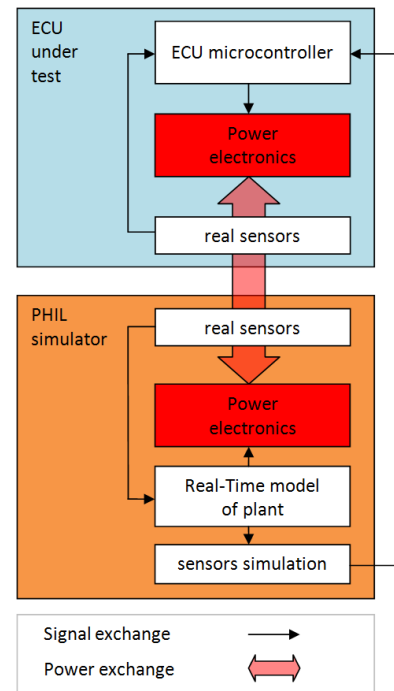


Figure 3. Schematic of a common PHIL simulator unit

III. MODEL BASED DESIGN OF DC MOTOR

A. Mathematical model

The plant model is based on DC motor extended equations, which are described below. Extending the basic equations (1) and (2) with friction and current ripple, we can obtain more accurate model of the plant, described by equations (3) and (4).

$$v_{input} = R_a i_a + L_a \frac{di_a}{dt} + v_{emf} \quad (1)$$

$$t_{emf} = J \frac{d\omega_r}{dt} + t_{load} \quad (2)$$

R_a	winding resistance	[Ω]
L_a	winding inductance	[H]
J	rotor inertia	[kg.m ²]
t_{emf}	EMF torque	[N.m]
t_{load}	total torque load	[N.m]
ω_r	rotor velocity	[rad.s ⁻¹]
v_{input}	input voltage	[V]
v_{emf}	back EMF voltage	[V]
i_a	armature current	[A]

$$v_{input} = R_a i_a + L_a \frac{di_a}{dt} + k_\phi \omega_r + f_{ripple}(\omega_r) \quad (3)$$

$$k_\phi i_a = J \frac{d\omega_{rotor}}{dt} + t_{extern} + t_{fr}(\omega_r) \quad (4)$$

k_ϕ	back EMF constant	[Nm.A ⁻¹]
$f_{ripple}(\omega_r)$	current ripple function	[V]
t_{extern}	external load	[N.m]
$t_{fr}(\omega_r)$	friction load function	[N.m]

Equations (3) and (4) are implemented in the MBD plant model which, is used for offline simulation and verification. The RT simulation plant model is based on equation (4).

B. Friction modeling

A friction model is included in the main plant model via $t_{fr}(\omega_r)$ function. Parameters can be set according to user needs. Three kinds of friction (Fig.4.) are simulated [12-14].

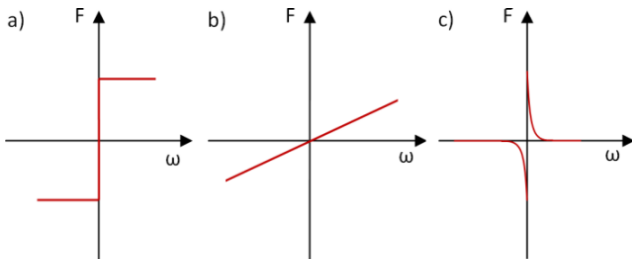


Figure 4. Friction model a) Coulomb, b) viscous, c) static

Equation of $t_{fr}(\omega_r)$ function is simplified and used in form (5), saturation output is in (0-0.9) interval.

$$t_{fr}(\omega_r) = B\omega_r + \text{sgn}(\omega_r)S_{fr} \frac{1}{0.1 + \text{sat}(\text{abs}(\omega_r))} \quad (5)$$

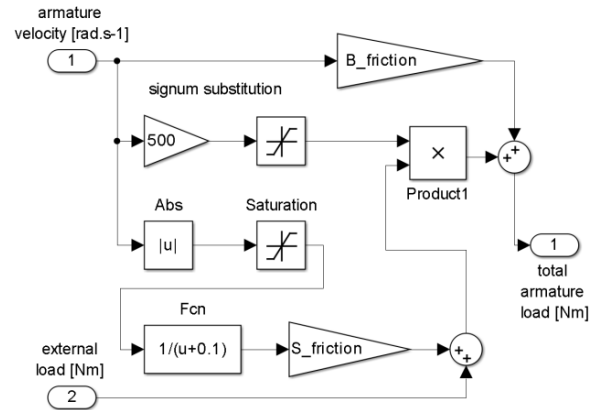


Figure 5. Simulink total armature torque load calculation block based on equation (5)

Fig.5. describes the Simulink model of total torque calculation depending on the armature velocity and external load. The result of the simulation is shown in (Fig.6.). Simulated friction curve is close to the Stribeck friction characteristic.

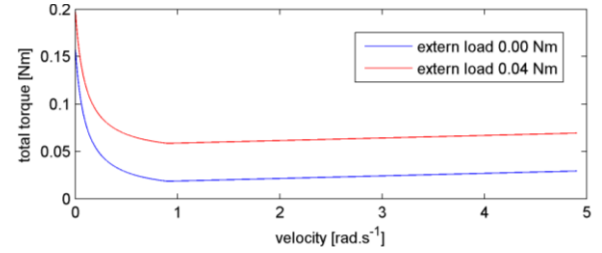


Figure 6. Total torque (friction+load) dependance on rotor velocity

Friction effect is dominant by velocities close to zero, when two surfaces are still and stucked together (Fig.4.c). Increasing the velocity saturates the $S_{friction}$ gain output at static value of the Coulomb friction (Fig.4.a). Linear contribution of viscous friction (Fig.4.b) is notable in (Fig.6.).

C. Current ripple

Many ECUs use algorithms and methods to obtain a velocity feedback from armature current waveform [8]. Due to commutation effect, low frequency current ripple waveform is superposed on the mean value of the measured current (Fig.7.).

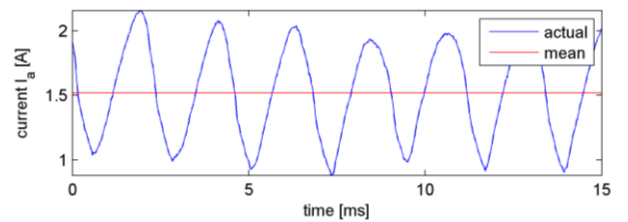


Figure 7. Current ripple measured on immersed fuel pump with open valve

Current ripple generator block is very simple. Its structure is described in (Fig.9.). The ripple coefficient determines the amplitude of the current ripple. The ripple coefficient has to be defined experimentally. Segment coefficient is set according to construction properties (number of brushes and segments) equation (6). The output is the current ripple superposed on current in the circuit.

The block diagram illustrates the current ripple control logic. It features several input signals: 'segment coefficient [-]' (value 3), 'omega [rad.s-1]' (value 1), and 'ripple coefficient [-]' (value 2). The logic proceeds as follows:

- The 'segment coefficient' (3) is multiplied by the output of the 'Discrete-Time Integrator' to produce 'Product1'.
- The 'omega' (1) signal is fed into an 'Interval Test' block, which also receives feedback from the 'current ripple [A]' output.
- The 'Interval Test' block outputs a signal to an 'OR' block (labeled 'Logical Operator1') and an 'Abs' block (labeled '|u|').
- The 'Abs' block output is multiplied by a 'Gain' of 1/200.
- The output of the gain block is added to the 'ripple coefficient' (2) at a summing junction (labeled '++').
- The result of the summing junction is fed back to the 'Interval Test' block and also passes through a 'Unit Delay' block (labeled $\frac{1}{z}$).
- The 'Unit Delay' block output is compared to a constant value of 6.2832 in a 'Compare To Constant' block.
- The 'Compare To Constant' block output is fed into the 'OR' block and also into the 'Discrete-Time Integrator'.
- The 'Discrete-Time Integrator' (labeled $\frac{K Ts}{z-1}$) receives inputs from 'Product1' and the 'Compare To Constant' block.
- The output of the 'Discrete-Time Integrator' is passed through a 'Trigonometric Function' block (labeled 'cos').
- The output of the 'Trigonometric Function' block is multiplied by the 'current ripple [A]' (value 1) at a final summing junction (labeled 'x') to produce the final 'current ripple [A]' output.

D. Final mathematical model

The diagram illustrates the electro-mechanical model, showing the interaction between the motor's electrical and mechanical components. It includes the following blocks and signals:

- Inputs:**
 - mechanical_load** (green box): load applied to axis.
 - U** (green box): input voltage.
 - C-** (yellow box): RIPPLE SET.
 - C-** (yellow box): SEGM. COEFF.
- Simple Ripple Generator:**
 - Inputs: RIPPLE SET, SEGM. COEFF.
 - Outputs: ripple coefficient [-], segment coefficient [-], current ripple [A], armature velocity [rad.s⁻¹].
- External Load:**
 - Input: mechanical_load.
 - Output: external load [Nm].
- Total Armature Load Calculation:**
 - Inputs: external load [Nm], current ripple [A].
 - Output: total armature load [Nm].
- Motor R L Circuit Model:**
 - Inputs: input voltage U, armature velocity [rad.s⁻¹].
 - Output: armature current [A].
- Electro-mechanical Model:**
 - Inputs: total armature load [Nm], armature current [A].
 - Output: armature velocity [rad.s⁻¹].

The diagram shows a feedback loop where the armature velocity from the electro-mechanical model is fed back into the simple ripple generator and the motor R L circuit model. The armature current from the motor R L circuit model is fed into the total armature load calculation, which also receives input from the current ripple. The total armature load calculation then feeds into the electro-mechanical model, which produces the armature velocity.

E. Real fuel pump measurement and simulation

Figure 10 consists of four subplots showing the comparison of motor simulation (blue line) and real motor measurement (red line) for armature current over time. The subplots are arranged vertically and show different time intervals.

- Top Subplot:** Shows the armature current [A] versus time [ms] from 0 to 45 ms. The current starts at 0, rises to a peak of approximately 7 A at 2 ms, and then decays with oscillations, settling around 2 A by 45 ms.
- Second Subplot:** Shows the armature current [A] versus time [ms] from 110 to 140 ms. The current exhibits high-frequency oscillations between approximately 1 A and 2.2 A.
- Third Subplot:** Shows the armature current [A] versus time [ms] from 0 to 45 ms. The current starts at 0, rises to a peak of approximately 15 A at 2 ms, and then decays with oscillations, settling around 5 A by 45 ms.
- Bottom Subplot:** Shows the armature current [A] versus time [ms] from 110 to 140 ms. The current exhibits high-frequency oscillations between approximately 2.5 A and 4 A.

176

Inrush current and zero velocity of the simulated motor cause notable differences in start up phase. Measurement and simulation is nearly the same in the steady state of running motor. The DC motor nonlinearity at stopped rotor and low voltage are also considered (Fig.12.).

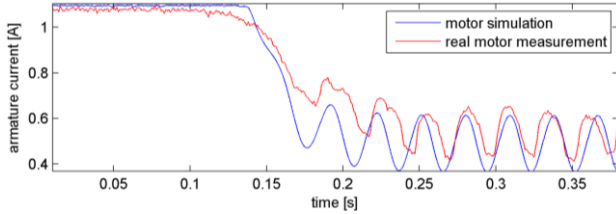


Figure 12. Current waveform during the rotor releas effect

IV. DESIGN OF THE MEU TESTING SYSTEM

A. Requirements

The MEU device must be able to sink or source the currents and voltages, which are expected during fuel pump operation. The MEU is expected to be used in other applications as well, therefore the maximum continuous input power of device was set to 250W which is higher than fuel pump nominal power. Maximum input voltage is 20V and maximal continuous current 13A (peak 20A).

A simple connectivity to the existing ECU (without any significant hardware changes) is required as well as an interface with outside environment using only two clamps. The electronics have to endure unsecure treatment. The simulation running in the RT core has to be as fast as possible to guarantee the stability with the ECU output PWM frequency up to 20kHz. Simulation step is expected to be 10 μ s. The MEU is not expected to simulate the behaviour of R L circuit, because those parts will be realized as real passive parts.

B. System topology design and verification

Primary idea of the simulator is described in (Fig.13.), where known replacement scheme of DC motor was realized with real hardware parts. Back-EMF voltage is generated by a chopper, which operates in bipolar switching mode with high frequency PWM carrier. Inductor and resistor are real parts or motor with mechanically locked rotor might be used. Closed loop hall-effect transducer is used for current sensing.

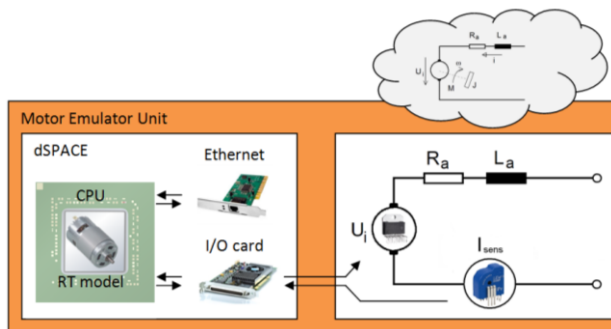


Figure 13. The MEU design idea

The design was verified by simulation of real electronic parts (transistor, diode, resistor and inductor) from Simulink SimScape Electronics library.

C. Hardware design of cost-effective testing system

The final design of the MEU is described in (Fig.14.). The chopper has been already developed in our mechatronics laboratory and is based on ISL83204 integrated chip. Current is measured by LEM transducer with output range of 0.5-4.5V (for current \pm 25A). The transducer has to be matched to the input of dSPACE AD converter (\pm 10V) by customized circuit.

For smoother voltage generation on chopper I. it is necessary to provide high frequency of PWM carrier. However, increasing the frequency leads to the resolution degradation, when using digital PWM generation method. For example dSPACE Autobox has 8-bit resolution, when PWM carrier is set to 100kHz. Therefore a customized circuit was designed to use analogue comparator method for PWM generation. The carrier frequency was set to 110kHz and the PWM generator was controlled by voltage from DAC with 14-bit resolution.

The RL circuit was made by a braked motor. Selector I. enables user to switch between the ECU under test and chopper II, which is controlled from dSPACE, using digitally generated PWM with user defined duty cycle and carrier frequency up to 20 kHz. Chopper II. is used for testing without ECU. Relay serves as a circuit breaker when some of these situations occur: over-current, failure of chopper or during new RT model download.

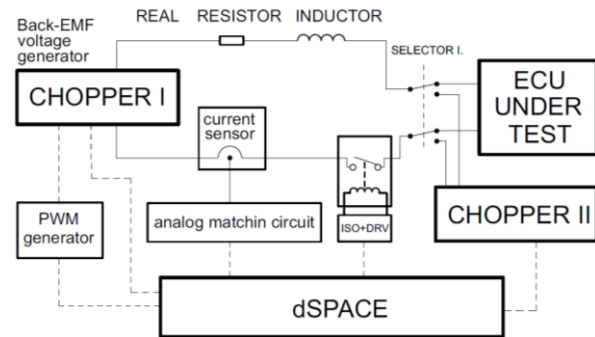


Figure 14. Block scheme of the final MEU including the choppers, R L circuit and dSPACE platform



Figure 15. Development stage of the PHIL emulator

D. MEU testing

The emulator unit was connected to the power supply (6V and 12V) and its responses were captured. Results are shown in (Fig.16.).

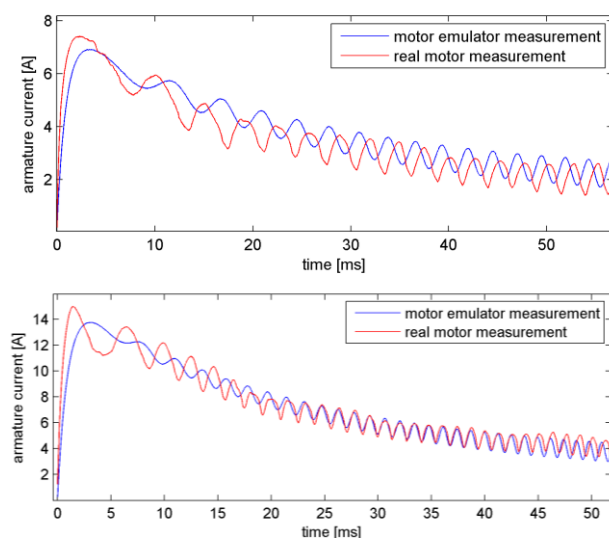


Figure 16. Measurements of the real motor in comparison with measurements of the DC motor emulator.

Some differences between current waveforms during fast dynamic start up are evident. Peak values, phase and shape of waveform depend on initial armature position, which is always unknown. In nearly steady state of rotation, the mean values of the currents are almost the same, and current ripple has the same frequency and nearly the same amplitude. The emulated current waveform is good enough to be used by ECU as a feedback for velocity or torque control purposes.

V. CONCLUSION

The user configurable DC motor emulator has been developed. Parameters of the real motor were measured and then used in simulation of the RT PHIL model. The motor emulator unit (MEU) ability to replace the real motor was proven by comparison of experimental measurement of the real motor and MEU. Parameters settings, virtual mechanical load and states of the virtual plant are controlled via ControlDesk graphical user interface. Model Based Design approach was used for verification of the concept idea during the development stage.

ACKNOWLEDGMENT

This work was supported by the European Commission within the FP7 project Efficient Systems and Propulsion for Small Aircraft "ESPOSA" contract No. ACP1-GA-2011-284859-ESPOSA, and by NETME CENTRE PLUS (LO1202) created with financial support from the Ministry of Education, Youth and Sports under the „National Sustainability Programme I“.

REFERENCES

- [1] Ahmed, O.A.; Bleijs, J.A.M., "Digital control of a fuel cell converter system: Verification, validation and test using a model-based design approach," Education and Research Conference (EDERC), 2010 4th European , vol., no., pp.52,56, 1-2 Dec. 2010
- [2] Rajib Mall: Real-Time Systems: Theory and Practice, Pearson Education, ISBN-13: 9788131700693, India, 2006
- [3] Bouscayrol, A., "Different types of Hardware-In-the-Loop simulation for electric drives," Industrial Electronics, 2008. ISIE 2008. IEEE International Symposium on , vol., no., pp.2146,2151, June 30 2008-July 2 2008
- [4] Jiri Skalicky, "Elektrické servopohon," FEKT, VUT v Brně, ISBN 80-214-1978-4, 2007
- [5] Andrs, O.; Hadas, Z.; Kovar, J., "Introduction to design of speed controller for fuel pump," Mechatronics - Mechatronika (ME), 2014 16th International Conference on , vol., no., pp.672,676, 3-5 Dec. 2014
- [6] Ali, Y.S.E.; Noor, S.B.M.; Bashi, S.M.; Hassan, M.K., "Microcontroller performance for DC motor speed control system," Power Engineering Conference, 2003. PECon 2003. Proceedings. National , vol., no., pp.104,109, 15-16 Dec. 2003
- [7] Jan B Nottelmann, "Sensor-less Rotation Counting in Brush Commutated DC motors," IDEAdvance Ltd © 2010
- [8] Li Yifan, "DC Motor Speed Calculation Based on Armature Current Measurement," Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on , vol.1, no., pp.818,820, 6-7 Jan. 2011
- [9] Yves Thurel, "Switched Mode Converters," CERN, 2004
- [10] Tabbache, B.; Aboub, Y.; Marouani, K.; Kheloui, A.; Benbouzid, M.E.H., "A simple and effective hardware-in-the-loop simulation platform for urban electric vehicles," Renewable Energies and Vehicular Technology (REVET), 2012 First International Conference on , vol., no., pp.251,255, 26-28 March 2012
- [11] Kazerani, M., "A high-performance controllable AC load," Industrial Electronics, 2008. IECON 2008. 34th Annual Conference of IEEE , vol., no., pp.442,447, 10-13 Nov. 2008
- [12] T. Tjahjowidodo, F. Al-Bender, H. Van Brussel, Friction Identification and Compensation in a DC Motor, 2005
- [13] Martin Hartl, The Measurement and Study of Very Thin Lubricant Films, FSI, VUT v Brně, ISBN 80-214-2224-6, VUTUM 2002
- [14] V. van Geffen, "A study of friction models and friction compensation," Technische Universiteit Eindhoven, Department Mechanical Engineering, 2009

Symbolic Kinematic and Dynamic Modelling toolbox for Multi-DOF Robotic Manipulators

Paolo Righettini; Roberto Strada; Ehsan Khadem Olama; Shirin Valilou;

Department of Engineering and Applied science, Università degli studi di Bergamo, Bergamo, Italy.

paolo.righettini@unibg.it, roberto.strada@unibg.it

e.khademolama@studenti.unibg.it, s.valilou@studenti.unibg.it,

Abstract— The objective of this article is to present a method to model the kinematics and dynamics of robot manipulators. A complete description of the procedure to model and control a Multi-DOF 3D robot manipulator is detailed and simulated using designed toolbox in MATLAB. Examples of path planning, symbolic dynamic derivation and control strategy designs are presented. Advanced path planning by constrained optimization has been developed and verified by this toolbox. One of the advanced position controllers for robot manipulators is the Sliding Mode Control with desired gravity and friction compensation which has been applied and verified in the toolbox. Force contacts and collisions are modeled in nonlinear near real situations. This helps designing more robust force controlled systems.

Keywords— Robot Manipulator; Multi-DOF Serial Robot; Kinematics and Dynamics Modelling; MATLAB toolbox;

I. INTRODUCTION

Robot manipulators are the main devices of technology and industrial advancements in few recent decades [1, 2]. They are hands of man in technology and so need to be as precise, quick and trustworthy as we need them. Manipulators are being taught in academic extensively but not so many practices because of their prices. New ideas come in mind of students when reading books and articles but as lack of experimental laboratories they could not work on them.

One special aspect of manipulators is designing control strategies and applying them on real robots. This aspect needs more special treatment as it is the key point to accuracy and repeatability of each robot. Many softwares and toolboxes are designed to help students of mechatronic to learn basics of robot manipulators [3-8]. But none of them are focused in a special aspect like control engineering designing and many of them are developed just for some special cases with limited DOF manipulators.

In control designing process, knowing of the fully parametric structure of the dynamics of any kind of robot manipulator is needed to study and design a successful strategy. As these requirements, the Advanced Robotic Manipulator Simulator “ADROMS” has been developed in University of Bergamo for studying on advanced robotics. ADROMS is a fully parametric-symbolic universal serial robot manipulator simulator which is written in MATLAB for control developments. Being the most automatic-ware and simultaneously flexible in practice were of the development

goals. In this paper the toolbox and its components will be described.

II. KINEMATICS AND DYNAMICS OF A MANIPULATOR[9]

A. Basic Transformation and Rotation

In this toolbox the Cartesian frame for rigid bodies has been adopted. A rigid body is described in space by its position and orientation with respect to reference frame ($O-xyz$). For each body is considered a point in it (usually center of mass) as origin of rigid body frame ($O'-x'y'z'$). By these considering, the orthogonal rotation matrix formulas are:

$$R_i^b = [\bar{x}', \bar{y}', \bar{z}'] = \begin{bmatrix} x'_x & y'_x & z'_x \\ x'_y & y'_y & z'_y \\ x'_z & y'_z & z'_z \end{bmatrix} \quad (1)$$

$$R_z(\alpha) = \begin{bmatrix} c\alpha & -s\alpha & 0 \\ s\alpha & c\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix}, R_y(\beta) = \begin{bmatrix} c\beta & 0 & s\beta \\ 0 & 1 & 0 \\ -s\beta & 0 & c\beta \end{bmatrix}, \quad (2)$$

$$R_x(\gamma) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c\gamma & -s\gamma \\ 0 & s\gamma & c\gamma \end{bmatrix}$$

In the toolbox three functions are for rotations:

`>> Rz(pi/3)`

Then if we consider a point in space it can be rotated to new point by:

$$P' = R_{xyz}(\Phi)P; \Phi = [\varphi, \vartheta, \psi] \quad (3)$$

By considering this rotation we can construct 12 different sets of angles allowed. Each set represents a triplet of Euler Angles. Two traditional sets are ZYZ and ZYX (Roll-Pitch-Yaw).

RPY angles originates from a representation of orientation in the aeronautical field to denote the typical changes of attitude of an aircraft. In the toolbox the function for these rotation and its inverse is the:

`>> R=tet2rpy([pi/3,pi/4,pi/6])`

`>> tet=rpy2tet(R)`

For non-minimal rotations the toolbox has functions including quaternion:

```
>> R=rkno2rot(rr,pi/3)
>> R=rkno2quat(rr,kno)
>> [rr,kno]=quat2rkno(R)
```

Considering O_1^0 be the vector describing the origin of frame 1 with respect to frame zero (base) and R_1^0 be the rotation matrix of frame 1 with respect to frame 0. Let also P^1 be vector of coordinates of P with respect to frame 1. By simple geometry position of P in frame 0 would be (Fig. 1):

$$P^0 = O_1^0 + R_1^0 P^1 \quad (4)$$

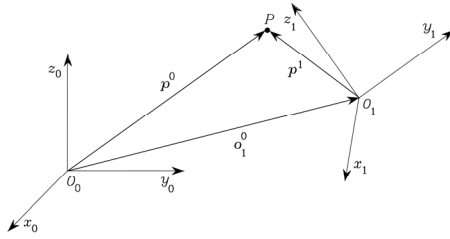


Fig. 1. Rigid body position and orientation with respect to base frame.

By this transformation and defining new 4 element point vector $\tilde{P} = \begin{bmatrix} P \\ 1 \end{bmatrix}$ we can make a homogeneous transformation matrix:

$$A_1^0 = \begin{bmatrix} R_1^0 & O_1^0 \\ 0^T & 1 \end{bmatrix} \quad (5)$$

$$\tilde{P}^0 = A_1^0 A_2^1 \dots A_n^{n-1} \tilde{P}^n \quad (6)$$

In the toolbox this transformation is available as:

```
>> htm(R,Ori)
```

B. Direct Kinematic

A manipulator or an arm (Fig. 2) consists of a series of rigid bodies (links) connected by means of kinematics pairs or joints. Joints can be essentially of two types: revolute (rotational motion) and prismatic (translation). The whole structure forms a kinematic chain. One end of the chain is constrained to a base. An end-effector (gripper, tool) is connected to the other end allowing manipulation of object in space. Each joint provides a mechanical structure with a single degree of mobility, represented by a joint variable q . Thus, the location of the first link depends on the value of the joint variable related to first joint q_1 . Since each joint connects two consecutive links, then the location of a link n depends on the corresponding joint variable q_n as well as on all the previous joint variables $q_1 \dots q_{n-1}$. The mechanical structure of a manipulator is characterized by a number of degrees of freedom (DOFs) which uniquely determine its posture. Each DOF is typically

associated with a joint articulation and constitutes a joint variable. The aim of direct kinematics is to compute the pose of the end-effector as a function of the joint variables.

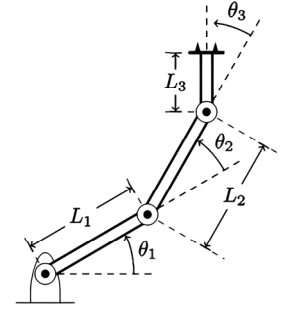


Fig. 2. A three-link planar manipulator.

Then the end-effector posture can be written as:

$$T_n^0 = A_1^0(q_1) A_2^1(q_2) \dots A_n^{n-1}(q_n) \quad (7)$$

C. Denavit Hartenberg Convention

The Denavit–Hartenberg parameters (also called DH parameters) are the four parameters associated with a particular convention for attaching reference frames to the links of a spatial kinematic chain. Jacques Denavit (Dr. Esai alumni) and Richard Hartenberg introduced this convention in 1955 in order to standardize the coordinate frames for spatial linkages [10].

This method uses 4 parameters to describe possible presentation of the end effector posture according to joints variables:

d_i : offset along previous z to the common normal.

θ_i : Angle about previous z , from old x to new x .

r_i or a_i : length of the common normal (aka a , but if using this notation, do not confuse with α). Assuming a revolute joint, this is the radius about previous z .

α_i : Angle about common normal, from old z axis to new z axis.

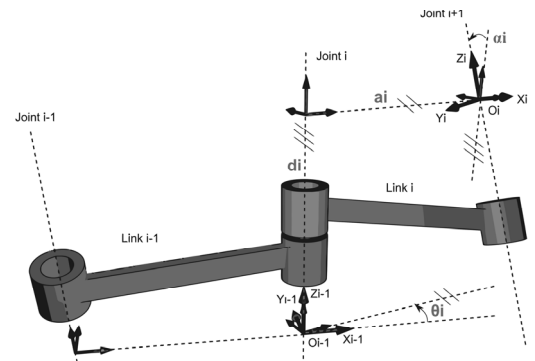


Fig. 3. Principals of Denavit-Hartenberg Standard.

In the toolbox homogeneous transformation of each linkage can be taken by function:

```
>> DH([q(1),L(1),pi/2,1])
```

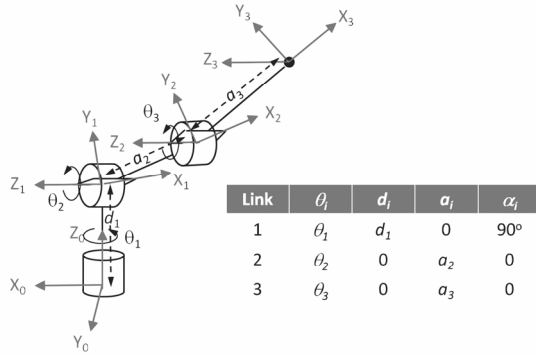


Fig. 4. An anthropomorphic arm with 3DOFs.

By using this function and a recursive approach the full kinematic of the system of (Fig. 4) can be derived as in script e001_Foward_Kinematic.m:

```
DHTable=[q(1),L(1),pi/2,0;q(2),0,0,L(2);q(3),0,0,L(3)];
SFK=symDHmodel(DHTable)

[cos(q2 + q3)*cos(q1), -sin(q2 + q3)*cos(q1), sin(q1), cos(q1)*(L3*cos(q2 +
q3) + L2*cos(q2))]

[cos(q2 + q3)*sin(q1), -sin(q2 + q3)*sin(q1), -cos(q1), sin(q1)*(L3*cos(q2 +
q3) + L2*cos(q2))]

[sin(q2 + q3), cos(q2 + q3), 0, L1 + L3*sin(q2 + q3) +
L2*sin(q2)]

[0, 0, 0,
```

D. Differential Kinematic

Based on the DH transformation, the posture of the end-effector can be written as (8). So we can estimate the slight differences in joint space variables with respect to slight differences in work space variables by (9).

$$P = FK(q) \quad (8)$$

$$\dot{P} = J(q)\dot{q} \quad (9)$$

E. Dynamics

Robot Dynamics is the study of the relation between the applied forces/torques and the resulting motion of an industrial manipulator. Computation of the time evolution of $\ddot{q}(t)$ (and then of $\dot{q}(t), q(t)$), given the vector of generalized forces (torques and/or forces) $\tau(t)$ applied to the joints and, in case, the external forces applied to the end-effector, and the initial conditions $q(t=t_0), \dot{q}(t=t_0)$ is said the Direct Dynamic Model. There are several reasons for studying the dynamics of a manipulator (simulation, analysis and synthesis, analysis of the structural properties). There is two approaches for the definition of the dynamic model:

Euler-Lagrange: First approach to be developed. The dynamic model obtained in this manner is simpler and more

intuitive, and also more suitable to understand the effects of changes in the mechanical parameters. The links are considered altogether, and the model is obtained analytically.

Newton-Euler: Traditionally the Newton-Euler equations is the grouping together of Euler's two laws of motion for a rigid body into a single equation with 6 components, using column vectors and matrices. These laws relate the motion of the center of gravity of a rigid body with the sum of forces and torques (or synonymously moments) acting on the rigid body.

F. Euler-Lagrange Differential Equation

For a system we can define two distinct function of kinetic and potential energy as $K(q, \dot{q})$, $P(q)$ respectively. Then the Lagrange function can be defined as:

$$L(q, \dot{q}) = K(q, \dot{q}) - P(q) \quad (10)$$

By this definition the Euler-Lagrange Differential Equation is:

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \right) - \frac{\partial L}{\partial q_i} = \tau_i; i = 1 : n \quad (11)$$

Each τ_i consists of:

$$\tau_i = \tau_{\text{internal}} - \tau_{\text{external}} - \tau_{\text{frictions}} \quad (12)$$

Velocity of any linkage can be written as:

$$\vec{V} = \vec{V}_{\text{com}} + \omega \vec{r} \quad (13)$$

So we can find any kinetic energy by:

$$K = \frac{1}{2} \sum_{i=1}^n m_i V_{\text{com}i}^T V_{\text{com}i} + \frac{1}{2} \sum_{i=1}^n \omega_i^T R_i \tilde{I}_i R_i^T \omega_i \\ = \frac{1}{2} \dot{q}^T M(q) \dot{q} \quad (14)$$

By considering (14), Inertia Matrix can be extracted by taking derivative of Lagrange function with respect to \dot{q} two times. For making matrix of Centrifugal and Coriolis effect the Christoffel Symbols (15) are used.

$$c_{ijk} = \frac{1}{2} \left[\frac{\partial M_{kj}}{\partial q_i} + \frac{\partial M_{ki}}{\partial q_j} - \frac{\partial M_{ij}}{\partial q_k} \right] \\ C_{kj} = \sum_{i=1}^n c_{ijk} \dot{q}_i \quad (15)$$

In deriving the dynamic model, the actuation system has not been taken into account. These dynamics consists of motors, gears and transmission systems. Actuation system it introduces additional nonlinear effects such as backlash, friction and elasticity. Full dynamic of the system can be written as:

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + G(q) = \tau_{\text{internal}} - \tau_{\text{external}} - \tau_{\text{frictions}} \quad (16)$$

In the ADROMS toolbox, the function $[M,C,P,Jac,Pos,Rotj,Jw]=symDynamic(m,q,dq,L,b,Ma,Jm,DHTable,g)$ computes dynamic model of all the types of serial manipulators based on Denavit Hartenberg table symbolically which is the most needed in control strategy designs. Script e003_Dynamic_Symbolic_Model.m shows an example of a two link or higher dynamic computation. Results of this computation are:

The M matrix:

$$\begin{bmatrix} Ma2*L1^2 + 2.0*Ma2*cos(q2)*L1*b2 + Ma1*b1^2 + Ma2*b2^2 + J1_33 + J2_33, & Ma2*b2^2 + L1*Ma2*cos(q2)*b2 + J2_33, \\ Ma2*b2^2 + L1*Ma2*cos(q2)*b2 + J2_33, & Ma2*b2^2 + J2_33 \end{bmatrix}$$

The C matrix:

$$\begin{bmatrix} -1.0*L1*Ma2*b2*dq2*sin(q2), & -1.0*L1*Ma2*b2*sin(q2)*(dq1 + dq2) \\ L1*Ma2*b2*dq1*sin(q2), & 0 \end{bmatrix}$$

The G matrix:

$$\begin{bmatrix} G*Ma2*(b2*cos(q1 + q2) + L1*cos(q1)) + G*Ma1*b1*cos(q1) \\ G*Ma2*b2*cos(q1 + q2) \end{bmatrix}$$

The full Jacobian matrix of kinematic:

$$\begin{bmatrix} -L2*sin(q1 + q2) - L1*sin(q1), & -L2*sin(q1 + q2) \\ L2*cos(q1 + q2) + L1*cos(q1), & L2*cos(q1 + q2) \\ 0, & 0 \\ 0, & 0 \\ 0, & 0 \\ 1, & 1 \end{bmatrix}$$

The Position matrix of kinematic:

$$\begin{bmatrix} L2*cos(q1 + q2) + L1*cos(q1) \\ L2*sin(q1 + q2) + L1*sin(q1) \\ 0 \end{bmatrix}$$

The Rotation matrix of kinematic:

$$\begin{bmatrix} cos(q1 + q2), & -sin(q1 + q2), & 0 \\ sin(q1 + q2), & cos(q1 + q2), & 0 \\ 0, & 0, & 1 \end{bmatrix}$$

III. CONSTRAINED SMOOTH PATH PLANNING

Vibrations may be produced if trajectories with a discontinuous acceleration profile are imposed to the actuation system [11]. In the forward kinematic by knowing any set of q you can determine the posture of the end effector in space. But the problem comes from specifying a posture in space and want the robot to follow it. As the formulation of the forward kinematic is highly nonlinear according to the joint variables, finding a smooth path in joint spaces based on the unknown reverse nonlinear posture to joint space variables is one of hard in robotics. In real life robotic, each joint has been constructed by a servo motor which can rotate in constrained portion of a full circle $[-\theta_1, +\theta_2] \subset [-\pi, \pi]$. In this case finding a smooth path is a real hard. Here we have proposed a solution with optimization approach. Consider forward kinematic (8).

Then we can construct a nonlinear optimal objective function as:

$$f(q) = (K(q) - P_{ref})^T (K(q) - P_{ref}) + \alpha(q - q_{init})^T (q - q_{init}) \quad (17)$$

$$s.t \{-\theta_1 < q < \theta_2\} \quad (18)$$

$$Grad_f(q) = \frac{\partial f(q)}{\partial q} = 2(K(q) - P_{ref})^T J(q) + 2\alpha(q - q_{in})^T \quad (19)$$

In this objective function we want to find q s which are subject to (18). The first part of the objective function reduces as the q finds the path P_{ref} , so represents the accuracy of the path and the second part chooses the optimal q according to initial selection q_{init} , so that we find the shortest path in the joints space variables. α Determines how much we want the optimality part effect on the objective function. If this variable goes to 1 the effect of the optimality would be the same as the path accuracy. As the nonlinear objective function is a kind of quadratic function with respect to nonlinear parts we choose Interior point methods (also referred to as barrier methods) which are a certain class of algorithms to solve linear and nonlinear convex optimization problems. For this optimization we need one more function which is the Gradient of the objective function (19).

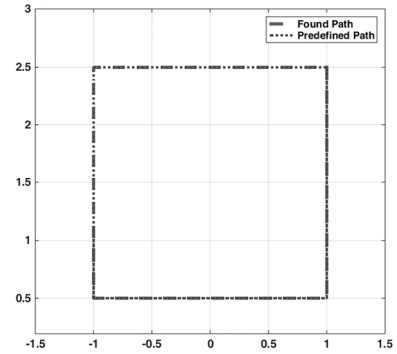


Fig. 5. A square path predefined and found by optimization.

Our method uses the previous found best q as the initial for new optimization. In the script e002_PathPlanning_2.m, a square path within the reachable of the workspace have been considered for a three-link planar manipulator with restriction in q s. Each joint just can move from $-\pi/3$ to $\pi - \pi/6$ radian. The path planning have been done by this constraints and the results are shown in (Fig. 5 and Fig. 6). As seen, the found path in joint space is constrained, smooth and with an error of less than $80\mu m$ (Fig. 7).

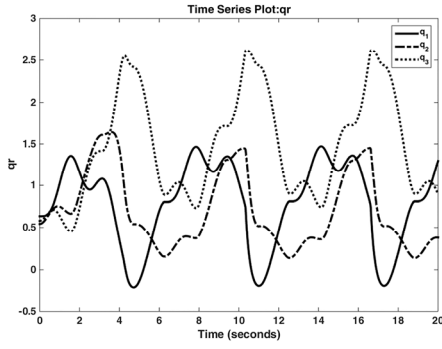


Fig. 6. Smooth path of q in joint space.

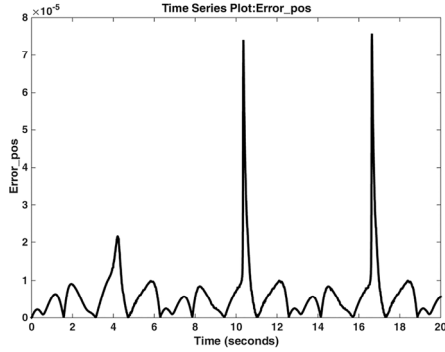


Fig. 7. Error in designed position which is less than $80\mu m$

IV. POSITION CONTROL

After path planning it is needed to apply this path in the dynamic structure of the robot manipulator. This is realized by designing a control box for robot manipulator dynamics to track the reference pre-designed path in joint spaces. There are many kinds of control strategies from simplest PID [12, 13] to complex model predictive control [14, 15], from classical controllers to new methods of neural networked [16] adaptive fuzzy controllers [17, 18]. All these controllers need an accurate and simple model to investigate their designs. In ADROMS it has been designed a Simulink version of the dynamic system, so working with any kind of control is easy and precise.

For illustration, a Sliding Mode Controller for position tracking with assumption of friction and gravity parts as uncertainty has been developed. As the dynamic model of the system is parametric and it does not need to be designed from scratch, it is easy just increasing or decreasing the dynamic model by defining just simple base parameters like Denavit-Hartenberg, masses centers of mass and inertia.

In the script `e004_Dynamic_Position_Control.m` and Simulink file `fullSimulink_FSMC.slx` a sliding mode control for a 3DOF for a 3D path is described and composed by SIMULINK. In this illustration the friction and the gravity potential part of the dynamic of the system has been considered as uncertainty. Even a high frequency uncertainty at least equal to 50% of the system has been proposed to dynamic. Sliding Mode Control could have rejected all these uncertainties at minimum time. Drawbacks of this sliding mode control is that its input control is highly chattering and if the uncertainty goes beyond that much considered in design time all the dynamic of

the system goes instabilities. Results of the simulation has been provided.

V. CONTACT FORCES

In physics, a contact force is a force that acts at the point of contact between two objects, in contrast to body forces. Contact forces are described by Newton's laws of motion, as with all other forces in dynamics. Pushing a car up a hill or kicking a ball or pushing a desk across a room are some of the everyday examples where contact forces are at work.

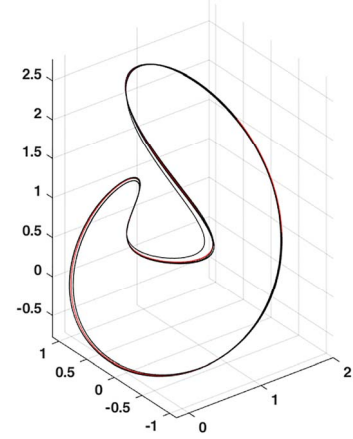


Fig. 8. Path in 3D which has been tracked by an anthropomorphic arm.

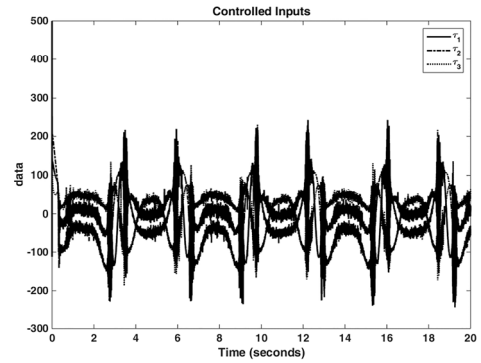


Fig. 9. Highly Chattering Controlled Input.

In the first case the force is continuously applied by the person on the car, while in the second case the force is delivered in a short impulse. The most common instances of force contacts include friction, normal force, and tension. According to forces, contact force may also be described as the push experienced when two objects are pressed together.

For simulation the contact and collision force on the end-effector while interacting with environment we have considered a semi rigid environment (Fig. 11). From mechanical principles, any interaction between two bodies can be simulated with three elements of spring, damper and inertia. These three elements are linear in low speed of physical systems. But by interfering high speeds of collision some nonlinear phenomena would be activated.

For calculation of the effect of the external forces on the system according to the equation (16) we can write the external force effect on joint space as:

$$\tau_{external} = \sum_i^m J^T(q) F_{i,ext}(p) \quad (20)$$

Where the m , is the number of external forces in workspace variables. Each external force vector can be written in the basis of the orthonormal vectors of the contact surface $F_{ext} = [F_n, F_t]$ each surface in workspace has its own formula known as constraint in workspace variables:

$$\Phi_i(p) = 0; i = 1:m \quad (21)$$

$$\begin{aligned} \frac{\partial \Phi(p(q))}{\partial t} &= \frac{\partial \Phi(p(q))}{\partial p} \frac{\partial p}{\partial q} \frac{\partial q}{\partial t} \\ &= A_p J(q) \dot{q} = A(q) \dot{q} = 0 \end{aligned} \quad (22)$$

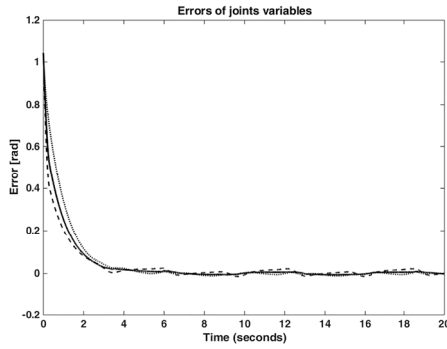


Fig. 10. Errors of the Dynamics as time goes on. Errors are about less than 0.07 radian.

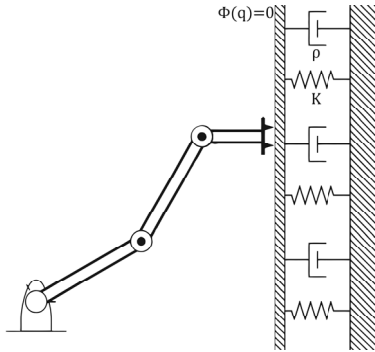


Fig. 11. Environment can be modeled as a non-rigid.

So the normal vector to the surface at point q can be derived

$$\vec{n}_p = \frac{A_p^T}{\|A_p\|} = \frac{J^{-T} A_q^T}{\|J^{-T} A_q^T\|} \quad (23)$$

The normal force on the end-effector from surface can be written as:

$$F_n = |f_n| \vec{n}_p \quad (24)$$

Now we start to construct a model for contacted surface. If we consider the surface as a non-rigid with appropriate values for elasticity of the surface then we can make a semi-rigid environment as (Fig. 11) Conventional formulation for normal force would be:

$$f_n = K \Phi(q) + \rho \|(\dot{p} \cdot \vec{n}_p)\| \quad (25)$$

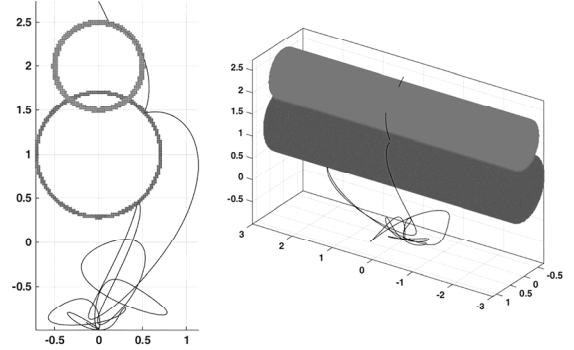


Fig. 12. Two cylinder constraints and collision of end-effector with them.

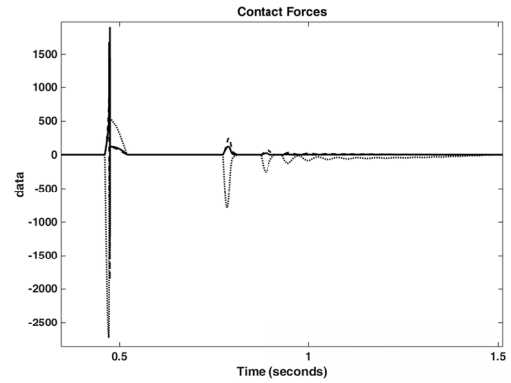


Fig. 13. Spike Forces calculated during collisions.

This formula consists of a spring with K which is less than 10^{+6} and ρ around 450 [19]. For the perpendicular tangent force we have:

$$\begin{aligned} F_t &= \mu(\dot{p}) |f_n| \vec{n}_t \\ \vec{n}_t &= \frac{\dot{p} - (\dot{p} \cdot \vec{n}_p) \vec{n}_p}{\|\dot{p} - (\dot{p} \cdot \vec{n}_p) \vec{n}_p\|} \end{aligned} \quad (26)$$

The $\mu(\dot{p})$, is a nonlinear function of \dot{p} :

$$\begin{aligned} \mu(\dot{p}) &= (\mu_s - \mu_k) \frac{-\tanh(\dot{p}_d - v_{th}) + 1}{2} + \mu_k \\ \dot{p}_d &= \|\dot{p} - (\dot{p} \cdot \vec{n}_p) \vec{n}_p\| \end{aligned} \quad (27)$$

ADROMS toolbox provides this modelling and simulation by script e005_Dynamic_Contact.m and Simulink of

Techniques for Monitoring and Predicting the OCV for VRLA Battery systems

Alessandro Mariani^{*1}, Kary Thanapalan¹, Peter Stevenson², Thomas Stockley¹, Jonathan Williams¹

¹Centre for Automotive and Power Systems Engineering (CAPSE) University of South Wales, Pontypridd CF37 1DL, UK

²Yuasa Battery (UK) Ltd, Rassau Industrial Estate, Ebbw Vale NP23 5SD, United Kingdom

E-mail*: alessandro.mariani@southwales.ac.uk

Abstract— This paper provides a simple and advanced technique to monitor and predict the state of health (SOH) and the open circuit voltage (OCV) for valve regulated lead acid (VRLA) battery cells. The underlying principal of the technique described in this paper employs a mathematical model that can simulate different pore geometries shape and a simple equation to predict the equilibrated cell voltage after a small rest period. The technique was tested and analyzed using results obtained from experiments conducted at the Yuasa Battery laboratories. Lead-acid battery system analysis was carried out with reference to the standard battery system models available in the literature. The results indicate that by using this technique, appreciable benefits can be accrued and it is possible to maintain high standard products in safe operating conditions

Keywords: Lead-acid technology; cell relaxation; system analysis; battery system; prediction mechanism

I. INTRODUCTION

Critical equipment in datacenters, bank, hospital, and many other industries must be supported by uninterruptible power supplies (UPS) to ensure their reliability. A rapidly expanding UPS market is spreading globally, requiring battery devices that can support high rate discharge applications in wide range of climate conditions [1, 2, 3, 4]. Valve regulated lead acid batteries are the technology of the choice in the majority of these applications. New product designs have been developed to meet the specific requirements of high rate UPS discharges and these have presented new challenges for the control of traditional manufacturing processes.

The lead-acid battery may not have the high energy density or fast cycle rates that the lithium and Nickel Metal Hydride (NiMH) technologies possess [5] but they are unparalleled in their ease of use, recyclability and low cost [6].

Generally the batteries are only required in case of an emergency. Therefore, it is of the utmost importance to ensure that, when called upon, the batteries are in a good operational state. Techniques that can provide information about the state of health of a battery, without resorting to full discharge tests, are useful to integrate

within the normal operating procedures of the site equipment. The current investigation sought to relate continuously measurable electrical characteristics of VRLA batteries with the detailed physical structure of their storage electrodes. A mathematical model may be used to distinguish the dynamic behavior of electrochemical reactions of passive elements system of lead-acid cells under an excitation of small amplitude. A variety of investigations have been conducted and published concerning the characterization of porous electrodes in order to supply information about their surface area and electrochemical utilization [7].

In this paper, firstly, a mathematical model was developed to identify the VRLA battery cell state of health. The model then validated with the use of experimental data obtained through the EIS (electrochemical impedance spectroscopy) technique. Secondly, to ensure that the cells are in a high SOC charged state, a monitoring system was developed for the OCV-SOC.

II. PREDICTION AND MONITORING TECHNIQUES

The mathematical model proposed in this work, is able to simulate the structure of different pore average geometries by comparing with the values calculated and from the impedance curve. The system related the coupling of the double-layer capacitance and the solution resistance in the pores with geometric effect at low frequencies in complex plane, transforming the complex valued functions in an equivalent real valued differential equation system. The novel model proposed can provide a better behaved system transformation compared with previous studies [8, 9], by simulating the appropriate crystal pore geometries for different state of charge conditions. Applying this mathematical model of different state of health products allows the identification of the real electrode pore geometries that penalize the overall battery efficiency.

The proposed method for monitoring and predicting the OCV for VRLA battery systems comprises two sections; the first is to identify the product to be tested and

analyzed and it can be done via the above mention technique. The second part is to determine the equilibrated OCV. Accurate prediction of OCV after a small relaxation period has been proven by Mariani et al. [10] and will be incorporated in this work.

A mathematical model will be developed and incorporated to the original technique developed by Mariani et al. [10]

The OCV analysis carried out in the previous work has indicated that it worked successfully, providing estimation errors of just 0.2%. Therefore, to ensure, the method will work for different cell type, regardless of its state of health, this paper investigates worst case scenarios by analyzing results of the low performance status VRLA battery cells.

The paper is organized as follows; firstly a mathematical model will be presented and is able to simulate the structure of different pore shapes and at different state of charge (SOC). The work then proceeds to investigate the worst case scenarios by analyzing results of the low performance status VRLA battery cells by using OCV-SOC prediction technique.

III. SYSTEM MODELING AND OPERATION

The analytical mathematical model proposed in this paper, is for the identification of the different stage of the process of individual elements of a VRAL battery system. It can be represented by the equivalent circuit shown in Fig. 1.

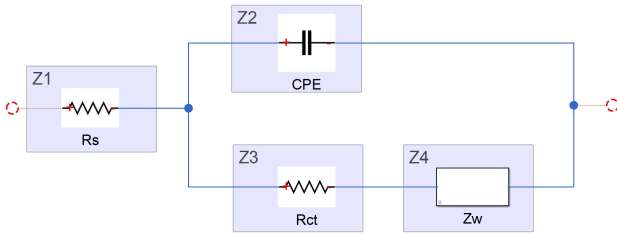


Fig. 1 Electrical equivalent circuit of diffusion process for VRLA battery

In Fig.1, Z1 represents (R_s) the charge transport in solution that is the ohmic resistance due to all the Pb involves in the cell battery design, the separator and electrolyte resistivity, Z2 is the charge transfer at the interface (Constant phase element - CPE) that is related to the porosity of the electrode, Z3 (R_{ct}) is the resistances of the charge transfer at the electrode, and Z4 is the Warburg impedances (Z_w) that characterizes the ion diffusion in the electrolyte and in the pores of electrodes. Since, Warburg impedances play a key role to estimate the SOH of the VRLA batter system, the investigation in this paper is focus on the description of the geometry pore shapes. In block Z4, it was considered the electrolyte resistance in the pore cavity and the electrolytic resistance in the pore as current flow, and it was related to the variation of the pore impedance Z with the angular

frequency w . For the experimental analysis it was decided to utilize Electrochemical Impedance Spectroscopy (EIS) technique since it is demonstrated to be a powerful and proven tool for studying AC impedance response [11]. The Z_w is substituted with ζ function that represents the analytical equations system to identify pore geometries shape. In the mathematical model, it was considered that the resistance of electrolyte as variable which is related to the electrode state of charge. The shape geometry is assumed as Ro function. Therefore the actual impedance can be written as:

$$Z = Ro * \zeta \quad (1)$$

For example, the Ro function for a cylindrical shape can be written as:

$$Ro = \frac{l}{\pi * K * r^2} \quad (2)$$

where, l is depth of poor cavity, K is electrolyte conductance (Ω^{-1}/cm) that measures the ability of sulphuric acid solution to conduct electricity and r is the outlet pore radius. Regarding the pore shape identification " ζ ", 0 was assumed to be as the bottom pore condition and 1 is the pore outlet state, as it can be seen from Fig.2.

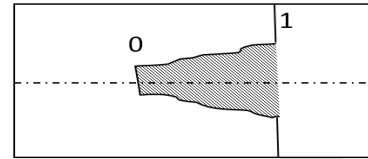


Fig. 2 Domain pore geometry

Now, consider that, $x \in [0,1]$ where $f(x)$ represents the form factor and it describes the deviation from the cylindrical shape of the pore. $\forall x, T > 0, T \in [0.01, 10000]$ the frequency variable is measured in Hz, but it is important to note that the frequency domain was converted in rad/s in the model. Thus, the general model will have the following form:

$$P'_{T(x)} = -Tf(x)S_T(x) \quad (3)$$

$$q'_{T(x)} = Tf(x)r_T(x) \quad (4)$$

$$r'_{T(x)} = \frac{1}{f(x)^2}P_T(x) \quad (5)$$

$$S'_{T(x)} = \frac{1}{f(x)^2}q_T(x) \quad (6)$$

with the initial values of

$$\begin{aligned} p_T, q_T, s_T &= 0 \\ r_T &= 1 \end{aligned} \quad x \in [0,1]$$

Therefore,

$$\zeta_{(T)} = \frac{(r_T(1)P_T(1)+S_T(1)q_T(1))+i(S_T(1)P_T(1)-r_T(1)q_T(1))}{P_T(1)^2q_T(1)^2} \quad (7)$$

Several variation of geometric shapes can be realized by using the modelling technique presented above. For illustrative purpose, three different pore geometries example results (cylinder, cone and bubble) are shown below.

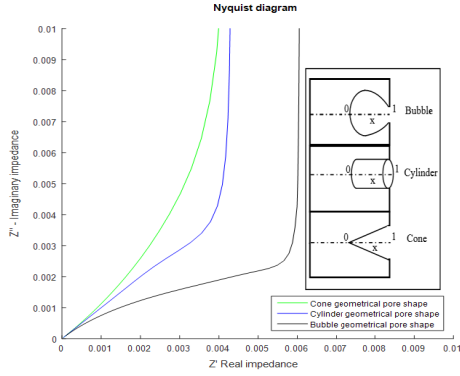


Fig. 3 Impedance curve of different pore shape

$$\text{Cylinder: } f(x) = 1 \quad (x \in [0,1])$$

$$\text{Cone: } f(x) = x + \varepsilon \quad (x \in [0,1])$$

$$\varepsilon \leq 0.01 \text{ and } \varepsilon \neq 0$$

$$\text{Bubble: } f(x) = \sqrt{(2r-x)x} \quad (x \in [0,1])$$

$$\text{with } \frac{1}{2} < r \leq 1$$

So far a mathematical model for the identification of real electrode pore geometry shape has been established. This model is then used to identify the product to be tested. The second part is to determine the OCV to monitor the OCV-SOC of the selected product.

A variety of literature has been published on various aspects of battery cell modeling and operation [12], and that can be used to simulate and estimate the voltage of a cell from a set of known parameters. Barsani and Ceraolo [13] defined a dynamic model of lead-acid batteries to use in the investigation of their discharge behavior. The general model of a lead-acid battery that represents the discharge equivalent network is shown in Fig. 4.

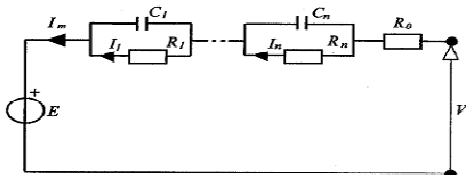


Fig. 4 Lead acid equivalent network for discharge electrochemical reactions [13]

The recovery voltage after charge and discharge is related to the resistance of R-C block ($R_1 \sim R_n$), and the voltage recovery time is dependent on the capacitance ($C_1 \sim C_n$) represented in the lead-acid equivalent network.

Furthermore, Barsani and Ceraolo [13] developed their mathematical model as a function of the battery SOC, dependent of the electrolyte concentration, operating temperature and OCV reading. It measures the initial voltage V_0 (voltage battery reading in a charged condition), discharged at constant current and recorded the OCV, V_3 and V_4 , until the battery reached a stable equilibrium represented in the Fig. 5 as V_1 value. After trying various methods for the identification of the lead-acid battery OCV, it was decided to conduct the investigation based on the previous work carried out by the authors [10]. With the low performance lead-acid technology, the internal cell equilibrium, at OCV is achieved in the early stages, so the coulometric technique was taken into account in this investigation.

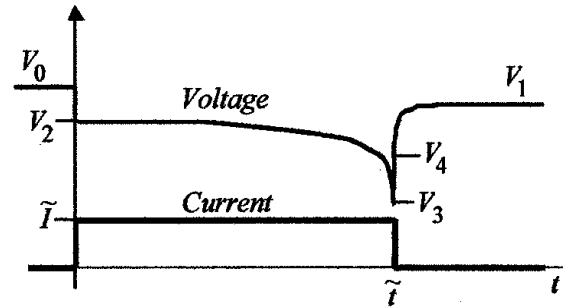


Fig. 5 Typical voltage and current profile for a constant current discharge [13]

The cell voltage can be derived in both the charge and discharge state by using equation (8) and equation (9).

$$V_b^d = E_0 - R i_b - K \frac{Q}{Q - i_b t} (i_b t + i_b^*) + A \exp(-B i_b t) \quad (8)$$

$$V_b^c = E_0 - R i_b - K \frac{Q}{i_b t - 0.1 Q} i_b^* - K \frac{Q}{Q - i_b t} + A \exp(-B i_b t) \quad (9)$$

Where, V_b^c and V_b^d are the cell voltage during charge and discharge. E_0 is battery constant voltage, R is internal resistance, Q is battery capacity, i_b^* is filtered current and $i_b t$ is actual battery charge. K is the polarisation resistance, value dependent on the conductivity of the electrolyte (H_2SO_4) in place of the battery. A and B are exponential zone amplitude and time constant inverse respectively.

The equilibrated OCV is given by;

$$V_{OC} = V_{tr} \pm K_v \quad (10)$$

where, V_{OC} is the equilibrated OCV, V_{tr} is the voltage at the time of measurement and K_v is a constant derived from the equation $V_{OC} - V_{tr}$.

The cells under study are composed of three main components: Positive electrode made from lead dioxide

(PbO₂), negative electrode made from lead (Pb) in a porous pasted form and the electrolyte is dilute sulphuric acid (H₂SO₄). The cell voltage is related to the strength of the electrolyte in the pores of the active material and very often the cut-off voltage is reached without using the full reserve of strong electrolyte available in the separator. This is especially true in high load conditions. During an open circuit state, the strong electrolyte will gradually diffuse from the separator into the pores of the active material, replacing the very weak electrolyte. This process can be seen by monitoring the OCV, where the voltage of the cell can be seen to increase slowly until after 4 hours voltage appears stable; this is known as the recovery factor. The recovery factor is dependent on the load applied to the cell; a higher load results in a greater difference between the load and equilibrated voltage.

IV. RESULTS AND DISCUSSION

The ability to estimate the equilibrated OCV at a short interval allows the OCV-SOC method to be used in a practical system. Furthermore, it is important to note that due to the simplicity of the method described in this work, the equation used in this paper to calculate the equilibrated OCV is so simple that can easily implemented in a real world application.

The experimental work conducted during this research work was performed on positive lead-acid electrode, from valve regulated lead-acid (VRLA) batteries, and are the representation of many tests carried out in Yuasa Ltd laboratory. The first part of the investigation is focused on the study of Warburg element model, that characterize the diffusion phenomenon at the electrodes interface, and that represent the limiting step of the entire electrochemical process in VRLA battery cells. The results, indicates that the difference in size and pores distribution across the electrodes, affect the electrochemical process limiting electrolysis during charge and discharge activity.

Validation of the mathematical model was carried out using EIS techniques for both standard and low performance VRLA cells at different state of charge. Two Yuasa SWL2500 (90 Ah) batteries were selected for the comparison test, discharged at constant power (2940 W) at different stages as shown in table 1.

Table 1 discharge setup

Product condition	Standard Positive positive product	Low performance product
Fully CHR state	Fully CHR state	Fully CHR state
Partially DCH	DCH (1.20 mins)	DCH (1.20 mins)
Low performance product less 1V/C	DCH (5.20 mins)	DCH (5.20 mins)
Standard product less 1V/C		DCH(11.19 mins)

Simulation analyses are carried out with the same discharge setup using the mathematical model for comparison. Figure. 6, shows the simulated results of the

behaviour comparisons of the standard and low performance products at different state of charge. Comparisons are made between the simulation results from the mathematical model and EIS experimental data. Correlation, in the main, is satisfactory but anomalies are present. Possible reasons for those anomalies are suggested. Overall the model results fit together with the EIS experimental data.

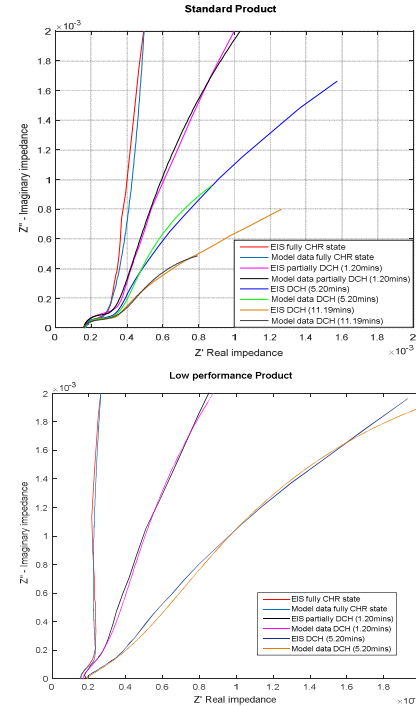
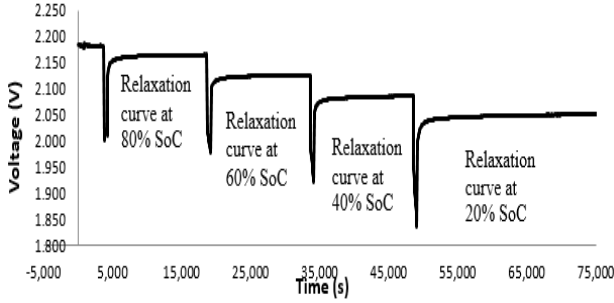


Fig.6 Nyquist diagram of behaviour comparisons of standard and low performance positive electrodes

Now, to determine that the mathematical model is a valid system to replicate real pore geometries, it was analysed to measure the difference between the standard and low performance product tested. The results indicated that the low performance pore radius was a much smaller compared with the standard product. This factor maybe the reason for the low efficiency of low performance product, since the smaller radius pore can easily cause a blocking effect that limited the electrolysis process.

The second part of the study was to prove that the more simplistic OCV prediction technique can be applied to low performance product. As the battery needs to be intact to ensure any generated results are representative of a production VRLA battery, the Silver-silver sulphate reference electrode technique was used. This technique measures across the ICC (internal cell connection) and the AGM (absorbed glass mat) to measure the voltage of each individual cell and is as unobtrusive as possible because access is gained through the vent valves. The advantages of the silver-silver sulphate technique over alternative methods are; a precisely defined electrode potential, relatively simple implementation and the availability in a variety of geometries at low cost [14]. As shows Fig.7 the cells were discharged at constant current to 80%, 60%, 40%, and 20% SoC. At the end of each 20% SoC, the cells were given a four hour rest period, where a voltage reading was taken every second. The cells under

investigation were discharged at different current rates 0.3C, 1C, and 3C to monitor the variation on relaxation resulting from the increase in load applied on the cell. The example shown in the Fig.8 illustrates the discharge curve of the low cell at 1C current rate. To ensure that the results were not discharge rate dependant, the tests were conducted at 0.3C, 1C and 3C for the low performance products. This was important because in a lead-acid cell,



the rate of discharge can drastically influence the amount of sulphuric acid consumed in the electrolyte.

Fig. 7 VRLA Cells discharge test setup

At low discharge rate, like 0.3C, the active material will be fully utilized by the discharge process, as the cells have enough time to utilize the significant quantity of strong electrolyte from the reservoir of the separator. During high rate discharge applications (like 3C rate), the chemical reaction occurs so fast that it does not permit the full utilization of the reservoir of electrolyte contained in the cell separator. Fig. 8 gives an overview of the cell OCV behaviour of two low performance products following a 1C discharge.

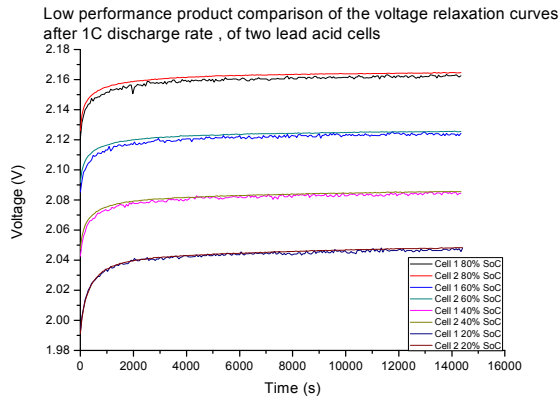


Fig. 8 Comparison of the voltage relaxation curves after 1C discharge, of two lead acid cells

The results of the low performance product OCV test have been compiled into Table 2. The cell voltage (V_{tr}) was measured after 30 minutes at open circuit, however, the true OCV measurement was not stable until the 240 minutes rest interval. For this type of cell the average voltage difference at open circuit from 30 to 240 minutes was 6.45 mV, and was therefore chosen as the K_v constant value. Equation (10) was then used to calculate the predicted OCV after the 240 minutes interval, so that

a comparison can be made with the measured (real) OCV after 240 minutes.

To aid in this explanation a worked example has been provided below for the 1C relaxation curve at 60% SoC.

$$V_{oc} = V_{tr} + K_v$$

$$V_{oc} = 2.118 + 0.00645$$

$$V_{oc} = 2.1245 \text{ OCV calculated}$$

$$\text{and } V_{real} = 2.1248 \text{ OCV real}$$

From this example the error between calculated and real was -0.0165%. Table 3 contains a comparison of the real and calculated OCV of the investigated cell for all test conditions. The maximum prediction error for the low performance product cell was calculated as $\pm 0.1623\%$ and can be seen in Table 3.

Table 2 Voltage measurements from low product relaxation test

Low performance SWL2500 Yuasa Battery Cell Voltage measurements from single cell relaxation test (V)						
Cell Test	0 min	30 min	60 min	120 min	180 min	240 min
0.3C 80% SOC	2.088	2.133	2.136	2.134	2.141	2.141
0.3C 60% SOC	2.074	2.091	2.092	2.095	2.096	2.097
0.3C 40% SOC	2.027	2.046	2.05	2.052	2.054	2.054
0.3C 20% SOC	1.974	2.003	2.006	2.009	2.011	2.012
1C 80% SOC	2.1226	2.1563	2.1584	2.1607	2.1623	2.1632
1C 60% SOC	2.0883	2.118	2.1208	2.1231	2.1237	2.1248
1C 40% SOC	2.048	2.0787	2.0822	2.0846	2.0855	2.0862
1C 20% SOC	2.0034	2.0422	2.0449	2.0484	2.0494	2.0504
3C 80% SOC	2.134	2.17	2.172	2.173	2.175	2.174
3C 60% SOC	2.115	2.147	2.148	2.151	2.152	2.152
3C 40% SOC	2.093	2.123	2.125	2.126	2.128	2.126
3C 20% SOC	2.065	2.096	2.099	2.1	2.101	2.101

Table 3 Comparison of the real and calculated OCV for the low product

Open Circuit Voltage (V)				
Cell Test	30 min	240 min (real)	240min (calc)	Error (%)
0.3C 80% SOC	2.133	2.141	2.1395	-0.0724
0.3C 60% SOC	2.091	2.097	2.0975	0.0215
0.3C 40% SOC	2.046	2.054	2.0525	-0.0755
0.3C 20% SOC	2.003	2.012	2.0095	-0.1267
1C 80% SOC	2.1563	2.1632	2.1628	-0.0208
1C 60% SOC	2.118	2.1248	2.1245	-0.0165
1C 40% SOC	2.0787	2.0862	2.0852	-0.0503
1C 20% SOC	2.0422	2.0504	2.0487	-0.0853
3C 80% SOC	2.17	2.174	2.1765	0.1127
3C 60% SOC	2.147	2.152	2.1535	0.0674
3C 40% SOC	2.123	2.126	2.1295	0.1623
3C 20% SOC	2.096	2.101	2.1025	0.0690

The worked example produced an error of just $\pm 0.1623\%$, so we have proved that the simple and effective OCV prediction method described in this paper is good in terms of performance. By contrast, Coroban et al. [15] proposed an online algorithm for prediction of the OCV. The accuracy of this method is dependent on the interval time used during the calculation (sampling time Δt) and on the gain derived experimentally (K factor).

The constant K is found using a calibration technique. The technique involves monitoring the battery voltage relaxation curve from the instant a rest period is entered to the time the cell reaches equilibrium. An average error of 1% to 5% was produced by using the methodology provided by [16]. A statistical analysis method can also be used to estimate the battery OCV as is evident in the work carried out by [17]. The work by [17] monitored the voltage curves following a charge step to specific SoC levels, with a varying current rate. In this case, the margin of error obtained when predicting the OCV was 8% to 10%. Bullock et al. [18] have proposed a model based on polynomial equations to obtain the relationship between acid molarity and OCV.

The authors used experimental OCV data acquired during a long self-discharge period at different ambient temperatures. This model is capable of predicting the OCV at each temperature condition to within an error of just 2%. Therefore, the simple and effective OCV prediction method, described throughout this research work, provided the OCV with the lowest error rate. This proves that the work in this paper can become a useful tool when used in concomitance with another mathematical method to predict the SoC and state of health (SoH) for electrochemical devices, with more precise results.

V CONCLUSIONS

In this paper, techniques for monitoring and predicting the OCV for VRLA battery system is described. Firstly, a mathematical model for the identification of real electrode pore geometry shape has been developed. This model is then used to identify the product to be tested via simulation analysis. For the model validation purpose, EIS experimental data was used. Comparisons are made between the simulation results from the mathematical model and EIS experimental data. Correlation, in the main, is satisfactory but anomalies are present. Possible reasons for those anomalies are suggested. Overall the model fit well with the EIS experimental data.

The second part is to determine the OCV to monitor the OCV-SOC of the selected product. To this end, a simple but effective methodology to predict the OCV after a small rest period is presented. It used a simple equation to predict the OCV. The OCV technique resulted in accurate readings when compared with other more complicated methods. By performing various tests using a simple equation it was possible to predict the OCV of a lead-acid cell. The equation was applied to the low performance product with a margin of error found from the real OCV and the calculated values was $\pm 0.1623\%$. Finally, it is concluded that from the results and the analysis presented in this paper, the simple techniques for monitoring SOH and predicting the OCV-SOC for VRLA battery systems give a greater accuracy.

References

- [1] Thanapalan, K., Williams, J. (2014). Development of a performance evaluation tool for hybrid vehicles system design. *i-manager's Journal on Instrumentation & Control Engineering*, 2 (4), 6 – 13.
- [2] Bitterlin, I. F. (2004). Standby-battery autonomy versus power quality. *Journal of Power Source*, 136 (2), 351-355.
- [3] Stephen, D. (1999). The K_d Model, Methods of Measurement, and Application of Chemical Reaction Codes. Office of Environmental Restoration U.S. Department of Energy, Washington, DC 20585.
- [4] Mariani, A., Thanapalan, K., Stevenson, P., Williams, J. (2013). Techniques for estimating the VRLA batteries ageing, degradation and failure modes. In 19th Int. Conf. on Automation and Computing, UK, 43-47.
- [5] Ribeiro, P. F., Johnson, B. K., Crow, M. L., Arsoy, A., Liu, Y. (2001). Energy Storage Systems for Advanced Power Applications. *Proceedings of the IEEE*, 89 (12), 1744-1746.
- [6] Divya, K. C., Ostergaard, J. (2009). Battery energy storage technology for power systems—An overview. *Electric Power Systems Research*, 79 (4), 511-520.
- [7] Skundin, A. M., Tsirlina, G. A. (2014). V. S. Bagotsky's contribution to modern electrochemistry. *Journal Solid State Electrochem.* 18:1147–1169 DOI 10.1007/s10008-014-2480-5.
- [8] Keiser, H., Beccu, K. D., Gutjahr, M. A. (1976). Abschätzung der porenstruktur poröser elektroden aus impedanzmessungen. *Electrochimica Acta*, 21 (8), 539-543.
- [9] DE Levie, R. (1989). On the impedance of electrodes with rough interfaces. *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry*, 261 (1) 1-9.
- [10] Mariani, A., Stockley, T., Thanapalan, K., Williams, J., Stevenson, P. (2014). Simple and Effective OCV Prediction Mechanism for VRLA Battery Systems. In the Proceedings of the 3rd International Conference on Mechanical Engineering and Mechatronics, Prague, Czech Republic, 1-10.
- [11] Masayuki, I., Keiichirou, H., Yoshinao, H., Isao, S. (2015). In-situ EIS to determine impedance spectra of lithium-ion rechargeable batteries during charge and discharge cycle. *Journal of Electroanalytical Chemistry*, 737 (1), 78-84.
- [12] Thanapalan, K., Williams J. G., Premier, G. C., Guwy, A. J. (2011). Design and Implementation of Renewable Hydrogen Fuel Cell Vehicles. *Renew. Energy Power Qual. J.*, 9, 310-315.
- [13] Barsali, S., Ceraolo, M. (2002). Dynamical Models of Lead-Acid Batteries: Implementation Issues. *Transactions on energy conversion, IEEE*, 17 (1), 16-23.
- [14] Ruetschi, P. (2003). Silver-silver sulphate reference electrodes for use in lead-acid batteries. *Journal of Power Sources*, 116 (1-2), 53-60.
- [15] Coroban, V., Boldea, I., Blaabjerg, F. (2007). A novel on-line state-of-charge estimation algorithm for valve regulated lead-acid batteries used in hybrid electric vehicles. *International Aegean Conference on Electrical Machines and Power Electronics ACEMP, Bodrum, Turkey*, 39-46.
- [16] Huet, F. (1998). A review of impedance measurements for the determination of the state-of-charge or state-of-health of secondary batteries. *Journal of Power Sources*, 70 (1), 59-69.
- [17] Sato, S., Kawamura, A. (2002). A New Estimation Method of State of Charge using Terminal Voltage and Internal Resistance for Lead Acid Battery. *Power Conversion Conference, PCC-Osaka*, 2, 565-570.
- [18] Bullock, K. R., Weeks, Bose, C. S. C., Murugesamoorthi, K. A. (1997). A predictive model of the reliabilities and the distributions of the acid concentrations, open-circuit voltages and capacities of valve-regulated lead/ acid batteries during storage. *Journal of Power Sources*, 64 (1-2), 139-145.

A Study of the Performance of the Combination of Energy Storage Fibres

Ruirong Zhang¹, Yanmeng Xu¹, David Harrison¹, John Fyson¹, Darren Southee² and Anan Tanwilaisiri¹

¹Cleaner electronics, College of Engineering, Design and Physical Sciences, Brunel University London, Uxbridge, UK

²Loughborough Design School, Loughborough University, Leicestershire, UK

Abstract—Fibre supercapacitors were manufactured using a motor-driven setup. Their electrochemical properties were characterised. The performance of two fibre supercapacitors in series or in parallel mainly followed the expected theoretical models. The electrochemical potential window of a series circuit of two fibre supercapacitors is 1.6 V, which is twice of the working potential of the individual fibre supercapacitors, and the charge-discharge current of two fibre supercapacitors in parallel is also twice of the current for a single one.

Keywords—fibre supercapacitors; energy storage fibre; series; parallel

I. INTRODUCTION

Supercapacitors, as energy storage devices, have a high power density, a high reversibility and a long cycle life. Recently, supercapacitors have been considered as a promising energy source for delivering peak power demands in electric vehicles, emergency power supplies and portable electronic devices [1-4]. The future trend of energy storage design is towards developing low-cost, light-weight, environmentally friendly and flexible energy storage devices. Some attempts have been made to manufacture flexible/wearable supercapacitors [2-7]. Bae et al. developed a high-efficiency fibre-based supercapacitor using ZnO nano-wires as electrodes. The supercapacitor with two fibre-based electrodes showed 0.21 mFcm⁻² and 0.01 mFcm⁻¹ in 1M KNO₃ electrolyte and a high specific capacitance of 2.4 mF/cm² and 0.2 mF/cm using PVA/H₃PO₄ as a gel-electrolyte [3]. Fu et al. presented a novel flexible fibre supercapacitor which consisted of two fibre electrodes using Chinese ink as the active material. A helical spacer wire inserted between two electrodes enables the efficient separation of the two fibre electrodes. This two-fibre structure supercapacitor showed good areal capacitance of 11.9-19.5 mFcm⁻² [5]. Our group has developed a coaxial fibre supercapacitors using Chinese ink as the active material [6, 7]. The manufacturing method has been improved from a hand-made process to a motor-driven apparatus for dip-coating to manufacture energy storage fibres. It has been shown that our fine fibre supercapacitors have good flexibility [6, 7], which is a big advantage for energy storage fibres those are to be woven into smart fabrics. However, this fibre supercapacitor is very fine and the energy storage ability is limited by the lack of enough active material. The working potential is limited because an aqueous electrolyte (H₃PO₄/PVA) was used in this flexible

structure [8]. In order to meet the requirements for the potential uses, the performance of combined circuits of the fiber supercapacitors in parallel or in series was studied in this work.

II. EXPERIMENTAL

A. Materials

Phosphoric acid (H₃PO₄, anhydrous) and polyvinyl alcohol (PVA, MW 146,000-186,000, > 99% hydrolyzed) were used without further purification. The gel electrolyte was made by dissolving 0.8 g H₃PO₄ and 0.8 g PVA in 10 mL deionized water. Another PVA gel used as an outer layer material was made by dissolving 1 g PVA in 10 mL deionized water. Copper wire (100 µm in diameter) was used as the core conductor material and conducting wire. Commercial Chinese ink was used as the active coating material.

B. Manufacturing of the energy storage fibre

Based on the working mechanism described elsewhere [1], fibre supercapacitors have been designed and manufactured [6, 7]. The coaxial single fibre supercapacitors consist of six layers, see Fig. 1. The central metal fibre and the silver paint are current collectors. Two active layers are made of Chinese ink and separated by a gel electrolyte layer, which serve as the energy storage electrodes. The outer PVA layer is a support layer to protect the whole structure and avoid the electrolyte evaporating.

An experimental setup of the dip coating method was designed and produced to manufacture energy storage fibre [7]. In the coating process, the thickness of each layer was controlled by the speed of the two-direction motor and times of coatings for each layer. In this work, each layer of the materials was coated 12, 3, 4, 2 and 2 times from inner to outer.

C. Characterisation of the electrochemical properties

The electrochemical performance of the energy storage fibre developed was characterised by galvanostatic charge/discharge (GCD), cyclic voltammetry (CV) and electrochemical impedance spectroscopy using a VersaSTAT 3 electrochemical workstation.

The capacitance was measured using the CV test and can be calculated using the following equation as:

$$C = \frac{Q_{\text{total}}/2}{\Delta V} \quad (1)$$

Where C is the capacitance and Q_{total} is the supercapacitor's charge in coulombs.

From a galvanostatic charge/discharge test, the capacitance C can be directly calculated from the following equation:

$$C = \frac{i \times \Delta t}{\Delta V} \quad (2)$$

Where i is the discharge current in amperes (A); ΔV is the voltage of the discharge (V).

III. RESULTS AND DISCUSSION

The fibre supercapacitors were manufactured and characterised [6, 7]. The highest specific capacitance that was reached was 34.5 mF/cm. The fiber supercapacitor with a thicker active layer also showed a good flexibility. The capacitance changed slightly when it was woven into a fabric or was bent on a glass rod [7].

For a single fiber supercapacitor, the energy storage ability and operating potential is limited by the decomposing potential of water or phosphoric acid. In order to meet the requirements for a higher potential use, the performance of combination circuits of the fiber supercapacitors in parallel and in series has been studied.

Fig. 1 shows the schematic structure of two single fibre supercapacitors (samples **a** and **b**) and their series and parallel combination circuits. In order to protect the fine fibre supercapacitor, a PVA gel was coated as the outer support layer. The copper core fibre and the silver paint layer were the current collectors, so another copper wire used as a conductor was glued with the silver paint layer. As shown in Fig. 2, the working potential of the two fibre supercapacitors in parallel is the same as the working potential of the individual single samples, and half of that of samples in series (1.6 V). The areas of CV curve of the parallel circuits is close to the total area of the two samples **a** and **b**. The area of CV curve of the series circuits would be half of the total area of these two samples. The capacitances calculated by equation (1) from the CV curves are shown in Table 1. The capacitance of the parallel circuit is 12.0 mF, which is almost the same as the sum of the capacitances (12.5 mF) of the two single samples. The capacitance of series circuit is 3.4 mF, which is higher than the theoretically expected series capacitance of 2.3 mF. As shown in Table 1, there is a considerable difference between the electrical series resistance and capacitance of sample **a** and sample **b** because they are of different length. Therefore the voltage is not equal across both capacitors which might account for the difference between the experimental and theoretical capacitance for the series case. These results illustrate that the electrical properties of the parallel combinations of the fibre supercapacitors showed a reasonable agreement with the theoretical models of series circuit combinations.

Fig. 3 shows the galvanostatic charge-discharge curves at the charge current of 0.2 mA for the two single

supercapacitors and the series circuit, and of 0.4 mA for the parallel circuit. The potential window of parallel circuit is set at 0.8 V, the same as the two single fibre supercapacitors, and half of the series circuit. As shown in the Table 1, the capacitances of the different cases calculated by equation (2) from the charge-discharge curve are generally following the trend which is in

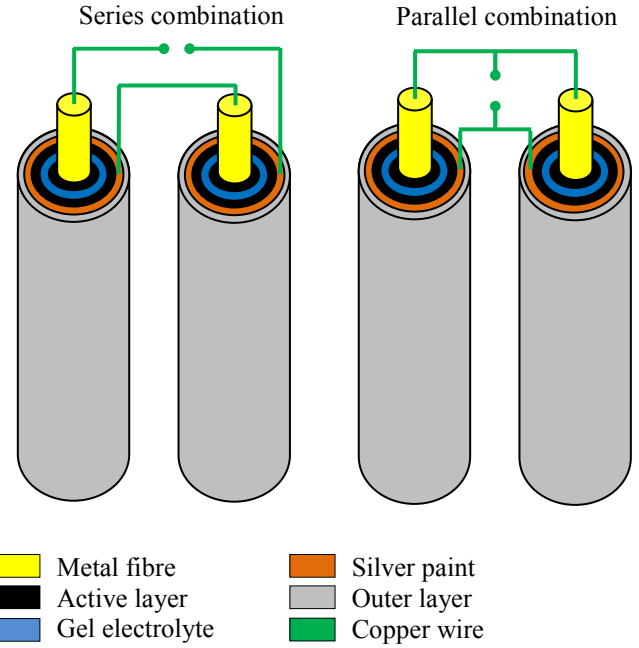


Figure 1. The schematic of two fibre supercapacitors combined in series or in parallel circuits.

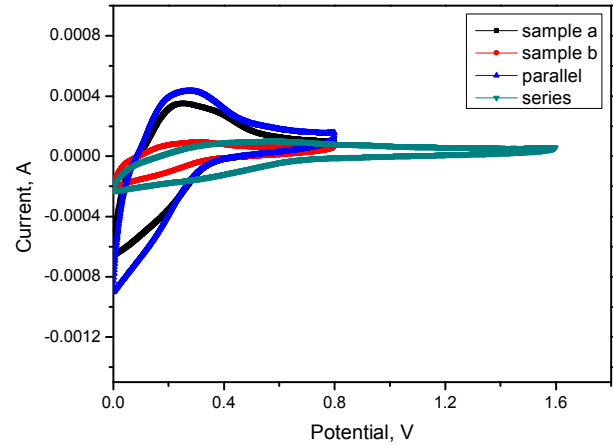


Figure 2. Cyclic voltammograms recorded at 20 mVs⁻¹ for two single fibre supercapacitors and their electrical combinations in series and parallel.

agreement with the theoretical models. Nyquist plots of the two single samples and the combinations in series or in parallel are shown in Fig. 4. It can be seen that the shapes of electrochemical impedance curves for the combination circuits were similar to those of the single fibre supercapacitors. This indicates that the ion diffusion processes for all samples are similar. The high intercept on the real axis is the electrical series resistance (ESR), which is shown in Table 1. For the case of the series circuit, the ESR is 75.2 Ω . This is very close to the sum of the ESR of the two samples **a** and **b**; the ESR of the parallel circuit is

about 15.1Ω , which is slightly less than the expected ESR of the parallel circuits of sample **a** and **b**. These results had a reasonable agreement with the theoretical models.

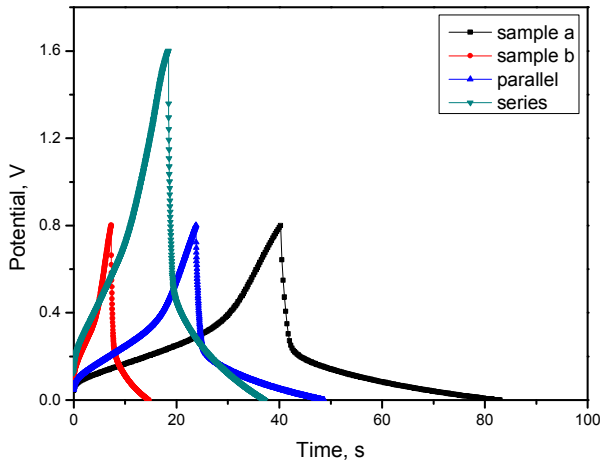


Figure 3. Galvanostatic charge-discharge curves of the 4th cycles of each case, for single and series circuit 0.2 mA was used, and 0.4 mA for the parallel circuit.

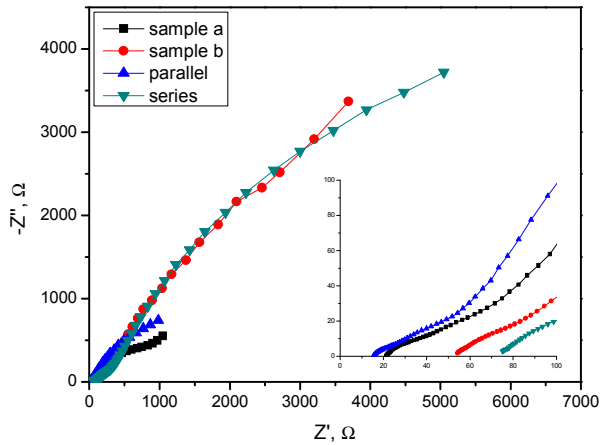


Figure 4. Electrochemical performances of two single fibre supercapacitors and their electrical combinations in series and parallel using a 4 mV AC modulation for a frequency range of 100 kHz to 0.01 Hz (insert shows an enlarged curve at a high-frequency region).

Table 1. Summary of electrical series resistances (ESR) from Nyquist plots and the capacitances from CV curve and galvanostatic charge/discharge curve with those calculated using theory for series and parallel circuits.

	Capacitance, mF (CV)		Capacitance, mF (GCD)		IMP, Ω	
	exp.	theory	exp.	theory	Exp.	theory
Sample a	9.4		13.2		21.2	
Sample b	3.1		2.2		54.3	
Series connection	3.4	2.3	2.8	1.9	75.2	75.5
Parallel connection	12.0	12.5	13.8	15.4	15.1	15.2

IV. CONCLUSIONS

Fibre supercapacitors were manufactured by a dip coating setup. The electrical performance of a

combination of two fibre supercapacitors in series or parallel has been studied. The current and potential range can be improved by connecting multiple fibre supercapacitors in parallel or in series to meet the power and energy requirements. This shows that these kinds of flexible energy storage fibres can be woven into other fabrics and connected into series or parallel to make smart textile materials with appropriate electrical performance for a desired use.

REFERENCES

- [1] R. Kötzt, M. Carlen, "Principles and applications of electrochemical capacitors," *Electrochim. Acta*, Vol. 45, pp. 2483-2498, 2000.
- [2] V.L. Pushparaj, M.M. Shaijumon, A. Kumar, S. Murugesan, L. Ci, R. Vajtai, R.J. Linhardt, O. Nalamsu and P.M. Ajayan, "Flexible energy storage devices based on nanocomposite paper," *Proc. Natl. Acad. Sci.* Vol. 104, pp. 13574-13577, 2007.
- [3] J. Bae, M.K. song, Y.J. Park, J.M. Kim, M. Liu and Z.L. Wang, "Fibersupercapacitors made of nanowire-fiber hybrid structures for wearable/flexible energy storage," *Angew. Chem. Int. Ed.* Vol. 50, pp. 1683-1687, 2011.
- [4] G. Milczarek, A. Ciszewski and I. Stepniak, "Oxygen-doped activated carbon fiber cloth as electrode material for electrochemical capacitor," *J. Power Sources*. Vol. 196, pp. 7882-7885, 2011.
- [5] Y.P. Fu, X. Cai, H.W. Wu, Z.B. Lv, S.C. Hou, M. Peng, X. Yu and D.C. Zou, "Fiber Supercapacitors Utilizing Pen Ink for Flexible/Wearable Energy Storage," *Adv. Mater.* Vol. 24, pp. 5713-5718, 2012.
- [6] D. Harrison, F. Qiu, J. Fyson, Y. Xu, P. Evans and D. Southee, "A coaxial single fibre supercapacitor for energy storage," *Phys. Chem. Chem. Phys.*, Vol. 15, pp. 12215-12219, 2013.
- [7] R. Zhang, Y. Xu, D. Harrison, J. Fyson, D. Southee and A. Tanwilaisiri, "Fabrication and Characterisation of Energy Storage Fibres," Paper presented at the 20th International Conference on Automation & Computing, IEEE, Cranfield, UK, pp. 228-230, September 2014.
- [8] X.Y. Zhang, X.Y. Wang, L.L. Jiang, H. Wu, C. Wu and J.C. Su, "Effect of aqueous electrolytes on the electrochemical behaviors of supercapacitors based on hierarchically," *J. Power Sources*. Vol. 216, pp. 290-296, 2012.

Study on the Inherent Characteristics of Planetary Gear Transmissions

Qing Tao^{1,2}, Jianxing Zhou, Wenlei Sun, Jinsheng Kang²

¹School of Mechanic engineering
Xinjiang University
Urumqi, CHINA 830046

²College of Engineering, Design and Physical Sciences,
Brunel University,
Uxbridge, Middlesex, UB8 3PH, UK

xjutao@qq.com; jianzhou82923@163.com; Sunwenxj@163.com; jinsheng.kang@brunel.ac.uk

Abstract— This article describes a novel modeling method for a rigid-flexible coupled planetary gear transmission system, which consists of rigid and elastic ring gear subsystems. The rigid gear subsystem model was established using the lumped parameter method while the elastic ring gear subsystem was established using the finite element method. According to the compatible state of deformation between mesh force and ring gear deformation, the natural frequency and vibration modes were determined, and the distribution rules of the natural frequency for this coupled system were established. The vibration modes were classified into six different modes based on the vibration characteristics of the system. This article provides a theoretical basis for the dynamic design of a rigid-flexible coupled planetary gear transmission system.

Keywords- Planetary gear transmission; Dynamic model; Rigid-flexible coupling; Natural frequency; Sensitivity

I. INTRODUCTION

The planetary gear transmission system is widely used in various machines and machinery equipment because of its compact structure, wide transmission ratio range, and high transmission efficiency. The mechanical behaviors and performance of a planetary gear transmission system play an important role in the performance of the whole machine.

Planetary gear transmission has become an indispensable key component of the main power system of high-speed heavy-duty gearings (e.g. Wind Turbine Gearbox). Shown in Fig.1.

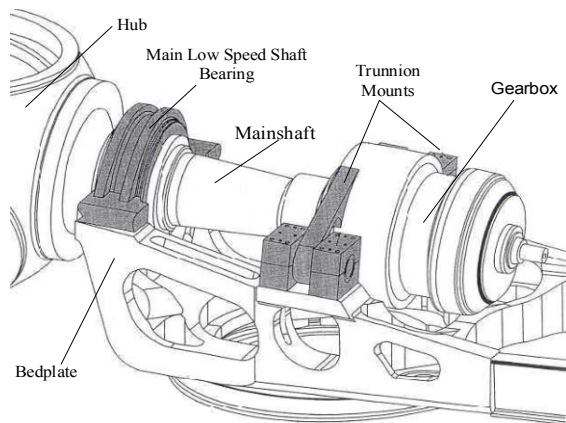


Figure 1. Typical Gearbox Mounting

The stable operation and dynamic performance of this transmission system directly determine the overall performance of the device. Studies on the dynamic characteristics of planetary gear transmission systems have been reported at home and abroad. The lumped parameter method is the most common modeling method used in existing relevant studies because of its accuracy and short calculation time[1]. Considering the effect of ring gear flexibility on the dynamic characteristics of planetary gear transmission systems, scholars have started to explore new modeling methods. Ambarisha et al.[2] established a model of a planetary gear transmission system using the FEA method, which simulates gear meshing through two-dimensional contact. A unique semi-FEA formula was used to simplify the calculation. Singh et al.[3] further established a three-dimensional finite element contact model to implement a systematic analysis on the load sharing characteristics of the system. Bajer et al. [4] established a dynamic contact model of a planetary gear transmission system, which consists of rigid and flexible components that are interconnected in a contact manner. Song Yimin et al. [5] divided the continuous flexible ring gear into rigid sections connected by equivalent springs. The flexibility of the ring gear reduces the low-order natural frequency of the system. Such reduction in size is related to the installation of a ring gear. Wu et al. [6] discovered that the flexibility of a ring gear introduces nodal diameter vibration mode into the system. This vibration mode can form a coupling effect with the vibration of other components. Kahraman et al. [7] studied the effect of ring gear thickness on its deformation, stress, and load sharing coefficient. Blending is the main deformation mode of a ring gear with a thin rim, and the maximum stress gradually shifts from the tooth socket to the tooth root as the rim thickens. However, rim thickness has no significant effect on load sharing performance. Avinash [8] discussed the effect of the rigidity of the planet and sun gears on load sharing performance. Although the distributed mass model can effectively present ring gear flexibility, the calculation is time consuming, and most existing studies adopted quasi-dynamic calculation [9,10].

The present study proposes a novel modeling method for a rigid-flexible coupled planetary gear transmission system. The natural frequency distribution law and vibration mode of the system are summarized.

II. ANALYTICAL MODEL

The planetary gear transmission system analyzed in this work is the 2K-H system, which includes three planet gears with uniform distribution. The mechanical model of the system is shown in Fig. 1, where r is the ring gear, s is the sun gear, p is the planet gear, c is the planet carrier, k is the rigidity, and θ is the intersection angle.

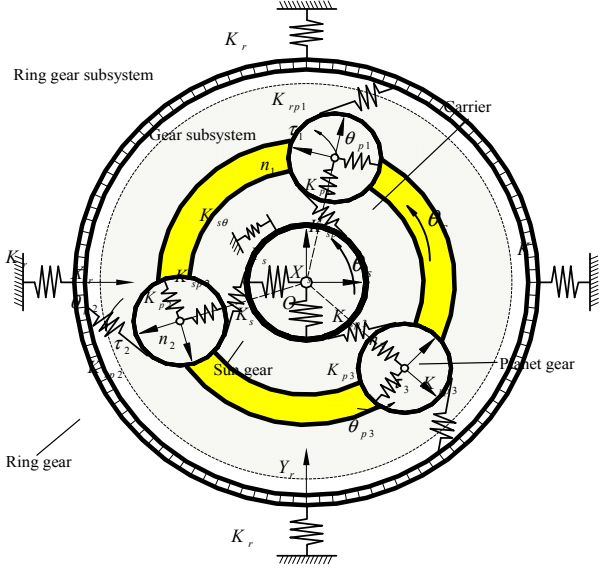


Figure 2. Mechanical model of the planetary gear transmission system

The support stiffness, torsional rigidity, and mesh stiffness of all the components were replaced by a spring. The center of the planet carrier was taken as the origin of coordinates, whereas the horizontal and vertical directions were viewed as the X- and Y-axes for system modeling. k_s is the flexible support stiffness of the sun gear, and $K_{s\Box}$ is the torsional rigidity of the sun gear. The planet gears are supported by bears. k_{pi} is the support stiffness of the i^{th} ($i = 1, 2, 3$) planet gear, and k_r is the support stiffness of the ring gear. The ring gear is fixed through four symmetric positions. In the model, k_{spi} is the mesh stiffness value between the sun and planet gears, and k_{tpi} is the mesh stiffness value between the planet and ring gears, respectively.

The rigid gear subsystem is composed of a sun gear, planet gears, and a planet carrier. In the generalized coordinates, x and y represent the lateral micro-displacement, and θ is the torsional micro-displacement. The finite element model of the ring gear was established using the finite element method. The coupling model between the rigid gear subsystem model and the ring gear model can be established through Equation (1).

$$\begin{bmatrix} \mathbf{M}_g & 0 \\ 0 & \mathbf{M}_e \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{X}} \\ \ddot{\mathbf{u}} \end{Bmatrix} + \begin{bmatrix} \mathbf{c}_g & 0 \\ 0 & \mathbf{c}_e \end{bmatrix} \begin{Bmatrix} \dot{\mathbf{X}} \\ \dot{\mathbf{u}} \end{Bmatrix} + \begin{bmatrix} \mathbf{k}_g & 0 \\ -\mathbf{R} & -\mathbf{k}_a + \mathbf{k}_e \end{bmatrix} \begin{Bmatrix} \mathbf{X} \\ \mathbf{u} \end{Bmatrix} = \begin{Bmatrix} \mathbf{P}(t) \\ \mathbf{F}_e(t) \end{Bmatrix} \quad (1)$$

where $[-\mathbf{R}]$ and $[-\mathbf{k}_a]$ are the coupled matrixes. The basic parameters of the planetary gear transmission system are listed in Table 1. The moment of inertia and the mass of all the gears were determined based on a solid

modeling using the three-dimensional UG NX software. The support stiffness of the components and the mesh stiffness between gears were calculated using the FEM [11].

TABLE I. PARAMETERS OF THE PLANETARY GEAR TRANSMISSION

SYSTEM			
Parameters	Sun gear	Planet gear	Ring gear
Number of teeth	35	31	97
Tooth width (mm)	42	42	42
Modulus (mm)	3	3	3
Pressure angle (°)	28	28	28
Modification coefficient	-0.02	0.02	0.02
Mass (kg)	2.7	2.5	5.3
Moment of inertia (kg·m ²)	0.008	0.007	0.178
Support stiffness (N/m)	8×10^7	7.5×10^8	2.7×10^9
Torsional rigidity (N·m/rad)	7.7×10^9	—	—
Comprehensive mesh stiffness (N/m)	8.3×10^8	9×10^8	

III. INHERENT CHARACTERISTIC ANALYSIS OF THE COUPLED SYSTEM

A. Natural frequency of the system

The inherent characteristics of the system can be converted into an eigenvalue problem.

$$([K] - \omega_i^2 [M])\{\phi_i\} = \{0\}$$

Then,

$$\omega_i^2 [M]\{\phi_i\} = [K]\{\phi_i\}$$

where ω_i is the natural frequency of the i^{th} order, and ϕ_i is the vibration mode of the i^{th} order.

The rigid-flexible coupled planetary gear transmission system includes the transverse and torsional degree of freedom (DOF) of the sun and planet gears as well as the DOF of the nodes on the flexible ring gear. In this case, the total natural frequency of the system increases significantly. The natural frequency can be divided into two frequency bands: 1) the low-frequency band, which refers to the coupled vibration frequency band between the rigid and elastic ring gear subsystems; and 2) the high-frequency band, which is the vibration frequency band of the ring gear. The low-frequency band includes 36 orders of the natural frequency of the system. According to Table 2, the natural frequency ranges between 274.28 and 7,653 Hz. The coupling of the ring gear and the drive system effectively enhances system flexibility. Consequently, the low-order natural frequency significantly decreases compared with the calculated result of the traditional model.

B. Vibration mode of the coupled system

The vibration mode of the rigid-flexible coupled planetary gear transmission system is considerably complicated. Coupled vibration between the rigid and

elastic ring gear subsystems occurs in the first 36 orders of natural frequency. It can be divided into the following according to the system vibration characteristics: pure nodal diameter vibration of the ring gear, coupling of the nodal diameter vibration of the ring gear and the transverse vibration of the central gear, coupling of the nodal diameter vibration of the ring gear and the torsional vibration of the central gear, coupling of the local ring gear vibration and the transverse vibration of the central gear, gear vibration, and local ring gear vibration.

TABLE II. NATURAL FREQUENCY OF THE SYSTEM

Vibration type	Frequency	Vibration mode
Pure nodal diameter vibration of ring gear	274.28	three-nodal diameter vibration
	1532.8	five-nodal diameter local vibration
	1724.3	six-nodal diameter vibration
	2760.9	seven-nodal diameter vibration
	3619.3, 3809.2, 3918, 3944.1, 4383	eight-nodal diameter vibration
	5390.2	nine-nodal diameter vibration
	313.46, 411.86, 556.64	Coupling of three-nodal diameter vibration of ring gear and transverse vibration of central gear
	876.99	Coupling of four-nodal diameter vibration of ring gear and transverse vibration of central gear
	1325	Coupling of local five-nodal diameter vibration of ring gear and transverse vibration of central gear
	1809.8	Coupling of six-nodal diameter vibration of ring gear and transverse vibration of central gear
Coupling of nodal diameter vibration of ring gear and transverse vibration of central gear	4173	Coupling of eight-nodal diameter vibration of ring gear and transverse vibration of central gear
	4947.3, 5162.2	Coupling of nine-nodal diameter vibration of ring gear and transverse vibration of central gear
	2956.6	Coupling of seven-nodal diameter vibration of ring gear and torsional vibration of central gear
	6776, 7653	Coupling of nine-nodal diameter vibration of ring gear and torsional vibration of central gear
Coupling of local ring gear vibration and transverse vibration of central gear	1164.4, 1301	Coupling of local ring gear vibration and transverse vibration of central gear

Gear vibration	6411	Torsional transverse vibration of central gear
Local ring gear vibration	997.68, 1406.1, 2282.3, 2460.2, 2586.6, 4172.9, 4382.7, 4473.6, 4815.4, 5979.3, 6231	Local ring gear vibration
Structural vibration of ring gear	⋮	Structural vibration of ring gear

These six vibration modes have the following characteristics:

1) Pure nodal diameter vibration of the ring gear

Only the nodal diameter vibration of the ring gear is observed. The other gears do not show any obvious vibration. Figure 3(a) shows the three-nodal diameter vibration of the ring gear.

Six types of the pure nodal diameter vibration of the ring gear are observed in the first 36 orders of natural frequency. Ten natural frequencies exist. The three-, five-, six-, seven-, and nine-nodal diameter vibrations have one single natural frequency, whereas the eight-nodal diameter vibration has five single natural frequencies.

2) Coupling of the nodal diameter vibration of the ring gear and the transverse vibration of the central gear

The coupling of the nodal diameter vibration of the ring gear and the transverse vibration of the central gear is observed. The torsional amplitude of the sun gear and planet carrier is zero. No transverse and torsional vibrations are observed on the ring gear. Figure 3(b) shows the coupling of the three-nodal diameter vibration of the ring gear and the transverse vibration of the central gear.

This vibration mode has six coupled vibrations with nine single-natural frequencies—three for the coupling of the three-nodal diameter vibration of the ring gear and the transverse vibration of the central gear, two for the coupling of the nine-nodal diameter vibration of the ring gear and the transverse vibration of the central gear, and one for the coupling of the four-nodal diameter vibration (five-, six-, and eight-nodal diameter vibrations) of the ring gear and the transverse vibration of the central gear.

3) Coupling of the nodal diameter vibration of the ring gear and the torsional vibration of the central gear

The coupling of the nodal diameter vibration of the ring gear and the torsional vibration of the central gear is observed. The transverse amplitude of the sun gear and the planet carrier is zero. No transverse and torsional vibrations are observed on the ring gear. Figure 3(c) shows the coupling of the nine-nodal diameter vibration of the ring gear and the torsional vibration of the central gear.

This vibration mode has two coupled vibrations with three natural frequencies, including two single natural frequencies for the coupling of the nine-nodal diameter vibration of the ring gear and the torsional vibration of the central gear as well as one single natural frequency for the coupling of the seven-nodal diameter vibration of the ring gear and the torsional vibration of the central gear.

4) Coupling of the local ring gear and the transverse vibration of the central gear

Local vibration is observed on the ring gear (some nodes are deformed, whereas the rest of the positions remain normal during vibration). This vibration is coupled with the transverse vibration of the central gear. However, the torsional amplitude of the sun gear and the planet carrier is zero. Figure 3(d) shows such coupling of the local ring gear vibration and the transverse vibration of the central gear.

This vibration mode has two coupled vibrations, which can be distinguished according to the vibration mode of the ring gear; both modes have one natural frequency.

5) Gear vibration

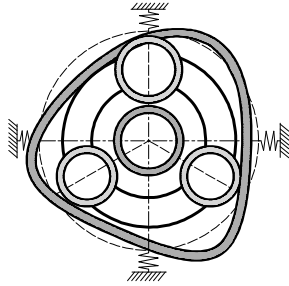
Only the overall ring gear vibration occurs; no structural vibration of the ring gear is observed. Figure 3(e) shows that the gear vibration mode is the same with the vibration mode of the rigid ring gear.

The torsional transverse vibration of the central gear is the only one gear vibration mode, which has one single natural frequency.

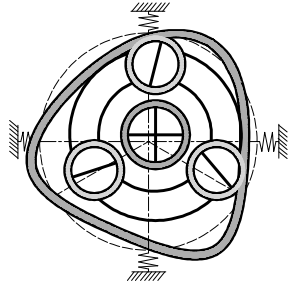
6) Local ring gear vibration

Only the local structure of the ring gear vibration is observed; the transverse and torsional amplitudes of the central gear are not obvious [Fig. 3(f)].

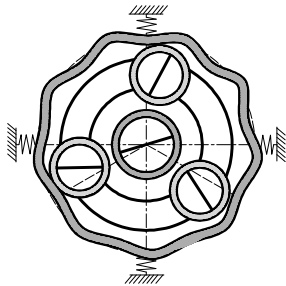
This vibration mode shows [11] single natural frequencies.



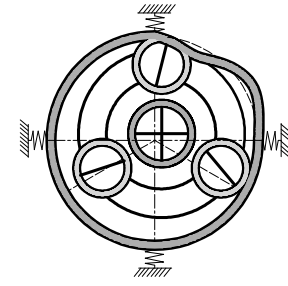
(a) Three-nodal diameter vibration of the ring gear



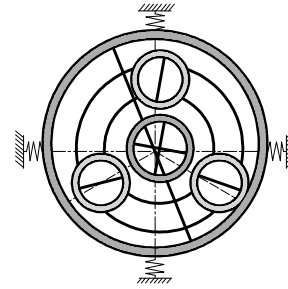
(b) Three-nodal diameter vibration of the ring gear + transverse vibration of the central gear



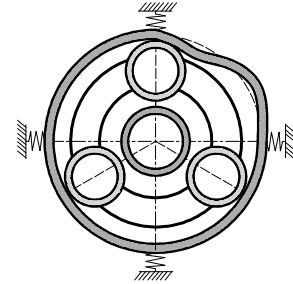
(c) Nine-nodal diameter vibration of the ring gear + torsional vibration of the central gear



(d) Local ring gear vibration + transverse vibration of the central gear



(e) Torsional transverse vibration of the central gear



(f) Local ring gear vibration

Figure 3. Vibration modes of the planetary transmission system when considering ring gear flexibility

IV. CONCLUSIONS

When ring gear flexibility is considered, the total natural frequency of the planetary gear system increases significantly. Meanwhile, the coupled vibration of the rigid gear subsystem and elastic ring gear subsystem is developed. The coupled system always produces a natural frequency near the torsional vibration frequency of the rigid gear subsystem under different ring gear thicknesses. However, ring gear vibration gradually changes from high-nodal diameter vibration to low-nodal diameter vibration as the ring gear thickens.

REFERENCES

- [1] Bu Zhonghong, Liu Geng, Wu Liyan. Planetary gear transmission dynamics is reviewed[J]. Journal of Vibration and Shock, 2010, 29(9): 161~165.

- [2] Ambarisha V.K., Parker R.G.. Nonlinear dynamics of planetary gears using analytical and finite element models[J]. Journal of Sound and Vibration, 2007, 302: 577~595.
- [3] Singh A.. Application of a system level model to study the planetary load sharing behavior [J]. Journal of Mechanical Design, 2005, 127(12): 469~476.
- [4] Bajer A., Demkowicz L. Dynamic contact/impact problems, energy conservation, and planetary gear trains[J]. Computer methods in applied mechanics and engineering, 2012, 191: 4159~4191.
- [5] Zhang Jun, Song Yimin, Wang Jianjun. Dynamic modeling for spur planetary gear transmission with flexible ring gear[J]. Chinese Journal of Mechanical Engineering, 2009, 45(12): 29~36.
- [6] Wu X., Parker R.G.. Modal properties of planetary gears with an elastic continuum ring gear[J]. Journal of Applied Mechanics, 2008, 75: 1~12.
- [7] Kahraman A., Ligata H., Singh A. Influence of ring gear rim thickness on planetary gear set behavior[J]. Journal of Mechanical Design, 2010, 132: 021002-1~8.
- [8] Avinash S. Application of a system level model to study the planetary load sharing behavior[J]. Journal of Mechanical Design, 2005, 127(1): 469~476.
- [9] Zhou Jianxing, Liu Geng, Ma Shangjun. Vibration and noise analysis of the gear transmission system[J]. Journal of Vibration and Shock, 2011, 30(6): 234~238.
- [10] Snežana Ćirić Kostić, Milosav Ognjanović. The noise structure of gear transmission units and the role of gearbox walls[J]. FME Transactions, 2007, 35: 105~112.
- [11] Tuma J. Gearbox noise and vibration prediction and control[J]. International Journal of Acoustics and Vibration, 2009, 14(2): 1~11.

Comparison of Contact Measurement and Free-Space Radiation Measurement of Partial Discharge Signals

A Jaber¹, P Lazaridis¹, Y Zhang¹, D Upton¹, H Ahmed¹, U Khan¹, B Saeed¹,
P Mather¹, M F Q Vieira², R Atkinson³, M Judd⁴, and I A Glover¹

¹Department of Engineering & Technology, University of Huddersfield, Huddersfield HD1 3DH, UK

²Department of Electrical Engineering, Universidade Federal de Campina Grande, Campina Grande, Brazil

³Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK

⁴High Frequency Diagnostics & Engineering Ltd, Glasgow G2 6HJ, UK

E-mail: Adel.Jaber@hud.ac.uk

Abstract—Two partial discharge (PD) measurement techniques, a contact measurement technique (similar to the IEC 60270 standard measurement) and a free-space radiation (FSR) measurement technique, are compared for the case of a floating electrode PD source. The discharge pulse shapes and PD characteristics under high voltage DC conditions are obtained. A comparison shows greater similarity between the two measurements than was expected. It is inferred that the dominant mechanism in shaping the spectrum is the band-limiting effect of the radiating structure rather than band limiting by the receiving antenna. The cumulative energies of PD pulses in both frequency and time domains are also considered.

Keywords- *Electromagnetic wave radiated; IEC 60270 measurement; Partial discharge; Radio frequency; Ultra high frequency.*

I. INTRODUCTION

Many electricity supply organizations around the world are facing growing energy demands and an ageing transmission and distribution infrastructure. The financial cost of replacing infrastructure is high which motivates cost-effective asset management of to maximize its useful lifetime whilst minimizing the risk of failure with consequent (large) costs. To facilitate efficient and reliable operation, continuous condition monitoring of the electrical equipment within substations is required.

PD measurement is an effective method to diagnose imminent failures due to insulation degradation with consequent substation outages [1].

IEC60270 is a standard for measuring the apparent charge of a PD pulse. It is a contact method requiring electrical connection to the test object. Apparent charge is the unipolar charge which, if injected into the terminals of the test object, results in the same reading on a measurement instrument as that due to the measured PD pulse. The IEC measurement therefore quantifies the time integral of PD current.

Free-space radiometric (FSR) measurement of PD uses a broadband antenna to receive the RF energy radiated by the accelerating charge comprising the PD current transient. In FSR measurements the received RF signal is

typically proportional to a time derivative of PD current [2].

Since both the principles and practical implementations of contact and FSR methods are quite different they may be expected to have different responses to the same PD event. The severity of PD is normally characterized by its intensity measured in picocoulombs (pC) of apparent charge. The nature of FSR measurements makes their use for the measurement of absolute PD intensity difficult, if not impossible. This is because the received signal amplitude depends on several factors which are unknown to greater or lesser extent and in at least one case is practically unknowable. In order of increasing difficulty to establish, these unknown factors include: (i) the path loss between radiating structure and receiving antenna, (ii) the polarization of the radiated field in the direction of the receiving antenna, (iii) the gain of the radiating structure in the direction of the receiving antenna and (iv) the radiated power. One of the main intentions of the work reported in this paper was to explore the possibility of using a contact PD measurement to empirically calibrate an FSR PD measurement.

As reported below, the utility of the contact method to calibrate an FSR measurement appears to be limited. The value of the measurements presented is therefore restricted to an observation of the degree to which the contact and FSR measurements are similar, at least in the context of a particular type of PD event, i.e. that arising from a floating-electrode discharge. The contact measurement reported here does not comply rigorously with the IEC60270 standard but the configuration of the measurement is very similar.

II. REPORTED RELATIONSHIP BETWEEN APPARENT CHARGE AND FSR MEASUREMENTS

Although both the contact and FSR methods can be used to identify insulation defects, FSR measurements have some practical advantages over contact measurements. Both, however, have limitations. For example, FSR techniques can provide information on the location of PD. However, this technique cannot easily quantify PD intensity. Contact measurements give an indication of the apparent charge (and therefore PD

intensity) and may also provide information about the nature of an insulation fault based on PD bandwidth and phase resolved discharge patterns, [3].

Judd et al., [3], proposed a new integrated approach to PD monitoring using the combined and simultaneous application of UHF and IEC60270 measurements. Initial results of the combined approach show that it may be possible to discriminate between different sources of PD using apparent charge and UHF signal energy. Zhang, et al., [4], investigated the correlation between RF measurements and the apparent charge of partial discharge. The results show a linear association between the amplitude of the RF signal and apparent charge in a positive half cycle. Ohtsuka et al., [5], investigated the association between apparent charge and FSR signals using measurements and an FDTD simulation. It is suggested that the PD charge quantity can be corrected by using the FSR method. Reid et al., [6], calculated the energy spectrum of the RF signal and showed that apparent charge together with the frequency distribution of the RF signal can prove useful as an identifier of PD and a means of defect characterization. Reid et al., [7], present phase resolved patterns of RF signals. The results show defect detection is possible if the system is trained by using both RF and contact measurement data. Reid et al., [8], recorded PD using contact and RF methods. Results indicate that the relationship between the methods creates characteristic patterns specific to defect types. Xiao, et al., [9], used RF and IEC time domain analysis measurements to investigate the assessment of insulation integrity for six types of defects. Results were inconclusive in respect of distinguishing differing defects using time domain analysis, but frequency domain analysis revealed information relating to the resonances of the radiating structure which might provide useful defect discrimination. Sarathi et al., [10], analyzed UHF FSR signals caused by the movement of conducting particles subject to a high voltage in transformer oil. (The magnitude of the UHF signal generated by a DC voltage was higher than that generated by an AC voltage.) It was concluded that it was possible to classify an incipient discharge from either an AC or a DC routine substation pressure test. Sarathi et al., [11], made conventional FSR measurements of PD in gas insulated switchgear and concluded that signal bandwidth is independent of applied voltage and operating pressure. Li et al., [12], described measurement results of PD under DC conditions. The results showed that discharge pulses have different shapes depending on the DC voltage and the main part of the discharge energy is situated below 100 MHz.

III. APPARATUS

Figure 1 is a schematic diagram of the measurement apparatus used to simultaneously obtain contact and FSR measurements for the same PD event.

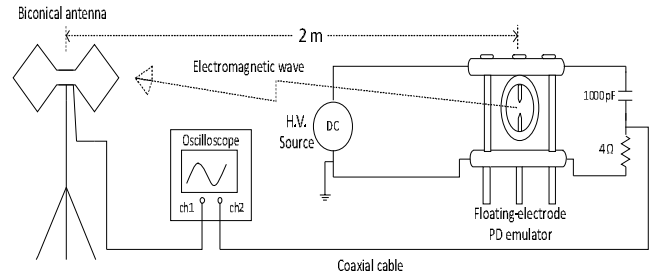


Fig. 1. Experimental apparatus.

PD is generated by applying a large DC voltage to a floating-electrode PD emulator with 1 mm gap between the HV and floating electrode.

The RF signal radiated from the emulator is captured using a wideband biconical antenna connected to one channel of a 4-channel, 4 GHz, digital sampling oscilloscope (DSO). The DSO sampling rate of each channel is 20 GSa/s. The antenna, oriented for horizontal polarization, is located 2 m from the PD source.

Figure 2 shows the floating-electrode PD emulator which is based on that described in [13]. The high voltage output of the power supply is connected to the upper electrode and the lower electrode is connected to earth. A floating metallic needle is located between the electrodes but is not (electrically) connected to either. The long dimension of the needle is parallel to the applied electric field. When the electric field is sufficiently large a corona discharge from the floating electrode is initiated, [14].

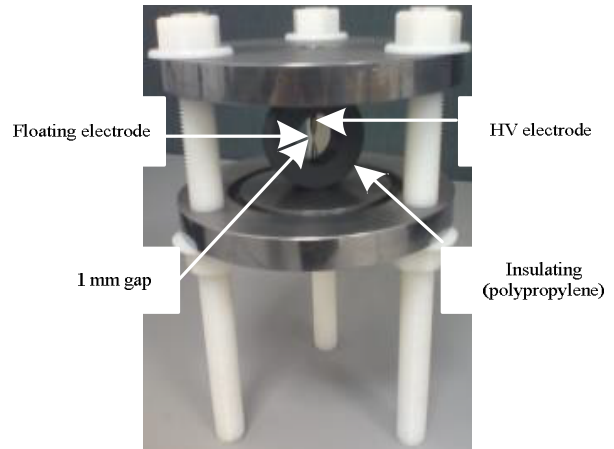


Fig. 2. Floating-electrode PD emulator.

IV. EXAMPLE EVENT

Figures 3 and 4 show a typical PD event captured by the FSR and contact measurements respectively. The voltage at which this PD event occurred was 6.2 kV. Figure 5 compares the two normalized measurements. This event is representative of many such measurements.

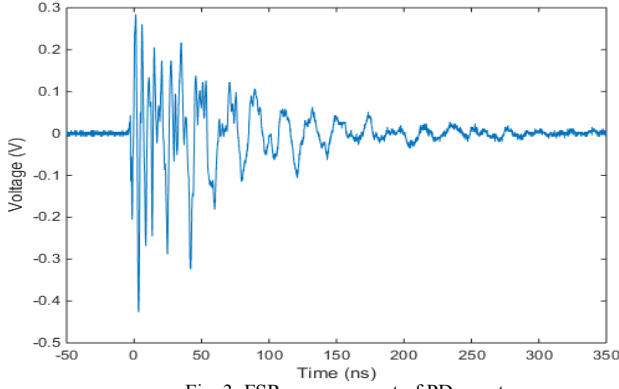


Fig. 3. FSR measurement of PD event.

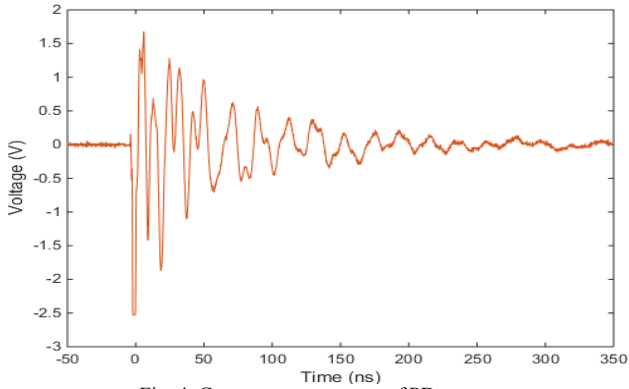


Fig. 4. Contact measurement of PD event.

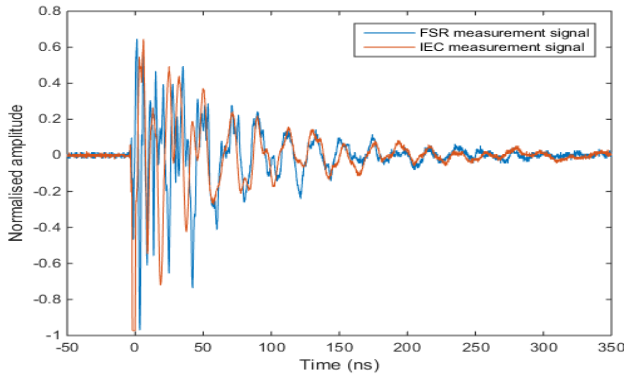


Fig. 5. Comparison of normalised FSR and contact measurements.

The similarity in the temporal decay of the two signals is more similar than was expected. Band limiting of the measurement due to the electromagnetic radiation and reception processes was expected in the case of the FSR measurement. The expectation was for a somewhat less severe band limiting in the case of the contact measurement resulting in shorter ringing. The inference drawn is that band limiting is dominated by the inductive and capacitive characteristics of the PD source rather than the frequency response of the FSR receiving antenna.

The total duration of the pulse is approximately 300 ns and the $1/e$ duration of the pulse envelope is approximately 70 ns. The normalised frequency spectra obtained by FFT analysis using 10^4 time samples are shown in Figures 6 and 7. Figures 8 and 9 show the frequency sub-bands and the fraction of energy that they contain. Most of the energy lies between approximately 10

MHz and 300 MHz, although the exact distribution of energy is not the same in the two sets of figures.

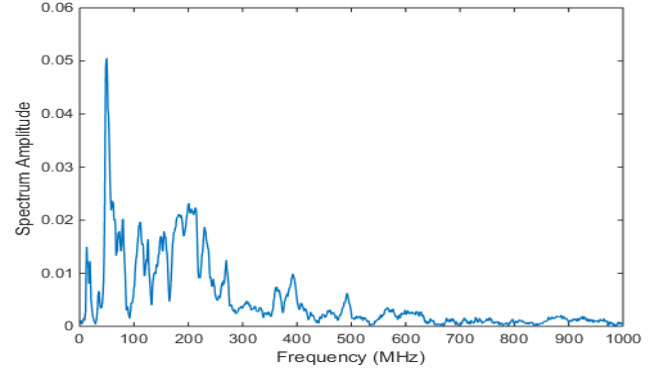


Fig. 6. Frequency spectrum of FSR measurement.

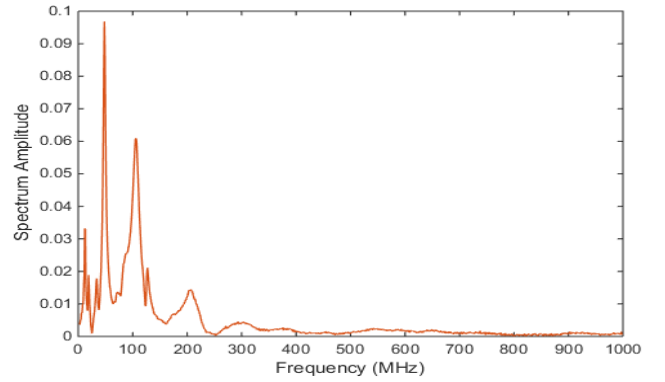


Fig. 7. Frequency spectrum of contact measurement.

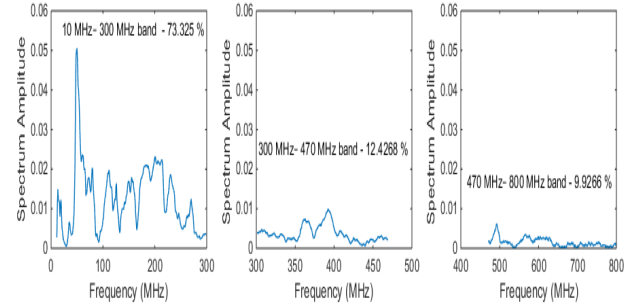


Fig. 8. Sub-band spectra and energy proportions for FSR measurement.

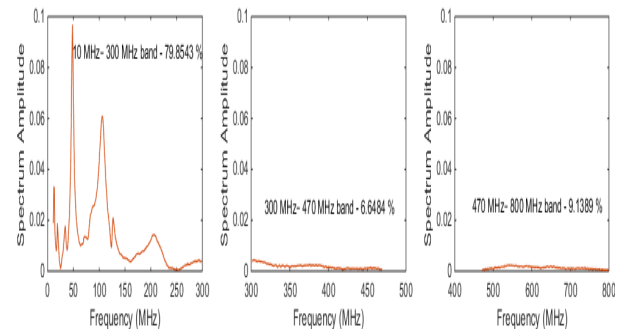


Fig. 9. Sub-band spectra and energy proportions for contact measurement.

The time-cumulative energies of the PD pulse measurements are shown in Figures 10 and 11. The total energy of the FSR pulse is 2.8×10^{-7} J and the total energy of the contact pulse is 1.2×10^{-5} J.

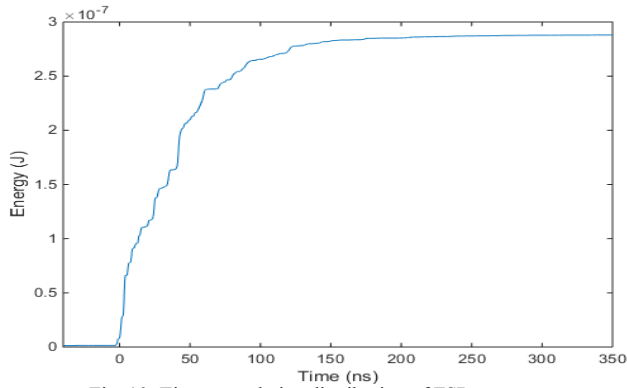


Fig. 10. Time-cumulative distribution of FSR energy.

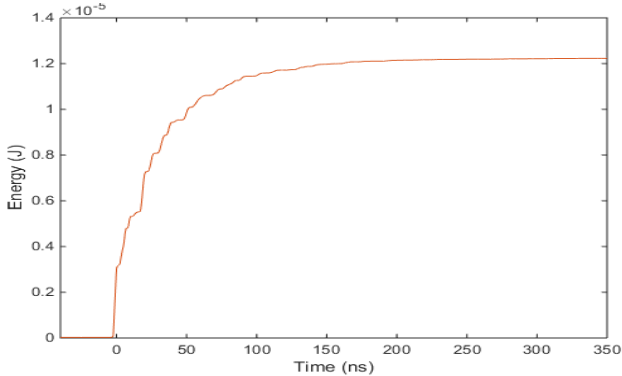


Fig. 11. Time-cumulative distribution of contact energy.

V. CONCLUSIONS

A comparison of PD signals captured using FSR and contact methods has been carried out. Preliminary results show greater similarity than was expected in the time domain signals. It appears that the signal characteristics in both cases are dominated by the lossy equivalent (resonant) circuit of the radiating structure. Normalised frequency spectra obtained by FFT analysis from the time domain pulses indicate that the main frequency content of the PD discharge is situated in the range of 10–300 MHz, i.e. 73% of the total energy for the FSR measurement and 80% of the total energy for the contact measurement.

REFERENCES

- [1] E. Iliana, J. Philip, and A. Ian, "RF-Based Partial Discharge Early Warning System for Air-Insulated Substation," *IEEE Transactions on Power Delivery*, vol. 24, pp. 20-29, 2009.
- [2] A. Reid, M. Judd, B. Stewart, D. Hepburn, and R. Fouracre, "Identification of multiple defects in solid insulation using combined RF and IEC60270 measurement," in *Solid Dielectrics, ICSD'07*. IEEE International Conference on, 2007, pp. 585-588.
- [3] M. D. Judd, A. J. Reid, L. Yang, B. G. Stewart, and R. A. Fouracre, "A new integrated diagnostic partial discharge monitoring strategy for HV plant items: combining UHF couplers and the IEC60270 standard," in *Electricity Distribution, 2005. CIRED 2005*. 18th International Conference and Exhibition on, pp. 1-4.
- [4] X. Zhang, J. Tang, and Y. Xie, "Investigation of the relationship between PD measured with RF techniques and apparent charge quantity of metal protrusion in air," in *High Voltage Engineering and Application (ICHVE)*, 2010 International Conference on, pp. 286-289.
- [5] S. Ohtsuka, T. Teshima, S. Matsumoto, and M. Hikita, "Relationship between PD-induced electromagnetic wave measured with UHF method and charge quantity obtained by PD current waveform in model GIS," in *Electrical Insulation and Dielectric Phenomena, 2006 IEEE Conference on*, pp. 615-618.
- [6] A. Reid, M. Judd, B. Stewart, and R. Fouracre, "Frequency distribution of RF energy from PD sources and its application in combined RF and IEC60270 measurements," in *Electrical Insulation and Dielectric Phenomena, 2006 IEEE Conference on*, pp. 640-643.
- [7] A. Reid, M. Judd, B. Stewart, and R. Fouracre, "Comparing IEC60270 and RF partial discharge patterns," in *Condition Monitoring and Diagnosis, 2008. CMD 2008*. International Conference on, pp. 89-92.
- [8] A. J. Reid, M. D. Judd, R. A. Fouracre, B. Stewart, and D. Hepburn, "Simultaneous measurement of partial discharges using IEC60270 and radio-frequency techniques," *Dielectrics and Electrical Insulation, IEEE Transactions on*, vol. 18, pp. 444-455, 2011.
- [9] S. Xiao, P. Moore, and M. Judd, "Investigating the assessment of insulation integrity using radiometric partial discharge measurement," in *Sustainable Power Generation and Supply, 2009. SUPERGEN'09*. International Conference on, pp. 1-7.
- [10] R. Sarathi, A. Reid, and M. D. Judd, "Partial discharge study in transformer oil due to particle movement under DC voltage using the UHF technique," *Electric Power Systems Research*, vol. 78, pp. 1819-1825, 2008.
- [11] R. Sarathi and R. Umamaheswari, "Understanding the partial discharge activity generated due to particle movement in a composite insulation under AC voltages," *International Journal of Electrical Power & Energy Systems*, vol. 48, pp. 1-9, 2013.
- [12] X. Li, G. Wu, X. Zhang, and S. Bian, "Partial discharge pulse shape detection and analysis under DC condition," in *Electrical Insulation Conference and Electrical Manufacturing Expo, 2007*, pp. 48-51.
- [13] B. Hampton, "UHF diagnostics for gas insulated substations," in *High Voltage Engineering, 1999. Eleventh International Symposium on (Conf. Publ. No. 467)*, 1999, pp. 6-16.
- [14] W. Liming, L. Hong, G. Zhicheng, L. Xidong, and N. Allen, "Sparkover behaviour of air gap with floating electrode combined with impulse and alternating voltages," in *Properties and Applications of Dielectric Materials. Proceedings of the 6th International Conference on*, 2000, pp. 641-644.

Diagnosis of Combination Faults in a Planetary Gearbox using a Modulation Signal Bispectrum based Sideband Estimator

Xiang Tian, Gaballa M. Abdallaa, Ibrahim Rehab,
Fengshou Gu and Andrew D. Ball
Centre for Efficiency and Performance Engineering
University of Huddersfield
Huddersfield, UK

Fengshou Gu, Tie Wang

School of Mechanical Engineering
Taiyuan University of Technology
Shanxi, China

Abstract—This paper presents a novel method for diagnosing combination faults in planetary gearboxes. Vibration signals measured on the gearbox housing exhibit complicated characteristics because of multiple modulations of concurrent excitation sources, signal paths and noise. To separate these modulations accurately, a modulation signal bispectrum based sideband estimator (MSB-SE) developed recently is used to achieve a sparse representation for the complicated signal contents, which allows effective enhancement of various sidebands for accurate diagnostic information. Applying the proposed method to diagnose an industrial planetary gearbox which coexists both bearing faults and gear faults shows that the different severities of the faults can be separated reliably under different load conditions, confirming the superior performance of this MSB-SE based diagnosis scheme.

Keywords—Planetary gearbox, Ball bearing, Modulation signal bispectrum, Combination fault diagnosis, Vibration signal.

I. INTRODUCTION

Planetary or epicyclic gearboxes are widely used for the power transmission of important machines such as helicopters, wind turbines, automobiles, aircraft engines and marine vehicles due to their large transmission ratios and strong load-bearing capacity. Gears and bearings are the critical mechanical components in planetary gearboxes. Early fault detection and diagnosis are significant to prevent any failures of either of these components which can lead to the failure of the entire system. Therefore, many advanced techniques have been investigated to analyze the vibration signals from planetary gearboxes for more accurate diagnosis.

Sawalhi et al. [1] proposed a method based on time synchronous averaging (TSA). Firstly, it isolates and then removes the deterministic components corresponding to each gear in the system by synchronous averaging, leaving a residual stochastic signal which should be dominated by bearing faults in some frequency bands. Then, the residual signal is applied to cepstrum pre-whitening for bearing fault detection. Vishwash et al. [2] used multi-scale slope feature extraction technique based on wavelet multi-resolution analysis, discrete wavelet transform (DWT) and wavelet packet transform (WPT), for fault diagnosis of gear and bearing. For planetary bearing fault diagnosis, Bonnardot and Randall et al. [3] presented an enhanced

unsupervised noise cancellation that uses an unsupervised order tracking algorithm to perform noise cancellation in the angular domain. To extract fault features of the rolling element bearing from the masking faulty gearbox signals, Tian et al. [4] explored a method based on WPT, Pearson correlation coefficient and envelope analysis. Elasha et al. [5] developed a method for defective bearings in a planetary gearbox by applying an adaptive filter, spectral kurtosis and envelope analysis to both AE and vibration signals. These efforts in improving data quality have shown different degrees of success in diagnosing fault types and severities.

However, these significant progresses in analyzing the vibration signals are made based on single type of fault cases largely and less attention is paid to multiple faults occurring concurrently which are becoming more significant as the structures of rotating machinery become of larger scale, of higher speed, and more complicated[6]. In addition, these studies usually focused more on noise reduction but with limited efforts on the utilization of multiple modulation characteristics in extracting the diagnostic information.

To fill these gaps, this paper presents a new method for combination fault detection of gear and bearing based on MSB-SE analysis of vibration signals, which has been demonstrated to be particular effective in highlighting sidebands and hence diagnosing faults on gears only [7]. The following content is organized as: Section 2 outlines the theoretical basis of the combination fault diagnosis based on the modulations between different vibration sources. Section 3 describes the experimental setups for validating the proposed method. Then, section 4 presents the diagnostic results and discussion. Finally, section 5 is the conclusion.

II. THEORETICAL BACKGROUND FOR DIAGNOSING THE COMBINATION FAULTS

A. Planetary gearbox vibration characteristics

A planetary gearbox is composed of a ring gear, a sun gear and multiple planet gears. Usually, the ring gear is stationary, a sun gear rotates around a fixed center, and planet gears not only spin around their own centers but also revolve around the center of the sun gear. The planet gears mesh simultaneously with both the sun gear and the ring gear. Due to these complicated gear motions, the

vibration signals generated by planetary gearboxes are more complicated than those by fixed shaft gearboxes. In addition, the planet phasing relationship, which is dependent on the number of planets, planet position angles, and the number of teeth of each gear, also adds complexity to vibration signals. In this section, the planetary gearbox vibration signal models will be introduced. The gear damage could produce the amplitude modulation and frequency modulation (AMFM) effects to the gear meshing vibration at corresponding fault characteristic frequencies [8].

Based on the theoretical analysis in [7], in steady working condition such as constant running load and speed, the vibration perceived by a sensor on the stationary ring can be represented with mutual modulations of both AM and FM phenomena. For a local fault, such as the crack and pitting on one of the tooth of sun gear, the signal model for the 1st sinusoidal component can be expressed as:

$$f(t) = [1 - \cos(2\pi f_{rs}t)][1 - \cos(2\pi f_{rc}t)][1 + A \cos(2\pi f_{sf}t + \phi)] \times \cos[2\pi f_m t + B \sin(2\pi f_{sf}t + \phi) + \theta] \quad (1)$$

on the planet gear

$$f(t) = [1 - \cos(2\pi f_{rc}t)][1 - \cos(2\pi f_{rc}t)][1 + A \cos(2\pi f_{pf}t + \phi)] \times \cos[2\pi f_m t + B \sin(2\pi f_{pf}t + \phi) + \theta] \quad (2)$$

and on the ring gear

$$f(t) = [1 + A \cos(2\pi f_{rf}t + \phi)][1 - \cos(2\pi f_{rc}t)] \times \cos[2\pi f_m t + B \sin(2\pi f_{rf}t + \phi) + \theta] \quad (3)$$

where f_{sf} , f_{pf} and f_{rf} is the fault characteristic frequency of the sun gear, planet gear and or ring gear respectively. f_{rc} and f_{rs} is the rotating frequency of the carrier and sun gear. f_m is the gear meshing frequency. θ , ϕ and ϕ are the initial phases of AM and FM respectively.

Therefore, consider the AMFM effects with the high orders of fault gear characteristic frequency nf_{sf} as the modulating frequency and with the higher orders of meshing frequency kf_m as the signal carrier frequency and f_{rx} as the corresponding component rotating frequency, the vibration spectral peaks will appear at the frequency locations of $kf_m \pm nf_{sf} \pm f_{rc}$ and $kf_m \pm f_{rx} \pm nf_{sf} \pm f_{rc}$ ($k, n = 1, 2, 3, \dots$) in the Fourier spectrum. From the analysis of vibration spectra, we can detect and locate the gear fault by monitoring the presence of magnitude increase of spectral peaks at the above mentioned frequency locations.

B. Characteristic Frequencies for Planetary Gear Fault Detection

According to reference [7], the rotation frequency of carrier can be calculated as

$$f_{rc} = \frac{Z_s}{Z_r + Z_s} f_{rs} \quad (4)$$

the planet gear frequency as

$$f_{rp} = \frac{(Z_p - Z_r)Z_s}{(Z_r + Z_s)Z_p} f_{rs} \quad (5)$$

and the meshing frequency as

$$f_m = (f_{rs} - f_{rc})Z_s = \frac{Z_r Z_s}{Z_r + Z_s} f_{rs} = Z_r f_{rc}, \quad (6)$$

where f_{rs} is the sun gear rotating speed; Z_r , Z_p and Z_s denote the number of teeth for the ring, planet and sun gear respectively.

As shown in many previous studies, detection and diagnosis can be carried out by examining the changes of characteristic frequencies around mesh frequency f_m and its harmonics. Considering that there are K number of planetary gears moving with the carrier, characteristic frequencies around meshing frequency can be calculated [9][10] for different local faults occurring on the sun gear

$$f_{sf} = \frac{f_m}{Z_s} = K(f_{rs} - f_{rc}) \quad (7)$$

on the planet gear

$$f_{pf} = 2 \frac{f_m}{Z_p} = 2(f_{rp} + f_{rc}) \quad (8)$$

and on the ring gear

$$f_{rf} = \frac{f_m}{Z_r} = K f_{rc}. \quad (9)$$

However, as shown in [10][11] only some of these expected sidebands will be apparent in the vibration spectrum when a planetary gearbox has faults due to the effects of constructive superposition of the vibration waves from the three gear sets, whereas other sidebands are hard to be seen because of the destructive effect of the superposition, and hence the latter have been largely neglected by previous studies when developing methods for fault diagnosis.

C. Characteristic Frequencies for Bearing Fault Detection

A ball bearing consists of an inner race, an outer race, several balls and a cage, which holds the balls in a given relative position. Race surface fatigue results in the appearance of spalls on the inner race, outer race or balls. If one of the races has a spall, it will almost periodically impact with the balls. The fault signature is represented by successive impulses with a repetition rate depending on the faulty component, geometric dimensions and the rotational speed. The period between impacts is different for all the listed elements and depends on the geometry of the bearing, the rotational speed and the load angle. For a fixed outer race bearing, the theoretical characteristic fault frequencies can be calculated using (10)-(13), and a derivation of these equations is presented in [12].

Fundamental cage frequency:

$$F_c = \frac{1}{2} F_s \left(1 - \frac{D_b}{D_c} \cos \alpha\right) \quad (10)$$

Outer race defect frequency:

$$F_o = \frac{N_b}{2} F_s (1 - \frac{D_b}{D_c} \cos \alpha) \quad (11)$$

Inner race defect frequency:

$$F_i = \frac{N_b}{2} F_s (1 + \frac{D_b}{D_c} \cos \alpha) \quad (12)$$

Ball defect frequency:

$$F_b = \frac{D_c F_s}{2 D_b} (1 - \frac{D_b^2}{D_c^2} \cos^2 \alpha) \quad (13)$$

where D_c is pitch circle diameter, D_b is ball diameter, α is contact angle, N_b is number of ball and F_s is shaft rotational frequency.

While the sensor is mounted on the gearbox housing, which is connected to or fastened to the ring gear directly, the bearing damage induced vibration has two main paths to go from its source to the sensor through solid mechanical components and their contacts. Through the first path, the vibration signal propagates from its origin to the gearbox casing, and then reaches the sensor. Whereas through the second path, the vibration signal follows a longer path, from its origin to the shaft firstly, then from the shaft go through the sun gear, planet gear and ring gear, after that from the ring gear to the gearbox casing, and finally to the sensor. Therefore, the vibration signal will be amplitude modulated by both sun gear rotating frequency and carrier rotating frequency as shown in (14).

$$f(t) = [1 - \cos(2\pi f_{rs} t)] \cos[2\pi f_{bx} t + \alpha] + [1 - \cos(2\pi f_{rc} t)] \cos[2\pi f_{bx} t + \alpha] \quad (14)$$

where f_{bx} is the characteristic frequency of bearing.

In practice there is always slight slippage, especially when a bearing is under dynamic loads and with severe wear. Therefore, these frequencies may have a slight difference from calculated ones above. In this paper, the experiment bearing is located on the output shaft of planetary gearbox which is closed to sun gear. It is a 6008ZZ deep groove ball bearing and its geometric dimensions that needed for fault frequency calculation are listed in Table I.

TABLE I. SPECIFICATION OF EXPERIMENT BALL BEARING

Parameter	Measurement
Pitch Diameter	54 mm
Ball Diameter	7.398 mm
Ball Number	12
Contact Angle	0

D. Modulation Signal Bispectrum

For a vibration signal $x(t)$ with corresponding Fourier Transform (FT) $X(f)$, the MSB-SE can be obtained by

$$B_{MS}^{SE}(f_c, f_s) = E \left[\frac{X(f_c + f_s) X(f_c - f_s) X^*(f_c) X^*(f_s)}{|X(f_c)|^2} \right] \quad (15)$$

where the product between the upper sideband $X(f_c + f_s)$, the lower sideband $X(f_c - f_s)$ and the normalized carrier component $X^*(f_c) X^*(f_s) / |X(f_c)|^2$ allows the sideband

effect to be combined and quantified without the effect of the carrier amplitude. Moreover, because of the average operation, denoted by the expectation operator $E[\cdot]$ in (15), the sideband products which associate with a constant phase value can be enhanced, while the noise and interfering components with random phases are suppressed effectively. This MSB based approach has been shown to yield outstanding performance in characterizing the small modulating components of motor current signals for diagnosing different electrical and mechanical faults under different load conditions [13][14][15]. Therefore, it is also evaluated in this study to extract the residual sidebands of vibration signal for the purpose of gear and bearing fault diagnosis.

III. EXPERIMENTAL SETUPS

To verify the effectiveness of MSB-SE based diagnosis, vibration signals were acquired from an in-house planetary gearbox test system. The maximum torque of planetary gearbox is 670 Nm, the maximum input speed is 2800 rpm and maximum output speed is 388 rpm. The schematic in Fig. 1 shows the position of the accelerometer that mounted on the outer surface of the ring gear and the position of experiment studied bearing.

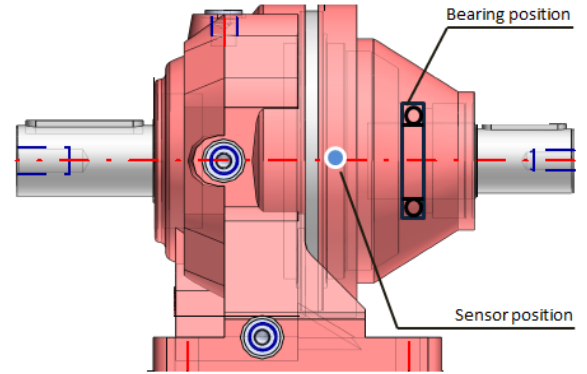


Figure 1. Schematic for a planetary gearbox

In the experiment, the planetary gearbox operates at 80% of its full speed under 5 load conditions (0%, 25%, 50%, 75% and 90% of the full load). The load setting allows fault diagnoses to be examined with variable load operations which are the cases for many applications such as wind turbine, helicopters etc. The vibration is measured by a general purpose accelerometer with a sensitivity of 31.9 mv/(ms⁻²) and frequency response ranges from 1Hz to 10kHz. All the data were logged simultaneously by a multiple-channel, high-speed data acquisition system with 100 kHz sampling rate and 16-bit resolution.



Figure 2. Tooth defects simulated on the sun gear and two kinds of inner race defect on deep groove ball bearing

Three cases of test were carried out to examine the combination faults. The first one is health case, in which

there is no defect on either the gear or bearing. The second one is for the combination fault of small bearing inner race defect and sun gear tooth defect. The third one is for the combination fault of large bearing inner race defect and sun gear tooth defect. For the convenience of discussion, these three cases are denoted as Healthy, CbFault1 and CbFault2, respectively. Fig. 2 shows the defects on sun gear and bearing inner races.

IV. DIAGNOSTIC RESULTS AND DISCUSSION

A. Spectrum Features of Vibration Signals

Fig. 3 shows the typical spectra for the three cases under the same load. They exhibit complicated patterns and high density of spectral component, which needs careful examination to find the components of interest. Three distinctive peaks close to the first three mesh

frequencies appear at $f_m + f_{rc}$, $2f_m - f_{rc}$ and $3f_m$ respectively, which agrees with the model prediction and that of previous studies[16][10][11][16]. However, there are also many distinctive peaks between two mesh frequencies. For example, the components at $2f_m - 6f_{sf} - 1f_{rc}$, $2f_m + 7f_{sf}$ etc. should not appear for a healthy planetary gearbox. The presence of these peaks may due to the gearbox manufacturing and installation errors. The green dash lines show the bearing inner race fault frequency and its harmonics. It is obvious that their amplitudes are quite small compared with the other components. This makes it difficult for reliable bearing fault diagnosis. Therefore, the modulation effects between the bearing fault frequency and other characteristic frequencies such as f_{rs} and f_{rc} are used for bearing fault diagnosis.

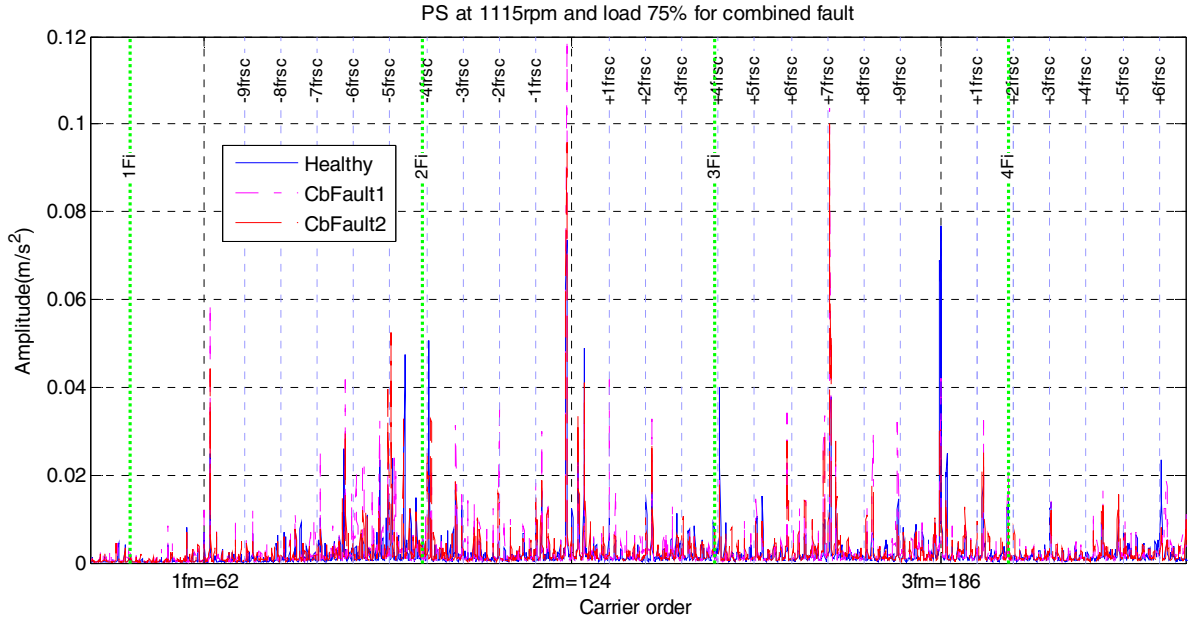


Figure 3. Spectra for different fault cases of the gearbox at 1115 rpm and 75% load.

B. MSB Features of Vibration Signals

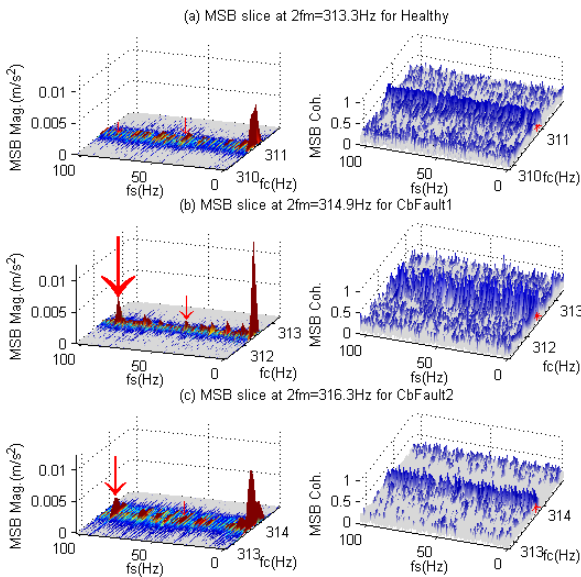


Figure 4. MSB results for different cases of the tests under 75% load.

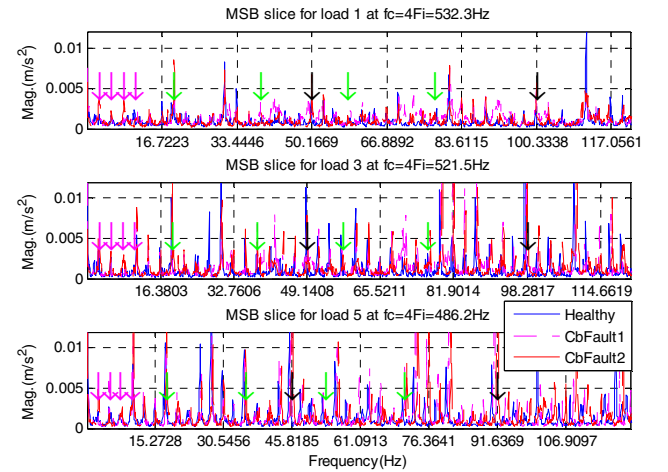


Figure 5. MSB slice for different cases at $f_c = 4F_i$.

Fig. 4 shows a typical MSB result for the three cases of test under 75% load. To show a clear change of the residual sidebands around mesh frequency $2f_m = 313\text{Hz}$, MSB and its corresponding coherence results are

presented in the bifrequency domain in the region of $f_c \leq 2f_m \pm 1 = 313 \pm 1\text{Hz}$ and $f_s < 100\text{Hz}$ to include the sidebands up to $6f_{sf}$.

Fig. 5 illustrates the MSB slices at $f_c = 4F_i$ for bearing fault detection. The pink, green and black arrows show the sideband at f_{rc} , f_{rs} and f_{sf} respectively. They show differences between baseline, small bearing inner race defect and large bearing inner race defect cases.

C. Diagnosis of Sun Gear Fault

The diagnostic results for the sun gear are presented in Fig. 6. From the results of residual sidebands obtained from the MSB slice at $2f_m - f_{rc}$, it can be seen that the amplitudes at f_{sf} show a good increasing trend with loads, which agrees with the load characteristics of gear transmissions. Moreover, these amplitudes show clear incremental differences between three tested cases under high load. Therefore, they can be used for obtaining fault diagnosis reliably.

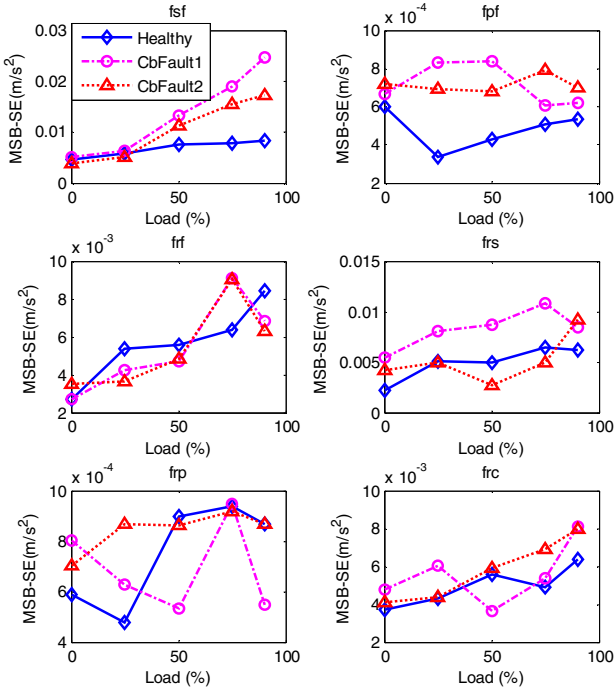


Figure 6. MSB-SE diagnosis results of the sun gear faults using the slice at $f_c = 2f_m - f_{rc}$.

The corresponding MSB coherence results are printed in Fig. 7, which can be used to assure the reliability of MSB-SE results. From the figure it can be seen that the MSB coherence is low at f_{pf} and f_{rp} , which indicates that there is no significant modulation phenomena at these two sidebands. It means there is no fault on planetary gear and ring gear. Meanwhile, the amplitude changes for other characteristic frequencies are also provided to assure the diagnostic results. These changes exhibit high fluctuations with the fault progression and the load increases, which is not consistent with the gear dynamic characteristics in that the fault usually causes higher vibrations and also increases with load. Therefore, they cannot be used to

indicate the corresponding faults but just caused by refitting errors.

Therefore, the fault location can be identified by checking the feature that the increase in residual sidebands occurs over several different loads simultaneously.

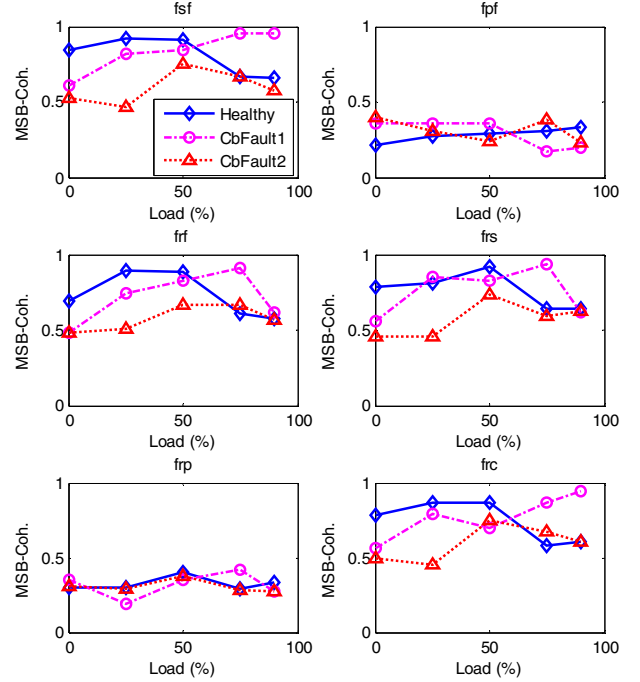


Figure 7. MSB coherence results of the sun gear faults using the slice at $f_c = 2f_m - f_{rc}$.

D. Diagnosis of Bearing Fault

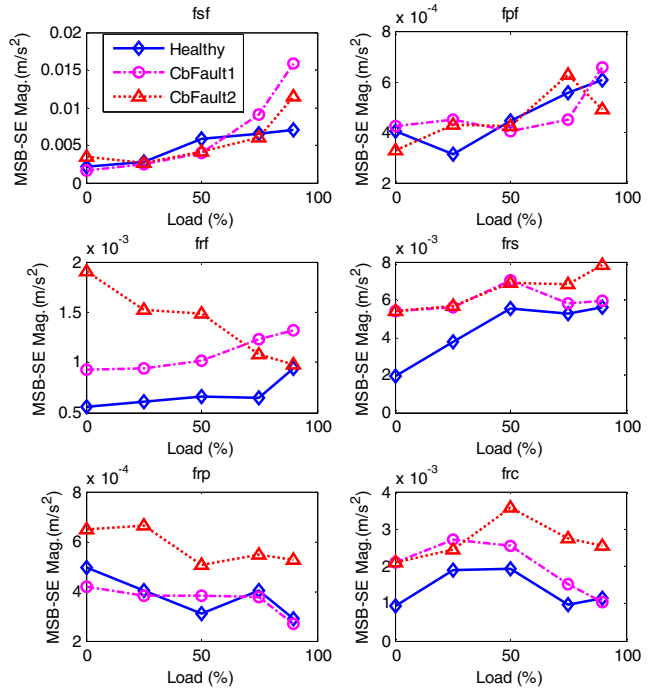


Figure 8. Averaged MSB-SE diagnosis results of the small bearing faults using the slices at $f_c = F_i$ and $f_c = 4F_i$.

From the spectrum in Fig. 3 it can be seen that the characteristic frequency of bearing fault and its harmonics may be interfered by the complex gearbox frequency components. Some harmonic amplitudes may be greatly reduced which is not conducive to bearing fault diagnosis. Therefore, only the harmonics with high coherence values will be selected for bearing fault diagnosis.

In this paper, MSB slices at $f_c = F_i$ and $f_c = 4F_i$ are selected for fault detection because of their high coherences. The averaged MSB-SE and MSB coherence results are presented in Fig. 8 and Fig. 9, respectively. Compared with other feature components, the features at f_{rs} and f_{rc} have higher coherence amplitudes, showing high potential of modulation effects between the fault characteristic frequencies and their closer interacting components. This thus confirms the presence of the inner race fault on the bearing. However, they can only separate small inner race fault from the larger under the high load conditions where the modulations are stronger.

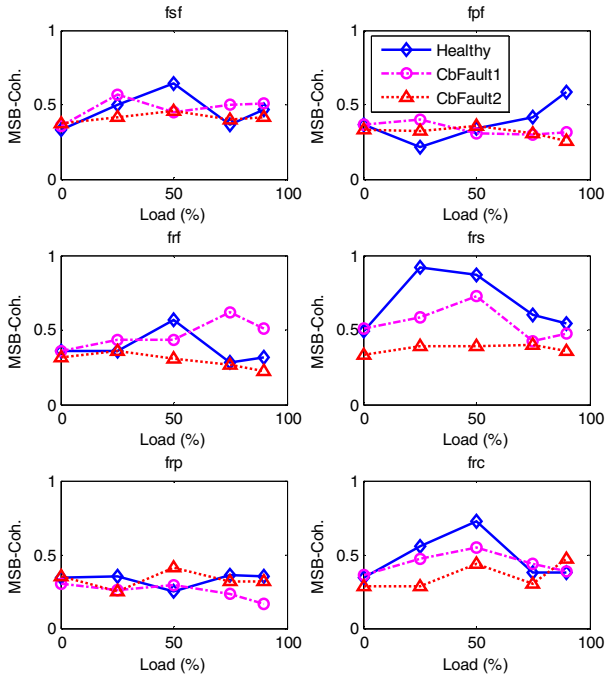


Figure 9. Averaged MSB coherence results of the small bearing faults using the slice at $f_c = F_i$ and $f_c = 4F_i$.

V. CONCLUSIONS

In this paper, a combination fault diagnosis method based on MSB-SE is developed for the monitoring faults on both bearing and gear in a planetary gearbox. MSB analysis is effective in suppressing random noise and decomposing the nonlinear modulation components in the measure vibration signals. Thus, the sideband amplitudes extracted by MSB-SE at the MSB slices relating to characteristic frequencies have more reliable information on the fault which causes the sidebands.

The method was verified with experimental data from a planetary gearbox with combined gear and bearing faults. The diagnostic results show that not only the types of the

combination faults: defects in bearing inner race and tooth breakages of sun gear can be separated but also the severity of the two faults can be estimated successfully under high load conditions.

REFERENCES

- [1] N. Sawalhi, R. Randall and D. Forrester, 'Separation and enhancement of gear and bearing signals for the diagnosis of wind turbine transmission systems', *Wind Energy*, vol. 17, no. 5, pp. 729-743, 2013.
- [2] B. Vishwash, P. S. Pai, N. S. Sriram, R. Ahmed, H. S. Kumar, and G. S. Vijay, "Multiscale Slope Feature Extraction for Gear and Bearing Fault Diagnosis Using Wavelet Transform," *Procedia Materials Science*, vol. 5, pp. 1650-1659, 2014.
- [3] Bonnardot, F., R. B. Randall, J. Antoni, and F. Guillet. "Enhanced unsupervised noise cancellation using angular resampling for planetary bearing fault diagnosis." *International journal of acoustics and vibration*, vol. 9, no. 2, pp. 51-60, 2004.
- [4] J. Tian, M. Pecht and C. Li, 'Diagnosis of rolling element bearing fault in bearing-gearbox union system using wavelet packet correlation analysis', Dayton, OH, 2012, pp. 24-26.
- [5] E. Faris, M. Greaves and D. Mba, 'Diagnostics of a defective bearing within a planetary gearbox with vibration and acoustic emission', in *The 4th International Conference on Condition Monitoring of Machinery in Non-Stationary Operations (CMMNO 2014)*, Lyon, France, 2014.
- [6] Q. Zhang, Q. Hu, G. Sun, X. Si and A. Qin, 'Concurrent Fault Diagnosis for Rotating Machinery Based on Vibration Sensors', *International Journal of Distributed Sensor Networks*, vol. 2013, pp. 1-10, 2013.
- [7] F. Gu, G. Abdalla, R. Zhang, H. Xu and A. D. Ball, 'A Novel Method for the Fault Diagnosis of a Planetary Gearbox based on Residual Sidebands from Modulation Signal Bispectrum Analysis', in *Comadem 2014*, Brisbane, Australia, 2014.
- [8] Z. Feng and M. Zuo, 'Vibration signal models for fault diagnosis of planetary gearboxes', *Journal of Sound and Vibration*, vol. 331, no. 22, pp. 4919-4939, 2012.
- [9] Y. Lei, J. Lin, M. Zuo and Z. He, 'Condition monitoring and fault diagnosis of planetary gearboxes: A review', *Measurement*, vol. 48, pp. 292-305, 2014.
- [10] L. Hong, J. Dhupia and S. Sheng, 'An explanation of frequency features enabling detection of faults in equally spaced planetary gearbox', *Mechanism and Machine Theory*, vol. 73, pp. 169-183, 2014.
- [11] M. Inalpolat and A. Kahraman, 'A theoretical and experimental investigation of modulation sidebands of planetary gear sets', *Journal of Sound and Vibration*, vol. 323, no. 3-5, pp. 677-696, 2009.
- [12] T. Barszcz and N. Sawalhi, 'Fault Detection Enhancement in Rolling Element Bearings Using the Minimum Entropy Deconvolution', *Archives of Acoustics*, vol. 37, no. 2, pp.131-141, 2012.
- [13] A. Alwodai, F. Gu, A. D. Ball, 'A comparison of different techniques for induction motor rotor fault diagnosis'. *Journal of Physics: Conference Series* 364, 1742-6596, 2012.
- [14] Z. Chen, T. Wang, F. Gu, H. Mansaf, A. D. Ball. 'Gear Transmission Fault Diagnosis Based on the Bispectrum Analysis of Induction Motor Current Signatures', *Journal of Mechanical Engineering*, vol. 48, no. 21, pp. 84-90, 2012.
- [15] F. Gu, T. Wang, A. Alwodai, X. Tian, Y. Shao and A. D. Ball, 'A new method of accurate broken rotor bar diagnosis based on modulation signal bispectrum analysis of motor current signals', *Mechanical Systems and Signal Processing*, vol. 50-51, pp. 400-413, 2015.
- [16] P. McFadden and J. Smith, 'An Explanation for the Asymmetry of the Modulation Sidebands about the Tooth Meshing Frequency in Epicyclic Gear Vibration', *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 199, no. 1, pp. 65-70, 1985.

A Study of Diagnostic Signatures of a Deep Groove Ball Bearing Based on a Nonlinear Dynamic Model

Ibrahim Rehab, Xiang Tian, Fengshou Gu and Andrew D. Ball

Centre for Efficiency and Performance Engineering,
University of Huddersfield, Huddersfield, UK

Abstract— For accurate fault detection and diagnosis, this paper focuses on the study of bearing vibration responses under increasing radial clearances due to investable wear and different bearing grades. A nonlinear dynamic model incorporating with local defects and clearance increments is developed for a deep groove ball bearing. The model treats the inner race-shaft and outer race-housing as two lumped masses which are coupled by a nonlinear spring formalized by the Hertzian contact deformation between the balls and races. The solution of the nonlinear equation is obtained by a Runge-Kutta method in Matlab. The results show that the vibrations at fault characteristic frequencies exhibit significant changes with increasing clearances. However, an increased vibration is found for the outer race fault whereas a decreased vibration is found for inner race fault. Therefore, it is necessary to take into account these changes in determining the size of faults.

Keywords- radial clearance; contact deformation; condition monitoring; bearing defects

I. INTRODUCTION

A large number of papers have been focused on developing signal processing techniques to detect and isolate faults of bearings with high degree of accuracy. But relatively few studies have presented a mathematical (physics-based) model, where faults can be simulated under different operating conditions rather than waiting for their natural occurrence, or alternatively having them seeded for laboratory testing.

Different mathematical models have been developed to study the dynamic effects on the roller bearing. In 1984, McFadden and Smith developed a model which described the vibration produced by a single point defect on the inner race of a rolling element bearing under constant radial load [1]. Purohit et al. (2006) [2] studied the radial and axial vibrations of a rigid shaft supported ball bearing. In the analytical formulation the contacts between the balls and the races are considered as nonlinear springs, whose stiffness is obtained by using the Hertzian elastic contact deformation theory. Culita et al. (2007) [3] presented the McFadden- Smith vibration model, one of the first valid models of vibration generated by single point defects in bearings. They presented how the defect is encoded by vibration in a more accurate and natural manner than previous models. A significant important contribution was brought by S. Sassi et al. (2007) [4]. They developed a numerical model with the assumption that the dynamic behavior of the bearing can be

represented by a coupled three-degree-of-freedom system. Upadhyay et al. (2009) [5] studied the dynamic behavior of a high speed unbalanced rotor supported on roller bearings with damping. The non-linearity in the rotor bearing system has been considered mainly due to Hertzian contact, unbalanced rotor effect and radial internal clearance. Patil et al. (2010) [6] presented an analytical model for predicting the effect of a localized defect on the ball bearing vibrations. In the analytical formulation, the contacts between the ball and the races are considered as non-linear springs. Dougdag et al. (2012) [7] presented an experimental verification of a simplified model of a nonlinear stiffness ball bearing in both static and dynamic modes and tested its capabilities to simulate accurately fault effects. Results of defects simulation and model behavior in statics and dynamics are compared to experimental results. Patel et al. (2013) [8] reported a theoretical and experimental vibration study of dynamically loaded deep groove ball bearings having local circular shape defects. The shaft, housing, raceways and ball masses are incorporated in the proposed mathematical model. Coupled solutions of governing equations of motion had been achieved using Runge-Kutta method.

The objective of this paper is to develop a relatively more realistic dynamic model of deep groove ball bearing in presence of different internal radial clearances and localized defects.

II. DYNAMIC MODEL WITH BEARING DEFECTS

A single ball bearing consists of a number of parts. The description of each component can lead to a simulation model with a large number of degrees of freedom (DOF). The free body diagram of the shaft bearing system is provided in Fig. 1. Moreover, the deep groove ball bearing (6206) is the studied bearing. The model of the study bearing system is carried out using springs and lumped masses. The proposed model incorporates the following realistic assumptions and considerations.

- Balls are positioned equi-spaced balls around the shaft and there is no interaction between them.
- There is no slipping of the balls during rolling on the surface of races.
- The mass of the ball is negligible because it is relatively small compared with other bearing parts refer to the Table I.

- The study bearing operates under isothermal conditions.
- Forces act in radial directions along X and Y axes.
- The mass of the inner race is included in the mass of the shaft, and the mass of the housing incorporates the mass of the outer race.
- The shaft-housing under study is modelled using two masses (M_s and M_h), which yields a 2-DOF system.
- Nonlinear Hertzian contact deformations are considered at the contacts formed between balls and races.
- Damping due to lubricant film is ignored.

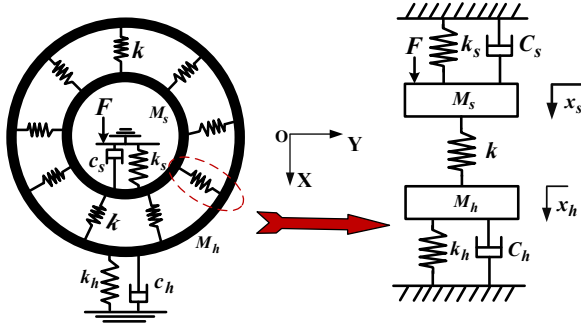


Figure 1. Free body diagram of the shaft-housing system

The geometric and physical properties of the studied bearing are presented in Table I.

TABLE I. GEOMETRIC AND PHYSICAL PROPERTIES USED FOR THE BALL BEARINGS

Geometric properties			Physical properties		
Symbol	value	Unit	Symbol	value	Unit
d_s	30	mm	E	210	KN/mm^2
d_i	37.48	mm	ν	0.3	-
d_o	56.45	mm	M_o	59.84	g
d_m	46.96	mm	M_i	32.4	g
D	9.48	mm	M_b	2.95	g
p_d	0.01	mm	M_s	5.26	Kg
N_b	9	-	M_h	1	Kg

where M_o , M_i , M_b , M_s and M_h are the masses of outer race, inner race, ball, shaft and housing respectively.

A. Load Deflection and Stiffness

Hertzian load deformation relationship is used in calculation of deformation at the contacts formed between ball and races of the bearing under investigation. The used relation is expressed as the following [9]:

$$Q = K\delta^n \quad (1)$$

where K is the load deflection factor or constant for Hertzian contact elastic deformation, δ is the radial deflection or contact deformation and n is the load

deflection exponent, $n = 3/2$ for ball bearing and $10/9$ for roller bearing [9].

The stiffness coefficient at the contacts formed between races and the i th ball are evaluated using the following relation [10, 11]:

$$K_{i,o} = \frac{2\sqrt{2}\left(\frac{E}{1-\nu^2}\right)}{3(\sum \rho)^{1/2}} \left(\frac{1}{\delta^*}\right)^{3/2} \quad (2)$$

where $K_{i,o}$ is inner and outer raceways to ball contact stiffness respectively, $\sum \rho$ is the curvature sum which is calculated using the radii of curvature in a pair of principal planes passing through the point contact. δ^* is the dimensionless contact deflection obtained using curvature difference (Refer Table 6.1, Rolling Bearing Analysis, Tedric Harris, Fourth Edition, 2001).

Total deflection between two raceways is the sum of the approaches between the rolling elements and each raceway [9], hence,

$$K = \left[\frac{1}{(1/K_i)^{1/(3/2)} + (1/K_o)^{1/(3/2)}} \right]^{3/2} \quad (3)$$

Fig. 2 shows the rigidly supported bearing subjected to radial load, the radial deflection at any rolling element angular position is given by [9]:

$$\delta_\phi = \delta_r \cos \phi - \frac{1}{2} P_d \quad (4)$$

where δ_r is the ring radial shift, occurring at $\phi = 0$ and P_d is the diametric clearance.

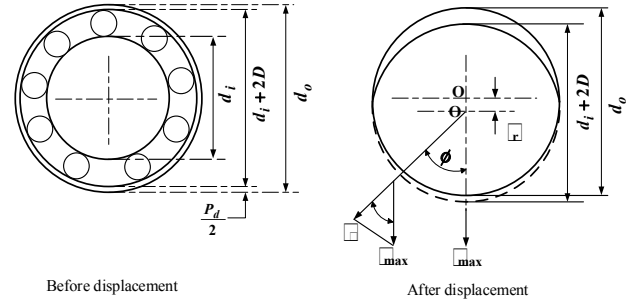


Figure 2. Radial deflection at a rolling element position [9]

If ϕ_0 is the initial position of the i th ball, the angular position ϕ_i at any time t is defined by the following relation:

$$\phi_i = \frac{2\pi}{N_b}(i-1) + \omega_c t + \phi_0, \quad i = 1, \dots, N_b \quad (5)$$

where the angular velocity ω_c of the cage can be expressed in terms of angular velocity ω_s of the shaft as:

$$\omega_c = \left(1 - \frac{D}{d_m}\right) \frac{\omega_s}{2}, \quad \omega_s = 2\pi f_s \quad (6)$$

The radial deflection at any rolling element angular position may be rearranged in terms of maximum deformation as follows:

$$\delta_\phi = \delta_{\max} \left[1 - \frac{1}{2\epsilon} (1 - \cos \phi_i) \right] \quad (7)$$

where $\varepsilon = \frac{1}{2} \left(1 - \frac{P_d}{2\delta_r} \right)$

Therefore, the contact force at any angular position is

$$Q_\phi = Q_{\max} \left[1 - \frac{1}{2\varepsilon} (1 - \cos \phi_i) \right]^{3/2} \quad (8)$$

In Fig. 3 it is clear that the overall applied radial load (F) is equal to the sum of the vertical components of the contact reactions of the rolling element loads. Mathematically, it is expressed as follows:

$$F = \sum_{\phi=0}^{\phi=\pm\psi_i} Q_\phi \cos \phi \quad (9)$$

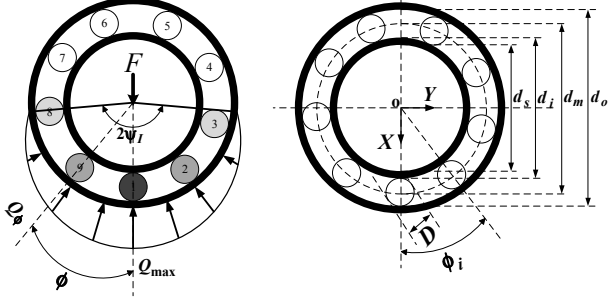


Figure 3. Load distributions in a ball bearing

The overall contact deformation δ for the i th ball is a function of the shaft displacement relative to the housing in the radial direction, ball position ϕ_i and the clearance c is provided by following expression.

$$\delta = (X_s - X_h) \cos \phi_i + (Y_s - Y_h) \sin \phi_i - c \quad (10)$$

where $c = P_d/2(1 - \cos \phi_i)$ is the internal radial clearance.

Since the Hertzian forces arise only when there is a contact deformation, the springs are required to act only in compression. In other words, the respective spring force comes into play when the instantaneous spring length is shorter than its unstressed length (the terms of δ should be positive), otherwise the separation between ball and race takes place and the restoring force is set to zero.

At the time of impact at the defect, a pulse of short duration is produced and it is accounted for by the term Δ additional deflection. Hence the expressions for δ is modified as:

$$\delta = (X_s - X_h) \cos \phi_i + (Y_s - Y_h) \sin \phi_i - c - \Delta \quad (11)$$

The total restoring force is the sum of the restoring force from each of the rolling elements. Thus the total restoring force components in the X and Y directions are

$$F_X = \sum_{i=1}^{N_b} K[\delta]^{3/2} \cos \phi_i \quad (12)$$

$$F_Y = \sum_{i=1}^{N_b} K[\delta]^{3/2} \sin \phi_i \quad (13)$$

B. Internal Clearance

Internal radial clearance is the geometrical clearance between the inner race, outer race and ball. Radial clearance is the play between the ball and raceway perpendicular with the bearing axis. The internal clearance

will significantly influence heat, vibration, noise, and fatigue life. For best rolling element bearing life and machine reliability, internal clearance at running conditions must be close to zero [9, 12]. Moreover, it is also necessary for the purpose of diagnosis to gain the knowledge of changes on the characteristic features when the clearance becomes large due to investable wear during the bearing service life.

The angular contact ball bearings are specifically designed to operate under radial and thrust load, and the clearance built into the unloaded bearing angle. In addition, there are five clearance groups, namely C2, C0 (Normal), C3, C4 and C5. Therefore, the radial internal clearances for radial contact deep groove ball bearing (6206) are presented in Table II [9].

TABLE II. RADIAL INTERNAL CLEARANCE FOR DEEP GROOVE BALL BEARING 6206 UNDER NO LOAD [ISO 5753]

Clearance values (μm)									
C2		C0		C3		C4		C5	
Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
1	11	6	20	15	33	28	46	40	64

C. Equation of Motion

Based on the assumption made, the governing equations for each mass in shaft and housing in X and Y directions can be developed. According to the motion direction in Fig. 1, the equations are:

$$M_s \ddot{X}_s + C_s \dot{X}_s + K_s X_s + \sum_{i=1}^{N_b} K[\delta]^{3/2} \cos \phi_i + F = 0 \quad (14)$$

$$M_s \ddot{Y}_s + C_s \dot{Y}_s + K_s Y_s + \sum_{i=1}^{N_b} K[\delta]^{3/2} \sin \phi_i = 0 \quad (15)$$

$$M_h \ddot{X}_h + C_h \dot{X}_h + K_h X_h - \sum_{i=1}^{N_b} K[\delta]^{3/2} \cos \phi_i = 0 \quad (16)$$

$$M_h \ddot{Y}_h + C_h \dot{Y}_h + K_h Y_h - \sum_{i=1}^{N_b} K[\delta]^{3/2} \sin \phi_i = 0 \quad (17)$$

D. Local Defect

1) Defect on the Inner Race

The location of defects on the inner race does not remain stationary since the inner race rotates at the speed of the shaft ω_s . Thus, the defect angle for the inner race α_{in} is defined as:

$$\alpha_{in} = \omega_s t \pm (W_{\text{defect}}/d_i) \quad (18)$$

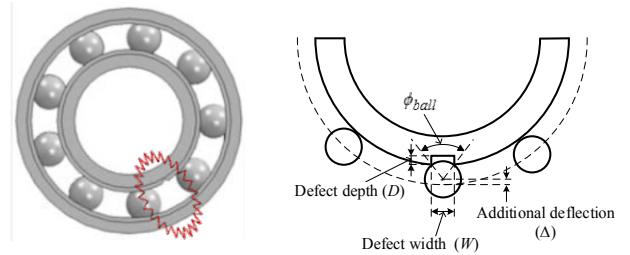


Figure 4. Additional deflection of ball due to defect on inner race

The rolling ball approaches the defect either in the loaded zone or the unloaded zone, therefore, the deflection δ of the i th ball varies. Additional deflection at the defect Δ of the ball when it passes through the defect is defined by the width of the defect as following:

$$\Delta = (D/2 - D/2 * \cos(0.5\phi_{ball})) \quad (19)$$

where ϕ_{ball} = width of the defect / radius of the ball.

The position of the i th ball in the defect zone is mathematically defined as:

$$(\omega_s t - W_{defect} / d_i) \leq \phi_i \leq (\omega_s t + W_{defect} / d_i) \quad (20)$$

2) Defect on the Outer Race

The defect on the outer race is located at an angle α_{out} from the X axis. The local defects on the outer race are normally found in the loaded region. Moreover, the stationary outer race means that the position of the defect usually does not change. Whenever a ball passes over the defect location, it has additional deflection Δ .

The angular position of i th ball passing the defect zone is mathematically defined by the following relation:

$$(\alpha_{out} - W_{defect} / d_o) \leq \phi_i \leq (\alpha_{out} + W_{defect} / d_o) \quad (21)$$

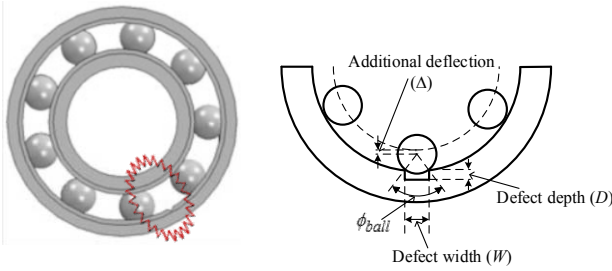


Figure 5. Additional deflection of ball due to defect on outer race

III. DIAGNOSTIC FEATURES WITH CLEARANCE

The above equations (14-17) are second order nonlinear differential equations. To solve these equations each of the second order equations is converted into two first order differential equations. The Runge-Kutta method is used to solve the first differential equation set in Matlab environment.

A. Initial Conditions

The governing equations of motion are solved based on the ball positions (5) and the deflection relation (11) at each step of time. Displacement in X and Y directions and velocity \dot{X} and \dot{Y} at time $(t + dt)$ are calculated. The time step (dt) of $6 \mu s$ for nine cycles has been considered. The initial displacements and velocities in X and Y directions are set to zeroes.

The shaft speed of 1500 rpm (25 Hz) with 1000 N radial load is considered as the normal operating condition of the bearing. Two defect sizes with width 0.6 mm and 2.0 mm on both the inner race and outer race will be studied subsequently, which are denoted as small and large faults respectively. The details of the calculation process are summarized in the flow chart of Fig. 6.

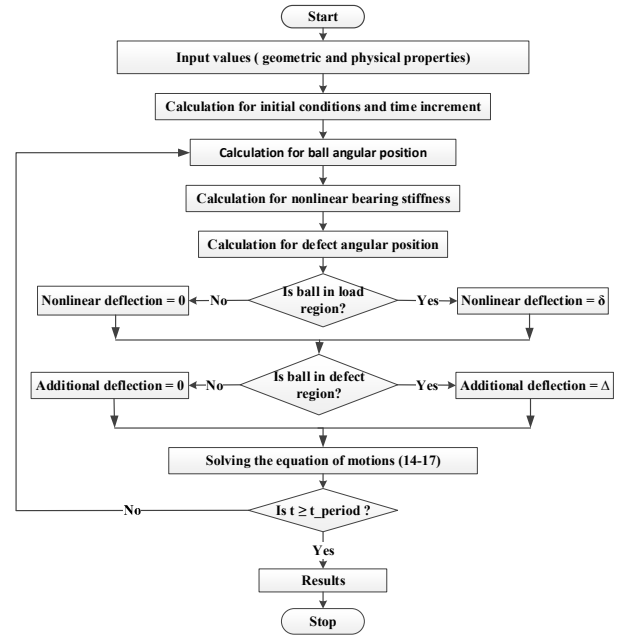


Figure 6. The flow chart of calculation process

B. Diagnosis of Outer Race Defect

The defect on outer race often occurs in loading zone and will have a constant angular position, usually corresponding to the applied direction of the external loading. Therefore, when a moving ball approaches to the defect same magnitude of impulse is expected for every time of contact between ball and defect. On the other hand, the defect may also appear at the start of loading zone due to poor lubricant conditions. Therefore, two outer race defect locations (0 degree and 320 degrees) have been examined for two defect widths 0.6 mm and 2.0 mm under four incremental different clearance values 1, 10, 30 and $60 \mu m$.

1) Outer Race Defect at 0 Degree

Fig. 7 presents vibration displacements of the housing in X direction for baseline, small defect and large defect when the clearance values are at 1, 10, 30 and $60 \mu m$. It can be seen that the baseline case, where there is no defect on the races, exhibit a clear increase in vibration displacement with clearance increments, showing that the amplitude of local becomes higher with larger clearances. Moreover, for the two defect cases, the periodic vibration amplitudes also show significant increases with clearances and large defect have higher amplitudes.

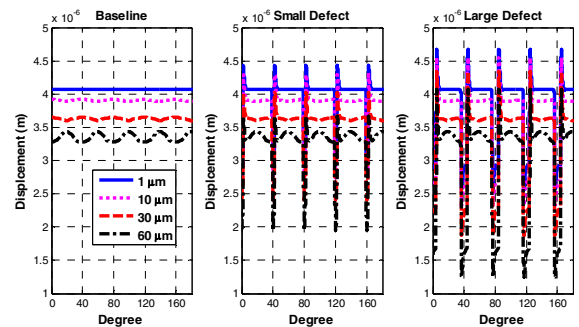


Figure 7. Housing displacements in X direction

To show the impact behavior, Fig. 8 shows the vibration velocities of housing in X direction of baseline, small defect and large defect at four different clearance values. For the baseline case, clear increase in vibration velocity amplitude can be seen as result of clearance increases. For the defect case, the impulse caused by the contact between ball and defect is repeated each 40 degrees. Moreover, both the impulse magnitude and duration increase with the defect size.

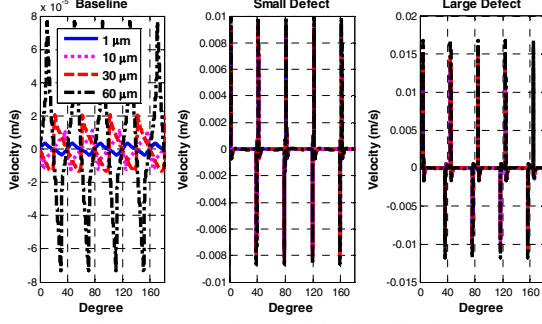


Figure 8. Housing velocities in x direction

The frequency spectra of housing acceleration in X direction of baseline, small defect and large defect cases are illustrated in Fig. 9. For the baseline the loading frequency which is the cage frequency multiply by the number of balls can be clearly appeared. For the defect case, the calculated ball pass frequency outer race (BPFO) is 89.8Hz. The fault feature frequency and its harmonics are obviously obtained. Moreover, the increase of the defect acceleration magnitude is caused by the defect size and nonlinear deflection.

$$BPFO = \frac{N_b}{2} F_s \left(1 - \frac{D}{d_m} \cos \phi\right) \quad (22)$$

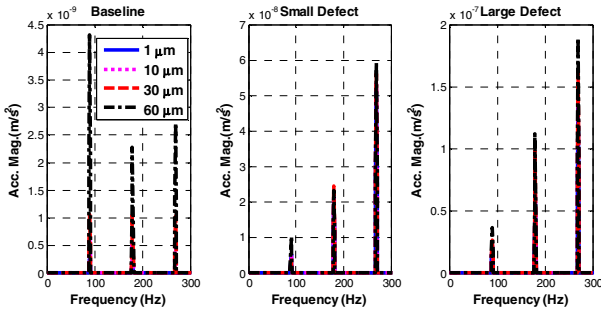


Figure 9. Housing acceleration spectrum in X direction

The acceleration amplitudes of housing in X direction of baseline, small defect and large defect cases are shown in Fig. 10. For the baseline case, the loading frequency amplitude is increased owing to large clearances. For the defect case, the BPFO magnitude and its harmonics are increasing with the increase of the clearance values. Additionally, the magnitude of the BPFO of the small defect is greatly affected by the loading frequency.

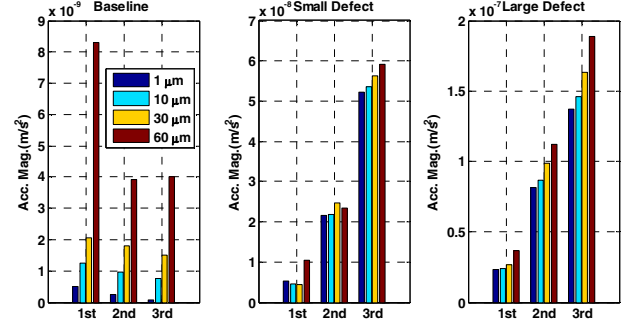


Figure 10. Acceleration amplitude on house of first three harmonics of fault characteristic frequency

2) Outer Race Defect at 320 Degrees

The vibration acceleration amplitudes of defective outer race at 320 degrees in X direction of small defect and large defect cases are shown in Fig. 11. The magnitude of the BPFO is greatly affected by the loading frequency. Furthermore, the BPFO magnitude and its harmonics are decreased with the increase of the clearance values.

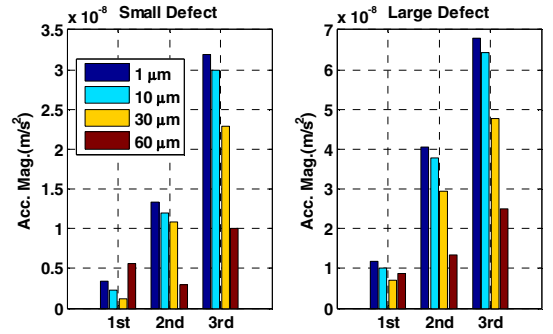


Figure 11. Acceleration amplitude on house of first three harmonics of fault characteristic frequency

The vibration acceleration amplitudes of different position of outer race defect show significant changes. When the defect is at 0 degree, the amplitude of vibration is maximum. As the position of the defect is changed away from this position, it is observed that the amplitude of vibration reduces. This variation can be seen in Fig. 10 and Fig. 11.

C. Diagnosis of Inner Race Defect

Rotating inner race defect generates a complicated vibration signal due to rotation of both defect and balls. The amplitude of the inner race defect is not constant due to the varying load on ball and defect contacts.

Vibration displacements of the housing in X direction at four different clearance values of small defect and large defect cases are presented in Fig. 12. As the rolling ball approaches the defect either in the loaded zone or the unloaded zone the periodic vibration amplitudes show significant variation.

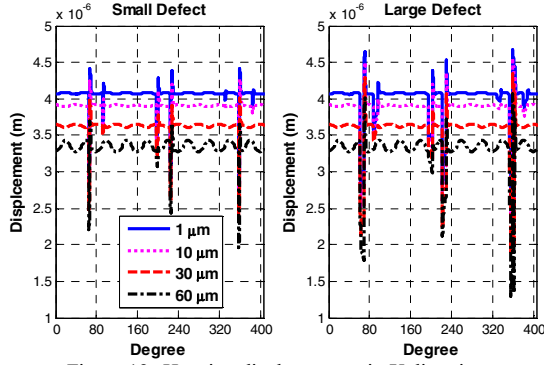


Figure 12. Housing displacements in X direction

The housing vibration velocities in X direction of small defect and large defect cases are presented in Fig. 13. It is clearly indicated that the velocity magnitude increases when the defect and ball contact accrue in the load zone and decrease in the unloaded zone.

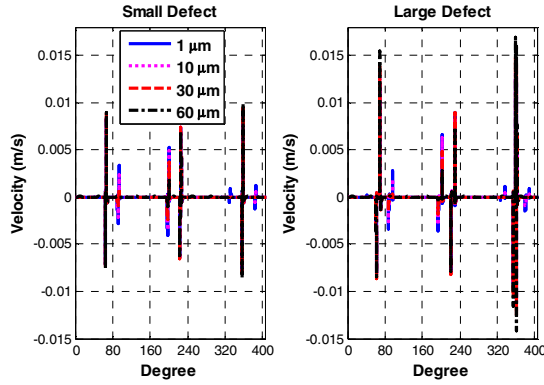


Figure 13. Housing velocities in X direction

The frequency spectra of housing vibration in X direction of small defect and large defect cases are shown in Fig. 14. The calculated ball pass frequency inner race (BPFI) is 135.198 Hz. The fault feature frequency and shaft rotational frequency and their harmonics are clearly visible. Furthermore, the inner race defect rotates at shaft speed, so the BPFI is amplitude modulated by the shaft rotational frequency. Therefore, peaks at frequencies $BPFI \pm f_s$ and their harmonics are also found.

$$BPFI = \frac{N_b}{2} F_s \left(1 + \frac{D}{d_m} \cos \phi\right) \quad (24)$$

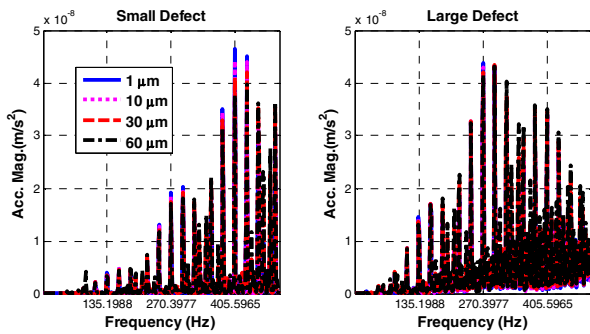


Figure 14. Housing acceleration spectrum in X direction

The acceleration amplitudes of small and large inner race defect cases are presented in Fig. 15. The defect

amplitude is decreasing with the increase of the clearance value. For the large defect, the third harmonic amplitude is greatly affected by the system frequency response.

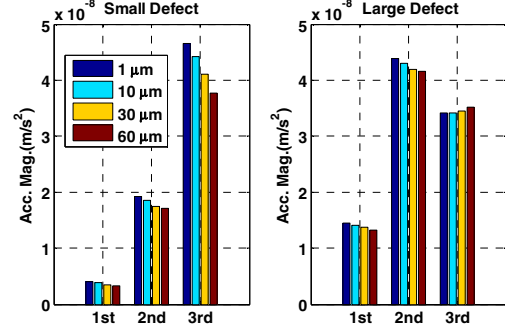


Figure 15. Acceleration amplitude on house of first three harmonics of fault characteristic frequency

IV. CONCLUSION

A dynamic model for deep groove ball bearings considering internal radial clearance as well as localized defects on inner and outer races is developed to obtain the vibration responses for bearing diagnosis. The bearing is modelled as a 2-DOF system, vibrations of shaft and housing in X and Y directions are studied. The vibration acceleration amplitudes and frequencies are simulated by solving the coupled nonlinear equation of motions using Matlab.

The model predicts that the vibration responses increase with internal radial clearances which will become large during bearing service period due to inevitable wear. As expected, the large the clearance the higher amplitude of the diagnostic feature for the outer race defects. In addition, the defect at loading zone produces higher amplitude.

However, the large clearance reduces the amplitudes of the feature for inner race defects. Therefore, the severity of inner race fault needs to be determined by taking into account bearing service duration and bearing grades.

REFERENCES

1. McFadden, P.D. and J.D. Smith, *Model for the vibration produced by a single point defect in a rolling element bearing*. Journal of Sound and Vibration, 1984. 96(1): p. 69-82.
2. Purohit, R. and K. Purohit, *Dynamic analysis of ball bearings with effect of preload and number of balls*. International Journal of Applied Mechanics and Engineering, 2006. 11(1): p. 77-91.
3. Culita, J., D. Stefanioiu, and F. Ionescu, *Simulation models of defect encoding vibrations*. Journal of Control Engineering and Applied Informatics, 2007. 9(2): p. 59-67.
4. Sassi, S., B. Badri, and M. Thomas, *A numerical model to predict damaged bearing vibrations*. Journal of Vibration and Control, 2007. 13(11): p. 1603-1628.

5. Upadhyay, S., S. Harsha, and S. Jain, *Analysis of nonlinear phenomena in high speed ball bearings due to radial clearance and unbalanced rotor effects*. Journal of Vibration and Control, 2010. 16(1): p. 65-88.
6. Patil, M.S., et al., *A theoretical model to predict the effect of localized defect on vibrations associated with ball bearing*. International Journal of Mechanical Sciences, 2010. 52(9): p. 1193-1201.
7. Dougdag, M., et al., *An experimental testing of a simplified model of a ball bearing: stiffness calculation and defect simulation*. Meccanica, 2012. 47(2): p. 335-354.
8. Patel, V.N., N. Tandon, and R.K. Pandey, *Vibration Studies of Dynamically Loaded Deep Groove Ball Bearings in Presence of Local Defects on Races*. Procedia Engineering, 2013. 64(0): p. 1582-1591.
9. Harris, T.A. and M.N. Kotzalas, *Rolling Bearing Analysis, Fifth Edition - 2 Volume Set*. 2006: Taylor & Francis.
10. Arslan, H. and N. Aktürk, *An investigation of rolling element vibrations caused by local defects*. Journal of Tribology, 2008. 130(4): p. 041101.
11. Karacay, T. and N. Akturk, *Vibrations of a grinding spindle supported by angular contact ball bearings*. Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics, 2008. 222(1): p. 61-75.
12. Oswald, F.B., E.V. Zaretsky, and J.V. Poplawski, *Effect of Internal Clearance on Load Distribution and Life of Radially Loaded Ball and Roller Bearings*. Tribology Transactions, 2012. 55(2): p. 245-265.

Pose Estimation Using Visual Entropy

Jianjun Gui*, Dongbing Gu and Huosheng Hu

*School of Computer Science and Electronic Engineering, University of Essex
Wivenhoe Park, Colchester CO4 3SQ, UK. Emails: jgui@essex.ac.uk*

Abstract—The trend of using visual information for pose estimation of camera becomes increasingly popular and diverse. In this paper, we propose a pose estimation based on visual information entropy. The constructed cost functions are robustness by natural and suitable to be applied in pose estimation in highly agile platforms. Especially, using the visual entropy or salient features as observation causes dramatically difference in computational time. The experiments based on real data from office environment shows that the entropy-based pose estimation using mutual information has huge potential in performance.

Index Terms—Pose tracking, Entropy, Mutual information

I. INTRODUCTION

As an easy means to observe the world, visual sensing has been widely adopted in all kinds of platform. Nowadays, it is very normal to find a tiny but high quality camera in our smart phones, laptops or even some electronic toys like drones. For academic research, “What can we get from the images?” is always ranked as a hot topic and has been developed to various branches according to different applications scenarios.

In computer vision community, camera poses estimation are often completed by the bundle adjustment [1] which can refine the environmental structure and camera poses at the same time. Visual odometry (VO) estimates 3D motion of camera sequentially, while in simultaneous localisation and mapping (SLAM), a global consistent map also should be built [2].

Among the trend of visual based pose estimation, there exists two major methods to deal with visual information. The most standard one is to extract a sparse set of salient features, then express them using feature descriptors. Although the feature-based methods come with an obvious limitation, i.e. only the scene contains a certain feature pattern (e.g. corners, lines) can be expressed and used, the successful rotation and scale invariant characters of some feature detectors and descriptors still make methods in visual SLAM or visual odometry (e.g. PTAM [3], monoSLAM [4]) be commonly accepted.

On the other hand, the direct methods recover the structure or motion directly from the intensity and/or gradient of the sequential images, on the occasion, the magnitude and/or direction of pixels can be used in pose estimation compared to only sparse features. The most prominent character of direct based methods is that all information in one image can

be exploited, even for the environments with little texture, few keypoints or huge impact of camera defocus and blur. Only recent years, some direct based visual odometry methods have been proposed and become popular. Methods in [5], [6] built a fully dense depth maps on per pixel basis and the computational ability of state-of-art GPU were adopted. Researchers in [7] cut down dense region to reduce the computational burden and proposed a semi-dense depth mapping method. The method in [8] combined direct information with keypoints repetitively to the enhance intra-frame tracking results.

As the value of intensity is quite sensitive to illumination change, direct based methods would trigger large difference in pixels even with small movements, which can lead to divergence in algorithm. However, the statistic information reflected by the intensity is more stable than the intensity value itself. And it happens that the conception of entropy in information theory gives the researchers a clue to deal with the intensity information. Considering two images as different signal sources, the alignment between them can be expressed by the value of mutual information. This idea has been widely used in image registration [9] or visual servo [10].

In this paper, we present two different entropy-based expressions for pose estimation. While one adopts direct information on per pixel basis, the other uses the traditional features as measurements. Experimental results show the efficiency and robustness for pose estimation and further mapping or global alignment. It is also revealed that the method using mutual information on the whole image is superior in computational complexity when compared to feature based methods.

This paper is structured as follows: in the first section, a generalized projection of spatial point is abstracted as a basic model, followed by an overview of the entropy regarding visual information as signals in section III. In section IV, a cost function is constructed using the entropy. Some experiments based on real data are analysed in the aspects of robustness, stability and efficiency in section V.

II. A GENERAL PROJECTION MODEL

A camera motion in free space can be modelled as in Figure 1. The two images are sequential where the motion ξ is small and rotation is remitted, or large span in space

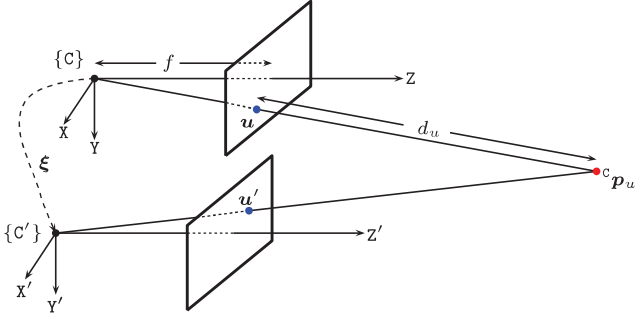


Fig. 1. A general projection model. A 3D point ${}^c p_u$ is projected into two image frames linked by motion ξ .

where ξ contains rotation and translation. A point u in the image is regarded as a pixel or a extracted feature and its paired u' is in the second image. Both of the points together with two camera centres and the corresponding spatial point ${}^c p_u$ are coplanar, forming geometry constraints for pose estimation. In order to reduce the computational complexity, the transformation is usually expressed in a non-redundant way.

A 3D rigid body transformation $T \in SE(3)$ denotes rotation and translation in 3D:

$$T = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \quad \text{with } R \in SO(3) \quad \text{and } t \in \mathbb{R}^3$$

The optimisation purpose in image alignment is to find the transformation T in each time step, i.e. T is regarded as the camera pose. A minimal representation for camera pose is better for optimisation purpose. The Lie algebra $se(3)$ corresponding to the tangent space of $SE(3)$ at the identity is used as the minimal representation. The algebra element is called twisted coordinates $\xi = [\omega^T \ v^T]^T \in \mathbb{R}^6$. The map from Lie algebra $se(3)$ to Lie group $SE(3)$ is the exponential map $T(\xi) = \exp(\psi(\xi))$ and its inverse map is the logarithm map $\psi(\xi) = \log T(\xi)$, where $\psi(\xi)$ is the wedge operator,

$$\psi(\xi) = \begin{pmatrix} [\omega \times] & v \\ 0 & 1 \end{pmatrix}$$

A 3D point with homogeneous vector ${}^c p_u$ in the camera frame maps to the image coordinate u via the pinhole camera projection model:

$$u = \pi({}^c p_u) = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} + \begin{bmatrix} f_u & 0 \\ 0 & f_v \end{bmatrix} \begin{pmatrix} x/z \\ y/z \end{pmatrix} \quad (1)$$

where u_0, v_0 and f_u, f_v are the principal point and focal length, respectively, representing camera intrinsic parameters which can be calibrated beforehand. Given a depth information d_u for a point u , the 3D point in the camera frame can be recovered from an image coordinate:

$${}^c p_u = \pi^{-1}(u, d_u) \quad (2)$$

III. VISUAL ENTROPY

A. Information entropy

Generally, the entropy is defined for two different levels of description of a given system: the macroscopic state of the system is defined by a distribution on the micro-states that are accessible to a system in the course of its thermal fluctuations. At one of these levels, the entropy S is given by the Gibbs' entropy formula, $S = -k_B \sum_i p_i \ln p_i$, where p_i is the probability that it occurs during the system's fluctuations and the quantity k_B is a physical constant known as Boltzmann's constant.

In information theory, entropy $H(X)$ defines the theoretical number of bits needed to encode a random variable X . This random variable could stand for an event, sample or character drawn from a distribution or data stream. Here, for our camera pose estimation, we consider the images or regions of interest as the random variables with possible valuables as the intensity of pixels whose possibility can be calculated from their histogram. Thus, if a discrete random variable X with possible values $\{x_1, \dots, x_n\}$, then the Shannon entropy $H(X)$ is given by:

$$H(X) = \sum_{i=1}^n P(x_i) I(x_i) = - \sum_{i=1}^n P(x_i) \log_2(P(x_i)) \quad (3)$$

where $I(x) = -\log_2(P(x))$ represents the self-information. For mathematical completeness, $0 \log_2(0) = 0$ should be introduced. Intuitively, the more values x are equally probable the bigger entropy $H(X)$ is, and the entropy reaches maximum: $H(p_1, \dots, p_n) \leq H(1/n, \dots, 1/n) = \log_2(n)$. While pose estimation matters sequential images rather than only one image itself, building up the relation between two random variables becomes quite important.

B. Joint entropy

Joint entropy $H(X, Y)$ describes the uncertainty associated with a set of variables, defining the theoretical number of bits needed to encode a joint system of two random variables X, Y with possible values $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_m\}$ respectively.

$$H(X, Y) = - \sum_{i=1}^n \sum_{j=1}^m P(x_i, y_j) \log_2(P(x_i, y_j)) \quad (4)$$

In the above expression, $P(x, y)$ is the joint probability of these values occurring together.

The joint entropy is bounded by $\max(H(X), H(Y)) \leq H(X, Y) \leq H(X) + H(Y)$, the second equality happens if and only if X and Y are statically independent, while the first inequality becomes equal when $Y \subseteq X$ or $X \subseteq Y$, which means X or Y can fully represent the other. In this case, a alignment problem can be viewed as finding the minimum of the joint entropy, thus making the two set of variables included by one another in a large extent.

However, it is not such simple for us just using the joint entropy to construct a cost function for pose estimation. In practical, adding variable Y into the system only increases the joint entropy $H(X, Y)$ but adding a variable set, which is exactly equal to original set X , i.e. $X = Y$, or only a constant set (probability is zero) would not add variability to the system. In the constant case, the set Y obviously can not express X , thus the alignment fails. However, the conception of mutual information has been proposed to solve this issue [11].

C. Mutual information

The mutual information, also formally called transinformation of two random variables, is a measure of the variables' mutual dependence. For two random variables X and Y , their mutual information is given by the following equation:

$$I(X, Y) = H(X) + H(Y) - H(X, Y). \quad (5)$$

Substituting equation (3) and (4), it yields to:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \left(\frac{P(x, y)}{P(x)P(y)} \right) \quad (6)$$

where $P(x, y)$ is the joint probability distribution function of X and Y , and $P(x)$ and $P(y)$ are the marginal probability distribution functions of X and Y respectively.

As we can see in equation (5), mutual information integrates the individual entropy of each variable and their joint entropy. If the random variables X and Y are independent, then $I(X, Y) = 0$; if they totally depends on each other, say $X = Y$, then $I(X, Y) = H(X) = H(Y)$. It can be easily checked that even for the set Y including most constant outliers, $I(X, Y) = 0$ due to $H(Y) = 0$, then the alignment will fail. Thus, the bigger the mutual information is, the better the aligned results would be. Compare to the sum of squared differences (SSD), this alignment method does not need to find the linear relation between two signals [12].

IV. POSE ESTIMATION METHODS

A. Pose estimation based on mutual information

The goal of visual-based estimation is to find the current pose through image alignment. Here the general random variables in above section become the images and the pixel intensity becomes possible value within image domain.

As stated in section II, a rigid body transformation ξ denotes rotation and translation and a point in 3D space with homogeneous vector ${}^c\mathbf{p}_u$ in the camera frame maps to the image coordinate \mathbf{u} via the camera projection. With this transformation ξ , the mutual information of two images is given by:

$$I(X, Y) = \sum_i \sum_j P(i, j, \xi) \log \left(\frac{P(i, j, \xi)}{P(i, \xi)P(j)} \right) \quad (7)$$

where i and j are pixel intensities in images \mathcal{I}_c and $\mathcal{I}_{c'}$, respectively. $P(i, \xi)$ and $P(j)$ are respectively the probability of the intensity i and j in the images, and $P(i, j, \xi)$ is the joint probability of two intensities. They can be computed as a normalized histogram:

$$\begin{aligned} P(i, \xi) &= \frac{1}{N} \sum_u \Phi \left(i - \frac{\lambda'}{\lambda} \mathcal{I}_c(\mathbf{u}, \xi) \right) \\ P(j) &= \frac{1}{N} \sum_{u'} \Phi \left(j - \frac{\lambda'}{\lambda} \mathcal{I}_{c'}(\mathbf{u}') \right) \\ P(i, j, \xi) &= \frac{1}{N} \sum_u \Phi \left(i - \frac{\lambda'}{\lambda} \mathcal{I}_c(\mathbf{u}, \xi) \right) \Phi \left(j - \frac{\lambda'}{\lambda} \mathcal{I}_{c'}(\mathbf{u}) \right) \end{aligned} \quad (8)$$

where N is the number of pixels in selected subset, λ is the gray levels of the original image and λ' is a desired value space with smaller scale to increase the computational efficiency and robustness [10]. $\Phi(\cdot)$ is B-spline function [13] with the character of unit result for integral operation, thus getting rid of renormalising.

Having the mutual information between images, we can acquire the best pose estimation through an optimization problem as

$$\xi^* = \arg \min_{\xi} (-I(\mathcal{I}_c(\xi), \mathcal{I}_{c'})). \quad (9)$$

This can be solved by Levenberg-Marquardt [14],

$$\xi = -\alpha (\mathbf{H}_{\xi} + \beta [\mathbf{H}_{\xi}]_d)^{-1} \mathbf{G}_{\xi}^T \quad (10)$$

where \mathbf{H}_{ξ} and \mathbf{G}_{ξ} are the Gradient and Hessian matrix of equation (9) with the parameters α and β . The Gradient can be computed as

$$\begin{aligned} \mathbf{G}_{\xi} &= -\frac{\partial I(\mathcal{I}_c(\xi), \mathcal{I}_{c'})}{\partial \xi} \\ &= -\sum_{i,j} \frac{\partial P(i, j)}{\partial \xi} \left(1 + \log \left(\frac{P(i, j)}{P(i)} \right) \right) \end{aligned} \quad (11)$$

with the joint probability that can be calculated further by chain rule in equation (8),

$$\frac{\partial P(i, j)}{\partial \xi} = -\frac{1}{N} \sum_u \frac{\lambda'}{\lambda} \frac{\partial \Phi}{\partial \mathcal{I}_c} \frac{\partial \mathcal{I}_c}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \xi} \Phi \left(j - \frac{\lambda'}{\lambda} \mathcal{I}_{c'} \right) \quad (12)$$

where $\frac{\partial \mathcal{I}_c}{\partial \mathbf{u}}$ is the image gradient and $\frac{\partial \mathbf{u}}{\partial \xi}$ is the interaction matrix at point \mathbf{u} , given by a the projection in section II:

$$\frac{\partial \mathbf{u}}{\partial \xi} = \begin{pmatrix} -1/d_u & 0 & u/d_u & uv & -(1+u^2) & v \\ 0 & -1/d_u & v/d_u & 1+v^2 & -uv & -u \end{pmatrix} \quad (13)$$

The depth value d_u above is topically estimated through a mapping process in VO or SLAM. Having the Gradient matrix the Hessian is approximately given by omitting a higher order term.

B. Pose estimation based on entropy-like cost function

For a general non-linear system with parameter state vector $\mathbf{x}(\xi)$ and other known inputs \mathbf{u} , an observation \mathbf{z}_i , $i \in 1, 2, \dots, N$ can be written as

$$\mathbf{z}_i = h(\mathbf{u}, \mathbf{x}(\xi)) + \mathbf{v}_i \quad (14)$$

where \mathbf{v}_i refers to the process noise in each observation. A relative squared residuals ϵ_i is defined as

$$\epsilon_i = \frac{\tilde{z}_i^2}{\sum_{j=1}^N \tilde{z}_j^2} \quad (15)$$

$$\tilde{z}_i = \mathbf{z}_i - \hat{\mathbf{z}}_i$$

where $\hat{\mathbf{z}}$ is the estimation of the observation containing the unknown estimation $\hat{\xi}$ and the relative squared residuals as the normalised factor can be regarded as some sort of the probability. Thus, we can write a normalised entropy-like cost function as follows:

$$H = -\frac{1}{\log N} \sum_{i=1}^N \epsilon_i \log \epsilon_i \quad (16)$$

with $H = 0$ if $\sum_{i=1}^N \epsilon_i^2 = 0$. The zero condition represents an exact matching and also completes the definition.

With all above definitions, a cost function can be written as

$$\xi^* = \arg \min_{\xi} \left(-\frac{1}{\log N} \sum_{i=1}^N \epsilon_i \log \epsilon_i \right). \quad (17)$$

An intuitive understanding of this cost function is that the best matching leading to less residuals and more outliers would make large residuals. In order to increase the robustness, the relative squared residuals are often wrapped by Huber-like function [15] to limit the upper bound of the denominator.

C. Feature based pose estimation

Here for comparison, we briefly list a weighted cost function to find the camera pose and adjust the features' position at the same time. Normally, the re-projection error is defined as the difference between a measurement \mathbf{z} and its estimate $\hat{\mathbf{z}}(\hat{\xi}, {}^w\hat{\mathbf{p}}_u)$:

$$\tilde{\mathbf{z}} = \mathbf{z} - \hat{\mathbf{z}}(\hat{\xi}, {}^w\hat{\mathbf{p}}_u)$$

The cost function $\eta(\hat{\xi}, {}^w\hat{\mathbf{p}}_u)$ is the sum of all squared errors $\tilde{\mathbf{z}}$ with a weighting matrix \mathbf{W} :

$$\eta(\hat{\xi}, {}^w\hat{\mathbf{p}}_u) = \sum_{i=1}^n \sum_{j=1}^m \tilde{\mathbf{z}}_{i,j}^T \mathbf{W}_{i,j} \tilde{\mathbf{z}}_{i,j}$$

where j from 1 to m is the index of points within a frame, and i is the number of frames indexing a set with size n .

Then, the optimisation problem is defined as

$$\xi, {}^w\mathbf{p}_u = \arg \min_{\xi, {}^w\mathbf{p}_u} \eta(\hat{\xi}, {}^w\hat{\mathbf{p}}_u)$$

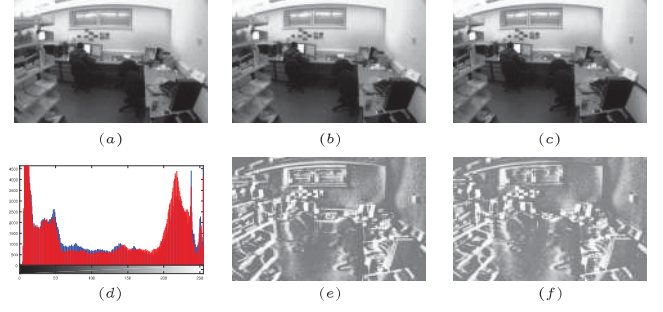


Fig. 2. Samples in sequential images. The first row displays three consecutive images; (d) is the histogram of (a) in red and (c) in blue, indicating the illumination variance; (e) is the error image of (a) and (b), while (f) is of (b) and (c), indicating the slight motion between consecutive frames.

V. RESULTS

This section presents some results of the process in pose estimation using different expressions mentioned above. We focus on the aspects of environmental condition change, average computational time before optimisation calculation, and the threshold of mutual information. The results indicate that with an equivalent accuracy, the entropy based methods enjoy the merits of robustness and computational efficiency and it is worth to further develop for applications in high flexible platforms like micro aerial vehicles (MAVs). The results were performed by applying consecutive images taken from our office environment. The frame rate is 30 Hz with the resolution of 640×480 and the lens is of 120° view angle with fish-eye distortion. All computational work was done in a laptop with Intel i7-5600U CPU and 8G RAM. Some sample sequential images in our trail are shown in Figure 2.

A. Robust to illumination

Some experiments took place under illumination change environment, for example turning off the lights during one trial. This simulates the situation which happens usually in real-time application: the camera moving from indoor to outdoor or merely used in dark environment. At our presented scene in Figure 3, there are four consecutive images including two under less illumination conditions. The matching lines between features show that after changing the illumination, matching pairs become less in number (decreasing 20 % as previous), limited in area (focusing on right area other than full image range) and more in mistake (outliers). The results indicate that in the direct-based expression, the pixel intensity are expressed in probability according to the histogram of whole image, which means except for the environment of totally dark, changing the illumination of the whole image would make less impact on the matching results compared to feature-based methods.

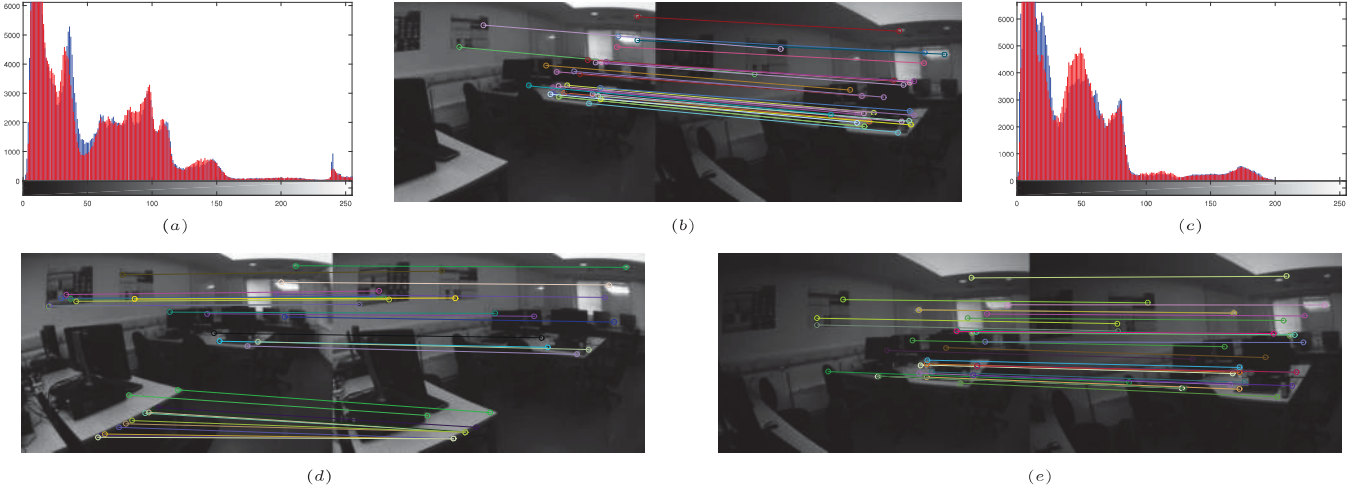


Fig. 3. Illumination changes during one experiment. (d) is the first two frames and their matching points under normal light condition, while (e) is the last two frames after turning off half of the lights in the lab. The histograms (a) and (c) are of (d) and (e) respectively. (b) shows the match pairs between a normal frame and a less illuminated frame with restricted matching region.

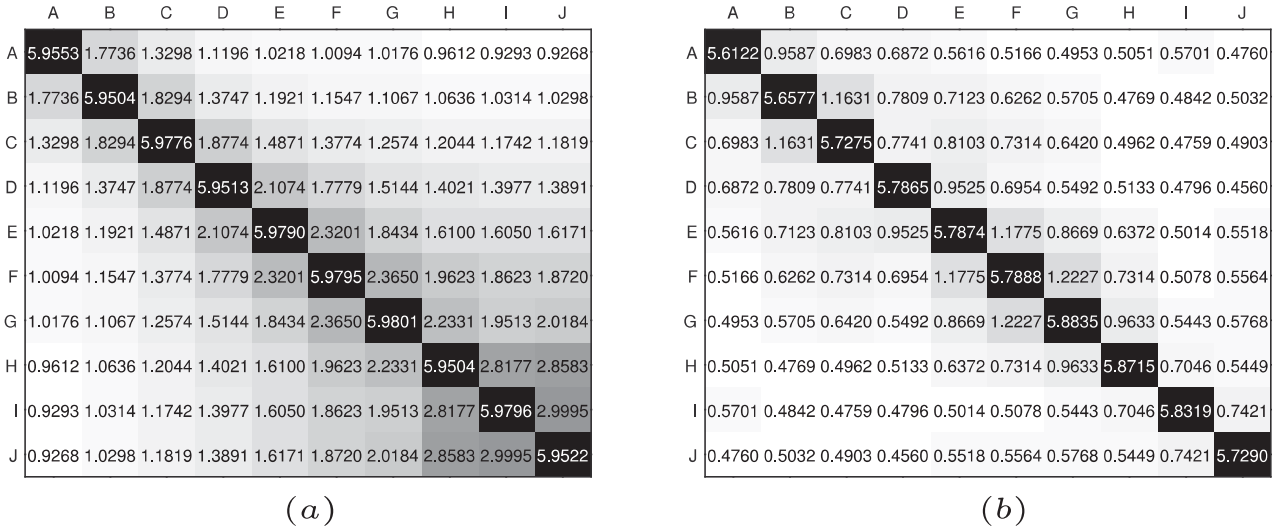


Fig. 4. Mutual information for two different trials. (a) 10 frames extracted from original sequence in every two frames; (b) 10 frames extracted in every ten frames. The value is the mutual information between the numbered frames from A to J.

B. Stable in mutual information

In order to show the availability and trend of mutual information, we formed two image sequences including 10 frames, which down-sampled from original sequence at 1/2 and 1/10 of original rate. In Figure 4, we can obviously view that values of mutual information wave around 2 between every other two frames (the diagonal line starting from B in Figure 4(a)), and the values jolt around 1 between every ten frames (the diagonal line starting from F in Figure 4(a) and B in Figure 4(b)). The values drop from average 5.8 to 2 dramatically between close frames but slightly between dis-

tant frames (four of five intervals). Thus, based on this value, a distance threshold as a constraint for motion optimisation can be set. The stable variance of mutual information also reveals the robustness character of entropy-based evaluation although the environment changes significantly.

C. Efficient in computation

Figure 5 shows the computational times of mutual information based on direct image information and standard SIFT process. The average computational time for mutual information for two trials are 0.0594s and 0.0440s, but for feature process they are 20.8772s and 18.5485s. The processes were

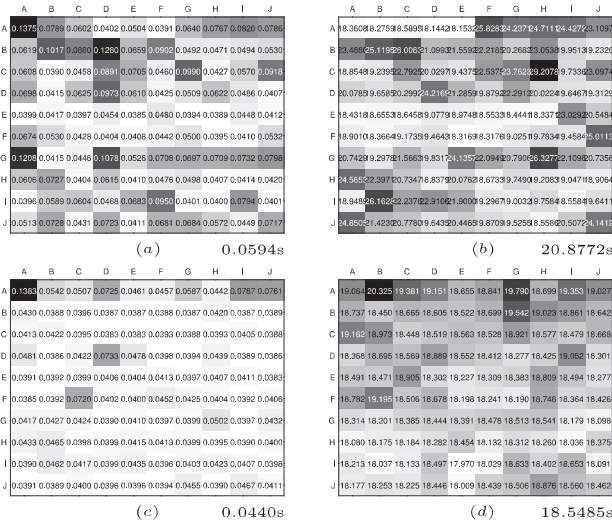


Fig. 5. Computational time map. The left column is the calculation time of mutual information between corresponding frames, while the right column is the computational time for the periods of feature process. The first row is the trial on the same image sequence and the second row is in another image sequence.

implemented in double directions, i.e. for every two images, they were computed from one to the other then reverse. The average time involves in all these calculations. It is worth mentioning here, if we exploit the same optimal method, the computational time in the optimisation period would be more or less the same. The most time consuming part is the feature extraction and expression, thus for clear comparison, we only consider the periods before real optimal calculation. These results indicate the huge potential of entropy based methods on direct image information.

VI. CONCLUSION

In this paper, we introduced the entropy-based pose estimation methods using visual information from a monocular camera. Stemming from information theory, entropy-like methods by natural have the robustness due to the use of probability, thus widely used to construct the cost function for optimisation. In another aspect, different visual observations and expressions used under the same conception of entropy will trigger dramatically different character in performance. Specifically, entropy-based optimisation depending on direct data from images or features extracted through a process will cost dramatically different computational consumption in whole algorithm. The entropy-based method using mutual information also shows the advantage in illumination variable environment. Through the experiments, we can see the huge potential in applying the entropy-based pose estimation methods in highly agile platforms like MAVs. In the future, we will further improve and analyse these methods and target at more real-time applications.

ACKNOWLEDGMENT

The first author has been financially supported by scholarship from China Scholarship Council.

REFERENCES

- [1] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment: modern synthesis," in *Vision algorithms: theory and practice*. Springer, 2000, pp. 298–372.
- [2] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *Robotics & Automation Magazine, IEEE*, vol. 18, no. 4, pp. 80–92, 2011.
- [3] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*. IEEE, 2007, pp. 225–234.
- [4] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 1052–1067, 2007.
- [5] M. Pizzoli, C. Forster, and D. Scaramuzza, "Remode: Probabilistic, monocular dense reconstruction in real time," in *Proc. IEEE Int. Conf. on Robotics and Automation*, 2014.
- [6] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtm: Dense tracking and mapping in real-time," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2320–2327.
- [7] J. Engel, J. Sturm, and D. Cremers, "Semi-dense visual odometry for a monocular camera," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1449–1456.
- [8] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in *Proc. IEEE Intl. Conf. on Robotics and Automation*, 2014.
- [9] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *Medical Imaging, IEEE Transactions on*, vol. 16, no. 2, pp. 187–198, 1997.
- [10] A. Dame and E. Marchand, "Entropy-based visual servoing," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 707–713.
- [11] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.
- [12] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International journal of computer vision*, vol. 24, no. 2, pp. 137–154, 1997.
- [13] A. Aldroubi, M. Unser, and M. Eden, "B-spline signal processing," *IEEE Transactions on Signal Processing*, vol. 41, pp. 821–849, 1993.
- [14] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in *Numerical analysis*. Springer, 1978, pp. 105–116.
- [15] P. J. Huber, *Robust statistics*. Springer, 2011.

Building a grid-point cloud-semantic map based on graph for the navigation of intelligent wheelchair

Cheng ZHAO*, Huosheng HU, Dongbing GU

School of Computer Science and Electrical Engineering, University of Essex CO4 3SQ, Colchester, UK

Email: IRobotCheng@gmail.com, hhu@essex.ac.uk, dgu@essex.ac.uk

Abstract—The automatic navigation of the intelligent wheelchair is a major challenge for its applications. Most previous researches mainly focus on 2D navigation of intelligent wheelchair, which loses many useful environment information. This paper proposed a novel Grid-Point Cloud-Semantic Multi-layered Map based on graph optimization for intelligent wheelchair navigation. For mapping, the 2D grid map is at the bottom, the 3D point cloud map is on the grid map and the semantic markers are labelled in them. The semantic markers combine the name and coordinate value of object marker together. For navigation, the wheelchair uses the grid map for localization and path planning, uses the point cloud map for feature extraction and obstacle avoidance, and uses the semantic markers for human-robot vocal interaction. A number of experiments are carried out in real environments to verify its feasibility.

Keywords—*Graph-based SLAM, Grid Map, Point Cloud Map, Semantic Labelling, Navigation*

I. INTRODUCTION

Intelligent wheelchair is a very useful assisted living application for the elderly and disabled people, which can decrease the pressure of ageing of population. The availability of automatical navigation of intelligent wheelchair is necessary in hospital, office and home. In the past decade, there are many notable successes for the simultaneous localization and mapping (SLAM) in robotics community. There are four mainly kinds of typical map now: grid map, point cloud map, topological map and semantic map. There are many navigation researches that focus on different type of maps.

The outline of grid and point cloud mapping is similar and it can be classified as filtering, smoothing and graph-based approaches. The state of filtering approach consists in current robot position and map, which can perform an online state estimation. When the new measurement is available, the estimate is augmented. The Kalman [1][2], particle [3][4], information [5][6] filter are very popular online SLAM methods. In term of smoothing approach [7][8] that relies on least-square error minimization estimate the whole trajectory of robot through the whole set of measurements. Recently, the graph-based SLAM [9][10][11] achieved many breakthroughs, which constructs a graph out of the raw sensor measurements. The graph consists of the nodes each of which represents the robot position with the measurements from this position and edges each of which represents the spatial constraint between two robot poses. The constraint is actually the transformation relating two robot poses and it can be calculated through odometry measurements or aligning the observations. After constructing the graph, the next step is graph optimization: calculating a most likely configuration that best satisfies the

constraints. However online processing for global loop closure detection is difficult in large-scale and long-term environment, the work in [12] proposed a graph-based memory management approach that only considers portions of map satisfy online processing requirements. The only difference between grid and point cloud mapping is that replacing the 2D scan matching and loop closure detecting with counterparts that operate 3D point cloud.

The work in [13] proposed a hybrid map that integrates grid and topological map for indoor navigation. The work in [14] proposed an approach to segment and detect the room spaces for navigation using range data in office environment. Using the anchoring technique, [15] labels the topological map with semantic information for navigation. The works in [16][17] introduced an approach to learn topological maps from geometric map by applying semantic classification procedure based on associative Markov networks and AdaBoost. Many work used visual features from camera sensor like [18] to extract semantic information via place categorization. The paper [19] integrated a robust image labelling method with a dense 3D semantic map. In addition, many work added rich semantic information to the map through human-robot interaction for navigation. In [20], a contextual topological map was built through making the robot follow the user and verbal commentary. The work in [21] built the semantic map through clarification natural language dialogues between human and robot. The system in [22] integrated laser and vision sensors for place and object recognition as well as a linguistic framework, which create conceptual representation of human-made indoor environment. The work in [23] summarized many multi-modal interactions such as speech, gesture and vision for semantic labelling, which can make the robot get rich environment knowledge without many pre-requisites features.

Different types of map have their own advantages and disadvantages for navigation. Grid map is easy to make path planning but it does not have enough 3D features and semantic information. Point cloud map is easy to get enough features information but it is hard to use for path planning because of high complexity. Semantic map can provide the knowledge of environment for human-robot interaction but the robot can not perform any action based on it. It is necessary to combine different type of map together in order to make the navigation of robot more efficiently. This paper proposes a grid-point cloud-semantic multi-layered map for the navigation of intelligent wheelchair and carries out some related experiments in real environment. The rest of this paper is organized as follows. Section 2 introduces the intelligent wheelchair platform and outlines the basic system workflow. Section 3 outlines the

approach deployed in this research. In section 4, some related experiment results are presented to verify the performance of the proposed approach. Finally, section 5 makes a brief conclusion and future work.

II. SYSTEM OVERVIEW

A. Robot Platform

The platform used in this work is a commercial intelligent wheelchair Spectra Plus equipped with a Kinect which provides RGB and depth information, a Hokuyo URG-04LX-UG01 Laser which provides 2D laser scan information and optical encoders which provide odometry transformation information. Computation is performed on an on board PC (Intel Pentium(R) 2.30 GHz, 8G RAM and Integrated Graphics) installed with Ubuntu 14.04 and ROS Indigo. The platform snapshot is shown in Figure 1.

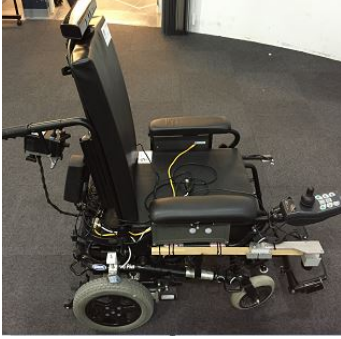


Fig. 1: The platform snapshot.

B. Architecture Overview

The mapping module can be broken up into three modules: Frontend, Backend and Map creation. The frontend mainly obtains the sensor data, extracts the features and calculates the geometric relationships between robot and landmarks. The backend mainly constructs a graph, performs loop closure detection and optimizes the graph structure. Finally, this module outputs a maximum likelihood solution of the robot trajectory. In addition, it also performs a model matching through comparing the object database with the features from RGB frames. If the model matching is successful, the mass centre position of the object will be calculated through the depth images. This position is combined with the object name together in a semantic marker. In terms of map creation, the 2D grid map and 3D point map are generated based on the trajectory. Then the semantic markers are labelled to them. Combining them together, the multi-layered map is finally generated.

The navigation module consists of localization, getting a destination, path planning. The wheelchair can perform the localization based on the trajectory from graph SLAM. Then the user tells the wheelchair the destination name using vocal interface. The wheelchair can get the position of destination from semantic markers. Finally, the wheelchair performs path planning using grid map and obstacle avoidance using point cloud map. The architecture overview is shown in Figure 2.

III. APPROACHES

A. Pose graph outline

The map of graph-based is a graph consisting of nodes and edges like $G = \{V, E\}$. The nodes [12] contain the odometry poses, laser scans, RGB and depth images of each location. In addition, they also save visual words which are used for loop closure detection. The edges are made of neighbors and loop closures, which save the geometrical transformation between nodes. When the odometry transformation between the current and previous nodes is generated, the neighbor edges are added to the graph. When a loop closure detection is found between the current node and one of the previous maps, the loop closure edges are added to the graph. After the construction of graph, the graph optimization is performed which calculates the most likely configuration that best satisfies the constraints of edges.

Suppose $x = (x_1 \dots x_2)^T$ is a vector of parameters which represent the configuration of nodes. ω_{ij} and Ω_{ij} represent the mean and the information matrix of the observation of node j and i respectively. For the state x , $f_{ij}(x)$ is the function which can calculate the observation of the current state. The residual r_{ij} can be calculated via equation (1).

$$r_{ij}(x) = \omega_{ij} - f_{ij}(x) \quad (1)$$

The amount of error introduced by constraints which are calculated by real or visual odometry, weighed by its information, can be obtained via equation (2).

$$d_{ij}(x)^2 = r_{ij}(x)^T \Omega_{ij} r_{ij}(x) \quad (2)$$

Assuming all the constraints to be independent, the overall error can be calculated through equation (3).

$$D^2(x) = \sum_{(ij) \in \phi} d_{ij}(x)^2 = \sum_{(ij) \in \phi} r_{ij}(x)^T \Omega_{ij} r_{ij}(x) \quad (3)$$

where $d_{ij}(x)^2$ is residual of the edge which connects node i and j in graph ϕ . So the key of graph-based SLAM is to find a state x^* that minimizes the overall error.

$$x^* = \underset{x}{\operatorname{argmin}} \sum_{(ij) \in \phi} r_{ij}(x)^T \Omega_{ij} r_{ij}(x) \quad (4)$$

Compactly, it is rewritten in equation (5)

$$x^* = \underset{x}{\operatorname{argmin}} \sum_{(ij) \in \phi} \|r_{ij}(x)\|_{\sum_{ij}}^2 \quad (5)$$

where $\sum_{ij} = \Omega_{ij}^{-1}$ is the corresponding covariance matrix of the information matrix.

B. Loop closure detection

The approach of loop closure detection [12][24] is used in this work. They use a Bayesian filter to evaluate loop closure hypotheses over all previous images. If the hypothesis reaches the pre-defined threshold H , a loop closure is detected. In order to compute the likelihood required by the filter, the ORB features are used for an incremental visual words considering the low computing power of the on board PC. If you have a powerful on board PC, the best choice is SIFT feature with GPU. For every 2D point from RGB image where the visual words are extracted, the relating 3D position can be calculated via calibration matrix and the depth image. So, every visual

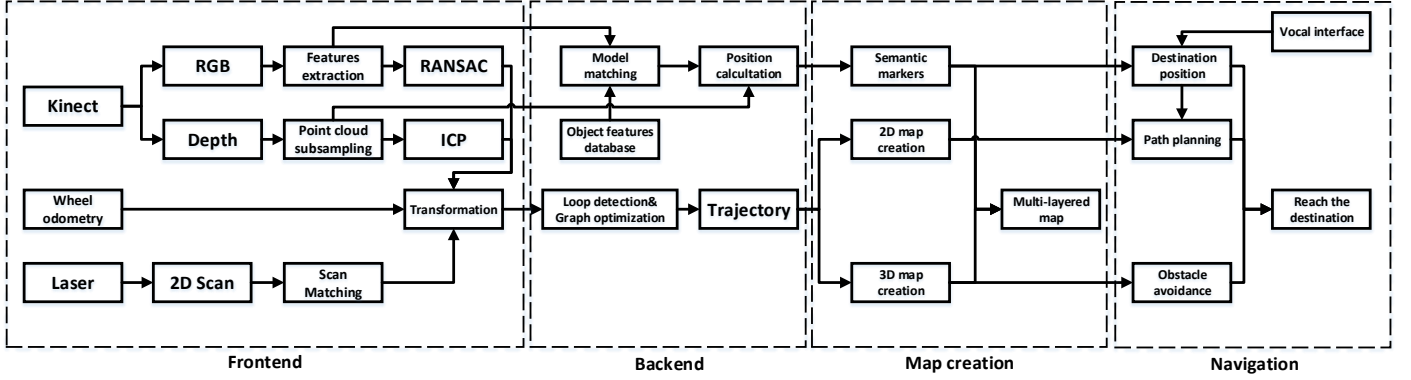


Fig. 2: Architecture overview.

words' 3D position can be obtained. Using the 3D visual word correspondences, every transformation between the matching images is computed via RANSAC if a loop closure is found. When the minimum of I inliers is found, the loop closure is successful and the related edge is added to the graph.

C. Graph optimization

For graph optimization, the G2O [25][26] framework is used in our work. This approach performs a minimization of non-linear error function that can be represented as a graph. G2O can minimize the error in a graph in which vertices are parameterized by transformation and edges represent constraints between vertices with associated covariance matrices. It maximizes the likelihood of the vertex parameters subject to the constraints using stochastic gradient descent. When the robot revisits a known part of the map, the loop closure edges in the graph can diminish the accumulated error introduced by odometry.

D. Map representation

Based on the trajectory outputted from graph-based SLAM, the laser scans are assembled together and the 2D grid map is generated at the bottom, meanwhile the RGB and depth images are assembled together and the 3D point cloud map is generated on the grid map. In terms of semantic labelling, firstly a database of objects' features is built. Then a model matching is performed comparing the extracted features from RGB images and the features in object database. If a model matching is successful, the centre of mass position of the found object can be calculated through corresponding depth image. This position is combined with object name in a semantic marker. Due to the low computing power of the on board PC, the ORB feature is selected as the feature extraction and description. Lastly, the semantic markers are labelled in the 2D and 3D map.

E. Navigation

"Where am I", "Where shall I go" and "How to get there" are the three central issues of navigation. For the first issue, the intelligent wheelchair can get its position through the trajectory

outputted from graph-based SLAM. For the second issue, the user can tell wheelchair the destination name using vocal interface through bluetooth earphone. The open source speech recognition toolkit Pocket Sphinx [27] is used for human-robot vocal interaction, which is easy to use for training (just upload the semantic text to the tool website). The voice will be transformed to text and then a string matching is performed. Finally the wheelchair can find the position of corresponding destination name from semantic markers. For the third issue, the wheelchair transforms the grid map to a cost map [28][29], and then uses Dijkstra algorithm to search the cost map to find a route that has minimal sum of the cost value. For obstacle avoidance, the obstacles and ground are segmented by normal filtering using point cloud map. All points with normal in the z direction are labelled as ground and all the others are labelled as obstacles. Lastly, movebase package [30] in ROS is used for the motion planning of intelligent wheelchair.

IV. EXPERIMENTS

A. 2D grid map

As shown in the Figure 3, the 2D grid map is generated using laser scans. The blue line represents the trajectory of the wheelchair and the small blue point represents the graph node. The yellow line connected with two nodes represents that there is a loop closure between them. There are many loop closures in A, B, C because the wheelchair rotated by 360 degrees there, while there are also many loop closures in D, E, F because the wheelchair moved with a loop trajectory.

B. 3D point cloud map

As shown in the Figure 5, the 3D point cloud map is generated using RGB-D information. The whole point cloud map are made of a large number of RGB-D frames obtained from every node in the graph. Those frames are aligned together through the transformation constrains.

C. Semantic labelling

The object database stores the ORB features of Baxter, Fox robot, Robot fish, Fly robot and Robot dog. By extracting the

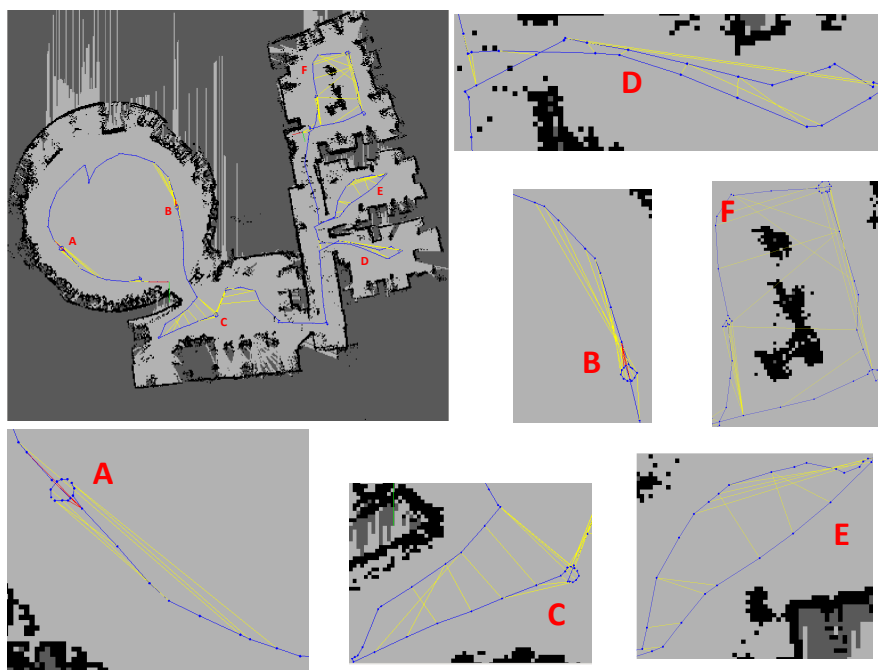


Fig. 3: 2D grid map.

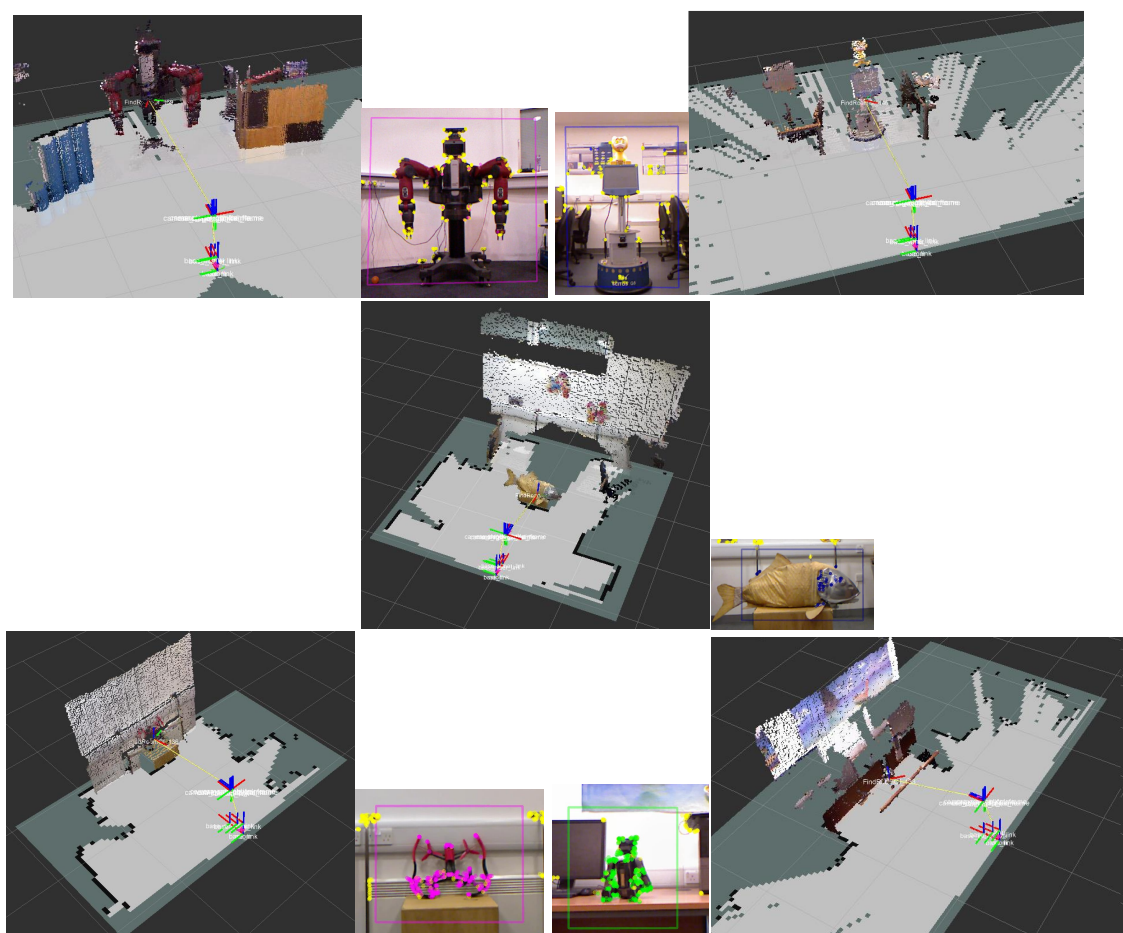


Fig. 4: 3D semantic labelling.

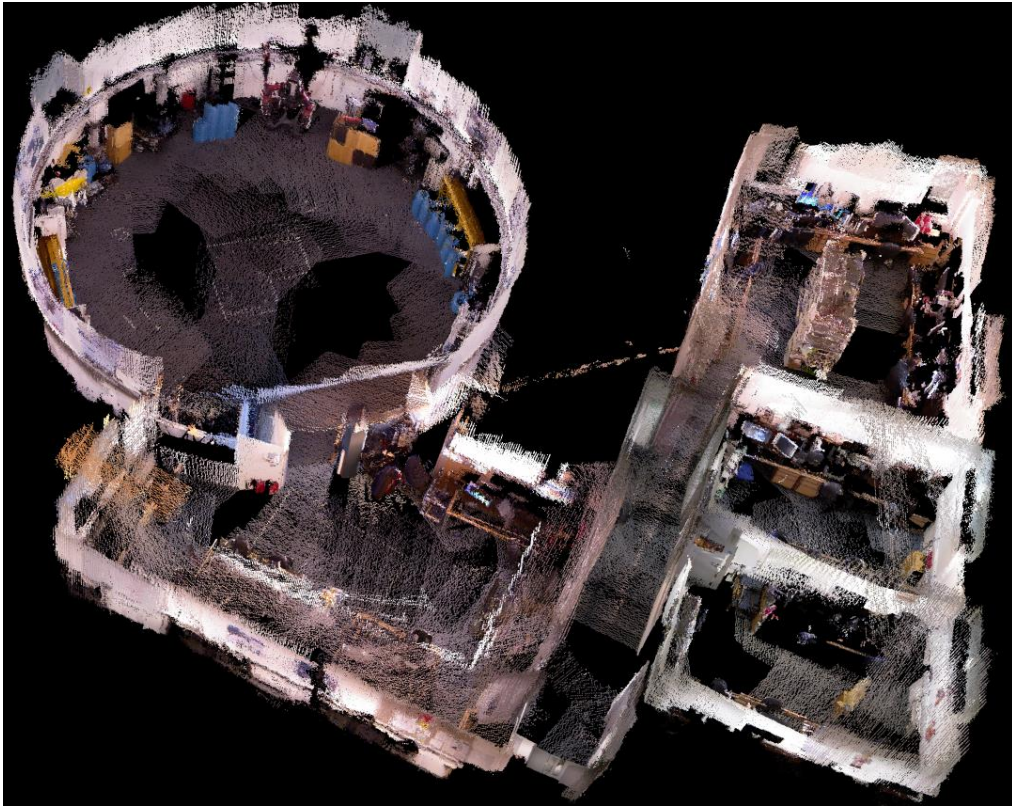


Fig. 5: 3D point cloud map.

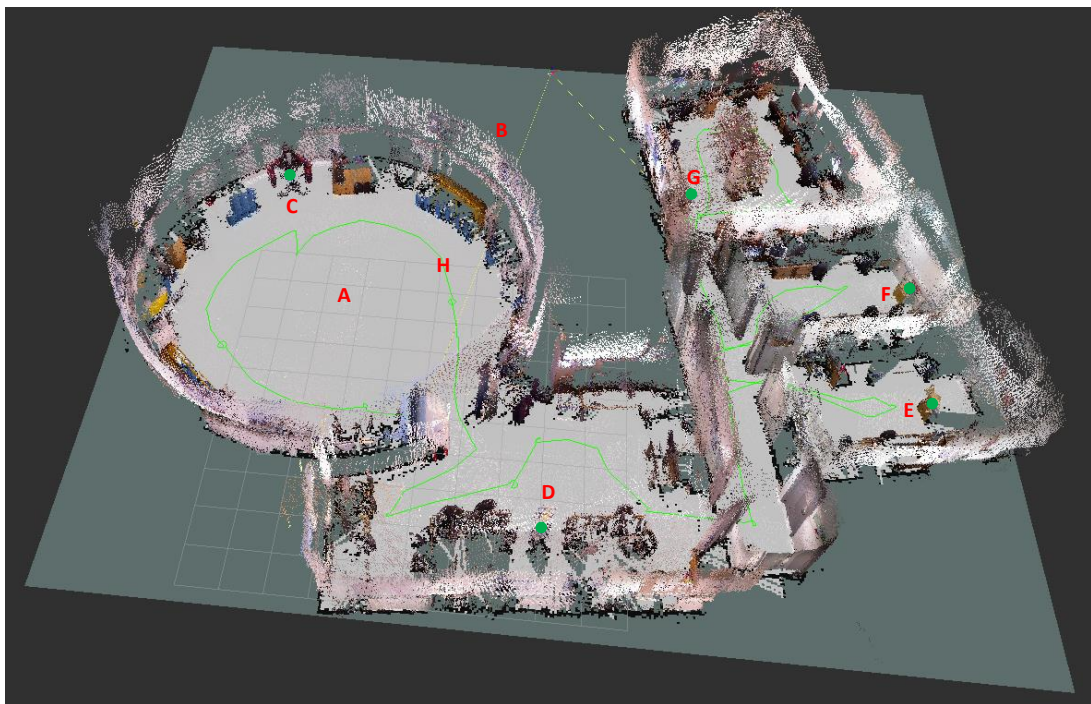


Fig. 6: Multi-layered map.

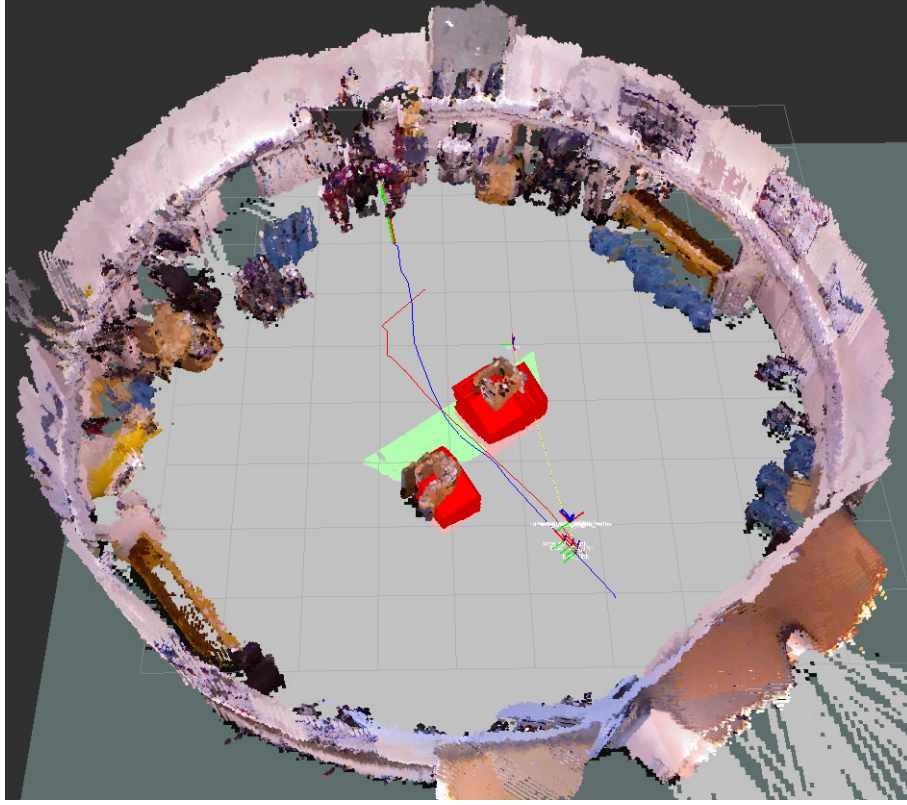


Fig. 7: Navigation in hybrid map.

ORB features from RGB images, these robots were found in the process of mapping, as show in the colorized box from Figure 4. The positions of their centre mass were calculated using the depth image. As shown in Figure 4, there is a yellow line connecting the wheelchair and the object marker. Finally, those semantic markers (X, Y, Z) will be labelled in the 3D point cloud map.

D. Multi-layered map

As shown in the Figure 6, the multi-layered map is generated: the grid map A is at the bottom, the point cloud map B is on the grid map and the semantic markers C, D, E, F, G are labelled as green points in the point cloud map. The green line represents the trajectory of the wheelchair.

E. Navigation

As shown in the Figure 7, in the point cloud map, the obstacles are inflated and labelled as red colour, the ground is labelled as green colour. The blue and red line on the grid map are the outcomes of global and local path planning respectively. All the mapping and navigation videos of experiments can be found <https://www.youtube.com/channel/UC-TI5R4MBKIXsGI0sFIJuXQ>

V. CONCLUSION

As there are some obviously drawbacks when the intelligent wheelchair performs navigation only using one single

type of map, in this paper a Grid-Point Cloud-Semantic Multi-layered map based on graph optimization for the navigation is proposed. The grid map at the bottom is used for localization and path planning, the point cloud map on the grid map is used for features extraction and obstacles avoidance, the semantic markers is used for human-robot vocal interaction. This hybrid map can make intelligent wheelchair perform navigation task more efficiently. In the future, our work will mainly focus on adding more rich semantic information to this hybrid map through 3D labelling.

ACKNOWLEDGEMENT

The 1st author is financially supported by scholarships from China Scholarship Council and University of Essex, U.K.

REFERENCES

- [1] C. P. Smith R, Self M, *Estimating Uncertain Spatial Relationships in Robotics*. Springer New York, 1990.
- [2] M. W. M. Gamin Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte, and M. Csorba, "A solution to the simultaneous localization and map building (SLAM) problem," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 229–241, 2001.
- [3] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM: A factored solution to the simultaneous localization and mapping problem," in *Proc. of 8th National Conference on Artificial Intelligence/14th Conference on Innovative Applications of Artificial Intelligence*, vol. 68, no. 2, 2002, pp. 593–598.
- [4] G. Grisetti, C. Stachniss, and W. Burgard, "Improved techniques for grid mapping with Rao-Blackwellized particle filters," *IEEE Transactions on Robotics*, vol. 23, pp. 34–46, 2007.

- [5] R. M. Eustice, H. Singh, and J. J. Leonard, "Exactly sparse delayed-state filters," in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*. IEEE, 2005, pp. 2417–2424.
- [6] S. Thrun, Y. Liu, D. Koller, A. Ng, Z. Ghahramani, and H. Durrant-Whyte, "Simultaneous Localization and Mapping with Sparse Extended Information Filters," vol. 23, no. 7, pp. 693–716, 2003.
- [7] F. Dellaert and M. Kaess, "Square Root SAM: Simultaneous Localization and Mapping via Square Root Information Smoothing," *The International Journal of Robotics Research*, vol. 25, no. 12, pp. 1181–1203, 2006.
- [8] E. Olson, J. Leonard, and S. Teller, "Fast iterative alignment of pose graphs with poor initial estimates," in *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2006, no. May, 2006, pp. 2262–2269.
- [9] S. Thrun, "The Graph SLAM Algorithm with Applications to Large-Scale Mapping of Urban Structures," *The International Journal of Robotics Research*, vol. 25, no. 5-6, pp. 403–429, 2006.
- [10] H. Strasdat, J. M. M. Montiel, and A. J. Davison, "Visual SLAM: Why filter?" *Image and Vision Computing*, vol. 30, no. 2, pp. 65–77, 2012.
- [11] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard, "3-D Mapping with an RGB-D camera," *IEEE Transactions on Robotics*, vol. 30, no. 1, pp. 177–187, 2014.
- [12] M. Labbe and F. Michaud, "Online global loop closure detection for large-scale multi-session graph-based SLAM," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 2661–2666.
- [13] S. Thrun and A. Bücken, "Integrating Grid-Based and Topological Maps for Mobile Robot Navigation," *Proceedings Of The National Conference On Artificial Intelligence*, vol. 13, no. August, pp. 944–950, 1996.
- [14] P. Buschka and A. Saffiotti, "A virtual sensor for room detection," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 1, no. October, pp. 637–642, 2002.
- [15] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. a. Fernández-Madriral, and J. González, "Multi-hierarchical semantic maps for mobile robotics," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2005*, pp. 3492–3497.
- [16] O. Mozos and W. Burgard, "Supervised Learning of Topological Maps using Semantic Information Extracted from Range Data," *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2772–2777, 2006.
- [17] E. Brunskill, T. Kollar, and N. Roy, "Topological mapping using spectral clustering and classification," *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3491–3496, 2007.
- [18] J. Wu, H. I. Christensen, and J. M. Rehg, "Visual place categorization: Problem, dataset, and algorithm," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009, 2009*, pp. 4763–4770.
- [19] Z. Zhao and X. Chen, "Semantic mapping for object category and structural class," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, no. Iros, 2014, pp. 724–729.
- [20] A. Diosi, G. Taylor, and L. Kleeman, "Interactive SLAM using laser and advanced sonar," in *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2005, no. April, 2005, pp. 1103–1108.
- [21] G.-J. M. Kruijff, H. Zender, P. Jensfelt, and H. I. Christensen, "Clarification dialogues in human-augmented mapping," *Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction - HRI '06*, p. 282, 2006.
- [22] H. Zender, O. Martínez Mozos, P. Jensfelt, G. J. M. Kruijff, and W. Burgard, "Conceptual spatial representations for indoor mobile robots," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 493–502, 2008.
- [23] G. Randelli, T. M. Bonanni, L. Iocchi, and D. Nardi, "Knowledge acquisition through human-robot multimodal interaction," *Intelligent Service Robotics*, vol. 6, no. 1, pp. 19–31, 2013.
- [24] M. Labbe and F. Michaud, "Appearance-based loop closure detection for online large-scale and long-term operation," *IEEE Transactions on Robotics*, vol. 29, pp. 734–745, 2013.
- [25] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "G2o: A general framework for graph optimization," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2011, pp. 3607–3613.
- [26] G. Grisetti, S. Grzonka, C. Stachniss, P. Pfaff, and W. Burgard, "Efficient estimation of accurate maximum likelihood maps in 3D," in *IEEE International Conference on Intelligent Robots and Systems*, no. MI, 2007, pp. 3472–3478.
- [27] D. Huggins-Daines, M. Kumar, A. Chan, A. Black, M. Ravishankar, and A. Rudnicky, "Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices," *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, pp. 185–188, 2006.
- [28] J. King and M. Likhachev, "Efficient cost computation in cost map planning for non-circular robots," in *Intelligent Robots and Systems*. St.Louis,MO: Ieee, Oct. 2009, pp. 3924–3930.
- [29] E. Marder-Eppstein, E. Berger, T. Foote, B. Gerkey, and K. Konolige, "The Office Marathon: Robust navigation in an indoor office environment," in *IEEE International Conference on Robotics and Automation*. Anchorage,AK: Ieee, May 2010, pp. 300–307.
- [30] B. P. Gerkey and K. Konolige, "Planning and Control in Unstructured Terrain," in *ICRA Workshop on Path Planning on Costmaps*, 2008.

Geometrical and statistical feature extraction of images for rotation invariant classification systems based on industrial devices

Rodrigo D. C. Silva, George A. P. Thé, Fátima N. S. de Medeiros
Depto. de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
Email: george.the@ufc.br

Abstract—In this work, the problem of recognition of objects using images extracted from a 3D industrial sensor is discussed. We focus in 7 feature extractors based on invariant moments and 2 based on independent component analysis, as well as on 3 classifiers (k-Nearest Neighbor, Support Vector Machine and Artificial Neural Network-Multi-Layer Perceptron). To choose the best feature extractor, their performance was compared in terms of classification accuracy rate and extraction time by the k-nearest neighbors classifier using euclidean distance. For what concerns the feature extraction, descriptors based on sorted-Independent Component Analysis and on Zernike moments performed better, leading to accuracy rates over 90.00 % and requiring relatively low time feature extraction (about half-second), whereas among the different classifiers used in the experiments, the support vector machine outperformed when the Zernike moments were adopted as feature descriptor.

Keywords—Invariant moments, Independent Component Analysis, Support Vector Machine, Multi-Layer Perceptron

I. INTRODUCTION

Automatic visual recognition of objects and people in a scene is a hot research topic worldwide, with many important applications being found in industry (e.g., counting, inspection and quality control), security (e.g., surveillance systems), urban environment (autonomous navigation, collision and accidents avoidance, traffic monitoring), etc [1].

In the industrial automation field, machine vision provides innovative solutions and helps improving efficiency, productivity and quality management, bringing competitiveness for solution providers [2]. As examples of industrial activities which have benefited from the application of machine vision technology on manufacturing processes, it can be cited [3] and references therein.

In general, it is desirable or required for the system to be able to work regardless of translation, rotation or scale transformations, what means achieving good performance in non-structured scenarios, though it may occur at the expense of data processing costs. Goal is, therefore, finding simple and efficient technical solutions.

Thanks to the recent developments in data acquisition, processing, and process control systems, efficiency of many industrial applications has been improved with the help of

automated visual processing and classification systems [4]. In this scenario, the classification between different objects (i.e., the ability of label assignment) is usually a complex problem for machines, because it involves many steps, e.g. image acquisition, preprocessing, feature extraction and classification itself. So, although the offer of shop floor equipments has increased a lot (in number and in performance) in the last decades, the majority of industrial computers still has processing and data storage limitations.

On the other side, high-performance computers, nowadays easily available commercially, allows for solving the complexity of the mathematics involved in those mentioned steps. Bridging this gap, the evolution of industrial data networks and integrated communication solutions made possible image-based supervision and on-line control in industry, provided that costly data processing be forwarded to remote stations.

To approach the problem of feature extraction in 2D object recognition, there are some traditional methods which rely mainly on the calculation of invariant moments (Hu, Zernike, Legendre, etc). One important property of the moments is their invariance under affine transformation. Moments are scalar quantities used to characterize a function and to capture its significant features. From the mathematical point of view, moments are projections of a function onto a polynomial basis [5].

More recently, a blind source separation technique named independent component analysis (ICA) has been used in many fields, from electrical power systems [6] to economic modeling [7], since it offers good feature description ability from a reduced set of descriptors [8]. In essence, it represents a given measurement (an image, a speech signal or whatever) as a linear composition of statistically independent components; one could therefore, use the independent components themselves or the coefficients of the linear composition as features of the input raw data.

In this context, in this work the issue of automatic inspection and supervised image classification is considered in both public image datasets and low-resolution images extracted from an industrial sensor. Primary goal is to evaluate the

performance of many feature descriptors based on invariant geometrical moments, comparing with feature extraction from independent component analysis. We also aim at investigating different classifiers, in order to point the best solution for image classification.

The paper is organized as follows: in Section II several approaches for feature extraction are described, whereas in Section III the classifiers are briefly reviewed. In the following, in Section IV the experimental setup and the datasets are described. Finally, results and discussion appear in Section V

II. FEATURE EXTRACTION

According to the literature, feature extraction is the problem of getting, from raw data, relevant information for classification purposes, thus achieving minimal within-class pattern variability while enhancing discrimination between classes. This is accomplished representing each image by a vector containing a set of features. In this section several feature descriptors based on geometrical moments, as well as that based on independent component analysis will be reviewed.

A. Hu moments

In 1962 Hu [9] introduced the concept of moment giving rise to the use of invariant moments and moment functions in the fields of image analysis and pattern recognition. In that seminal paper, seven nonlinear functions which are translation, scale and rotation invariant were introduced for computing the center of mass of a given image, and that of a certain region (in case of a binary mask). As pointed out in [10], the drawback of such approach is that the kernel function of geometric moments of order $(p+q)$ is not orthogonal, leading the geometric moments to suffer from information redundancy, as well as from noise sensitivity for higher-order moments.

B. Zernike moments

The issue of information redundancy can be overcome by the use of orthogonal moments for image representation. This has been known since the early 80's [11] and the Zernike moments were one of the foremost approaches adopted (along with Legendre moments, discussed next). In the discrete form, mn -th order moment is written as:

$$Z_{mn} = \frac{n+1}{\lambda_N} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} V_{nm}(x, y) I_{xy}, \quad (1)$$

where n is the order of the radial polynomial, λ_N is a normalization factor accounting for the amount of pixels inside the unit circle where the basis function, V_{nm} is evaluated and I_{xy} is the image matrix. For better description, refer to [11].

C. Legendre moments

It consists in a recursive relation of the p -th order Legendre polynomial, and takes the following discrete form:

$$L_{pq} = \lambda_{pq} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_p(x_i) P_q(y_j) I_{ij}. \quad (2)$$

It is calculated in the interval $[-1, 1]$ and, therefore, the pixels are scaled to the $-1 < x, y < 1$ region, giving rise to the x_i, y_i normalized coordinates. Again, λ_{pq} is a normalization constant and I_{ij} represents the $N \times N$ intensity image matrix. Finally, P_q and P_p are Legendre polynomials. For additional details, refer to [11].

D. Fourier-Mellin moments

This is another class of rotation-invariant orthogonal moments, which in the discrete form is calculated as:

$$O_{pq} = \frac{p+1}{\pi} \sum_{i=0}^{N-1} \sum_{k=0}^{N-1} I(x_i, y_k) M_{pq}^*(x_i, y_k) \Delta x_i \Delta y_k \quad (3)$$

where x_i and y_k are normalized coordinates of pixels in the region where polynomials are evaluated, M_{pq}^* is the complex conjugate of the orthogonal polynomials and $I(x_i, y_k)$ is the N -sized intensity image matrix.

E. Tchebichef moments

In contrast to the moments described so far, which rely on the definition of a region-of-interest for the evaluation of the discrete form of a given function, Tchebichef moments as introduced by [12] do not require numerical approximations, since the basis functions set is orthogonal in the discrete domain of the image coordinates. These moments are obtained as:

$$T_{pq} = \frac{1}{\rho(p, N) \rho(q, N)} \sum_{i=0}^{N-1} \sum_{k=0}^{N-1} t_p(x) t_q(y) I_{xy}, \quad (4)$$

where p and q define the order of the polynomial, ρ is a normalization function dependent on the moment order and on image size, whereas t_p and t_q are Tchebichef polynomials. For further details, refer to [12].

F. Bessel-Fourier moments

It is a more recent class of polar-coordinates based orthogonal moments, introduced in [13]. These are written as:

$$B_{nm} = \frac{1}{2\pi a_n} \sum_k b_{nk} C_{pq}. \quad (5)$$

In this expression, a_n is a normalization constant, b_{nk} is a term accounting for the zeroes of the first-order Bessel function and C_{pq} contains the complex moments dependent on the image matrix of interest. Additional details can be found in [13].

G. Gaussian-Hermite moments

Although it is not new as a class of continuous orthogonal moments, recently attention has been given in the context of rotation and translation invariance [14]. They are defined as $M = H I H^T$, where I is the image matrix, superscript T denotes transpose matrix and H represents the Gaussian-Hermite polynomials having σ scale factor:

$$H(x; \sigma) = \frac{e^{-\frac{x^2}{2\sigma^2}} H_p(\frac{x}{\sigma})}{\sqrt{2^p p! \sqrt{\pi} \sigma}}. \quad (6)$$

In the above equation, H_p is the Hermite polynomial.

H. Independent component analysis

According to [15], it is a statistical signal processing technique whose goal is to linearly decompose a random vector into components which are as independent as possible. The basic definition of ICA considers a set of observations of random variables $x_1(t), x_2(t), \dots, x_n(t)$ and the assumption that they are generated as a linear mixture of independent components $s_1(t), s_2(t), \dots, s_n(t)$, according to

$$x = A(s_1(t), s_2(t), \dots, s_n(t))^T = As, \quad (7)$$

where A is the unknown mixture matrix. For a better explanation on ICA model and underlying requirements, refer to [8]. ICA applications on pattern recognition of rotated images require as training step the random variables to be the training images. Letting x_i to be a vectorized image, we can construct a training image set x_1, x_2, \dots, x_n , with n random variables which are assumed to be the linear combination of the m unknown independent components s , in such a way that the coefficients are given by the elements of the mixture matrix. When ICA is applied to feature extraction, the columns of A_{train} contain the main feature vectors of the training images, being used therefore as input to the classifier along with the mixture matrix of the image under test, A_{test} .

According to the literature, the efficiency of the ICA algorithm is very dependent on preprocessing steps [8]. Indeed, in applications where rotation invariance is a requirement, we have observed and recently proposed to perform an ordering transformation of the input vectorized images to achieve better ICA feature extraction. This work is under review, but the technique will be applied to the analysis of the present manuscript for completeness and for comparison to the other feature descriptors. From now on, it will be referred to as *ICA_{sort}*.

III. CLASSIFICATION

After completing the feature extraction, the final stage of any image processing system contains the classification step, in which each sample is labeled or assigned to a new or existent class; in this step, the better the data representation provided by the feature descriptor, the better the assignment step will be. However, also the influence of the classifier itself to the classification efficiency plays a role (efficiency here understood as a measure of its ability to distinguish the interclass similarity whereas bypassing eventual intraclass differences). To accomplish with that, we chose 3 different classification approaches, which will be described next.

A. *k*-Nearest neighbors classifier

The *k*-Nearest neighbor is a classifier where the learning is based in analogy. The training samples are formed by n -dimensional vectors, and each element of this group is a point in n -dimensional space.

To determine the class of an element which does not belong to the training set, the *k*-NN classifier searches for k elements of the training set that are closest to this unknown element, i.e. those whose separation correspond to the smallest

distances. These k elements are called *k*-nearest neighbors. In this article, Euclidean distance is used as metric for evaluating the adjacency [16].

B. Artificial Neural Network

Artificial Neural Network (ANN) can be seen as a parallel distribution process, inspired on how biological neurons process information. It is composed of a large number of highly interconnected processing elements (neurons) working to solve specific problems.

For problems that are not linearly separable, it is possible to efficiently train networks built with intermediate layers, the so-called Multilayer Perceptron network (MLP). A typical MLP network has three main features: the neurons of the intermediate layer have a sigmoid-like activation function, the network has one or more intermediate layers and the network has a high degree of connectivity. [17]

C. Support Vector Machine

Support Vector Machine (SVM) is a technique for classification and regression that uses a nonlinear mapping to transform the original training data into a higher dimension where a separation hyperplane is better built.

The main idea consists in getting a hyperplane optimum, i.e. hyperplanes which maximize the margin separating the classes, in order to separate training patterns of different classes by minimizing the number of errors in the training group. However, usually the application data is not linearly separable. Thus, the SVM algorithm transforms the nonlinear input characteristics to a space in which linear methods can be applied, thus transforming the data to a space where they can be linearly separable.

Although the training time of even the fastest SVMs can be extremely slow, they are highly accurate, owing to their ability to model complex nonlinear decision boundaries. They are much less prone to over fitting than other methods. [16], [18]

IV. EXPERIMENTAL SETUP

A. Public database

As mentioned in the Introduction, the comparison among several feature descriptors and classifiers will be made for different datasets; two of them are public database.

The first database is named dataset A and has 77 images obtained from the database of the Ming Hsieh Department of Electrical Engineering of the University of Southern California. Each image was rotated with 5° step, from 0° to 360° , thus forming 73 samples for every image. Those corresponding to 0° have been used for training, and the remaining, for testing. Some samples are shown in Figure 1.

In the second one, named dataset B, we considered a texture database with different patterns. The Brodatz album available in [19] has 112 texture images, which have been resized from 640×640 to 128×128 pixels. Here again we rotated images with 5° step, from 0° to 360° , thus forming 73 samples for every image. Once more, those corresponding to 0° have been

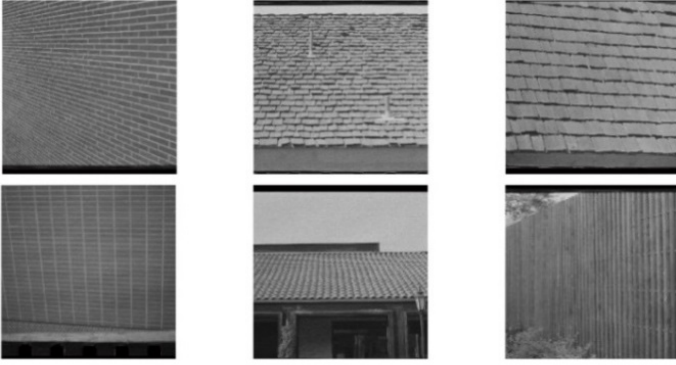


Fig. 1: Samples from University of South California public database.

used for training, and the remaining, for testing. Some samples of this dataset is shown in Figure 2.

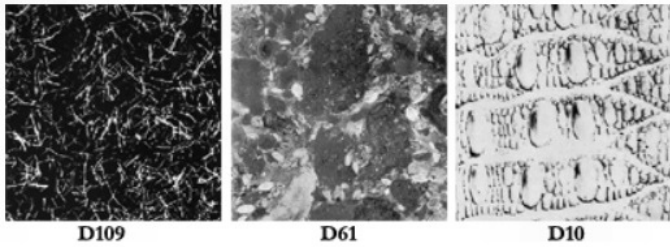


Fig. 2: Samples of Brodatz album.

B. Private database

The third database, dataset C, contains pictures of three small packages, just different in size, which were acquired after randomly rotating the packages on the conveyor belt, until getting 150 samples. This was done in a bad illuminated scenario, as it can be seen in the poor quality of images in Figure 3. The home-made experimental setup for acquisition of dataset C is illustrated in Figure 4.

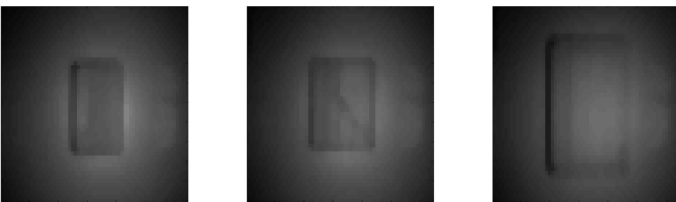


Fig. 3: Pictures of three packages with dimensions 15 x 10.5 x 7.2 cm, 15 x 14 x 6 cm and 21.5 x 16.2 x 9.6 cm, respectively, as acquired from the ifm sensor.

The conveyor is driven by an AC motor with PowerFlex 40P frequency inverter from Allen-Bradley©. This drive is connected to the outputs of the PLC Micrologix 1200©, for remote configuration.

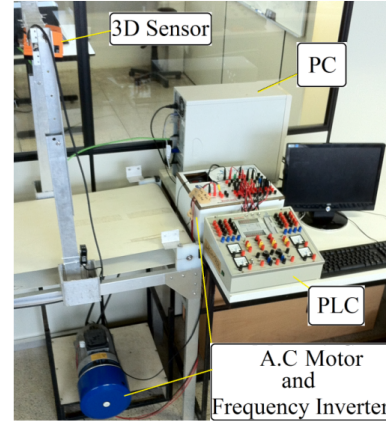


Fig. 4: Experimental apparatus for image acquisition and classification system.

There is also an optical sensor connected to the PLC responsible for triggering the image acquisition whenever an object of interest enters the sensing zone of the industrial camera, 3D effector pmd E3D200 from ifm Electronics©. This is a 50 x 64 resolution camera which uses the time-of-flight (TOF) principle of a matrix of light-emitting diodes to estimate 3D surfaces. It contains Ethernet interface and through remote procedure calls it allows for implementation of real-time applications of classification algorithms remotely.

Communication and data exchange between the running application and the shop floor equipments occurs through Object Linking and Embedding for Process Control Server (OPC) technology. In this solution we use RSLinx Rockwell Automation© to communicate the data managed by the PLC Micrologix 1200©.

V. RESULTS

A. Performance of feature descriptors

To make a fair and clear comparison among the descriptors, initially the feature vectors were presented only to the k-NN classifier using euclidean distance. The classifier was trained and tested 50 times with the same database. The size of the training sets was changed from as low as 10% of the whole available database up to 80%. The experimental results for the dataset C can be seen in the Figure 5 below, whereas those for dataset A and B are plotted in Figures 7 and 6. In these figures, the following legend was adopted: H - Hu Moments, Z - Zernike, L - Legendre, FM - Fourier-Mellin, T - Tchebichef, BF - Bessel-Fourier, GH - Gaussian-Hermite, ICA - Independent Component Analysis and *ICAsort* - ordered Independent Component Analysis (see Section II-H).

Also the time spent for extracting features were evaluated, as well as that for training and running the classifier, and are summarized in Tables I. These results refer to the partition case in which 10% of the dataset was separated for training.

From the table and figures, a number of comments can be made:

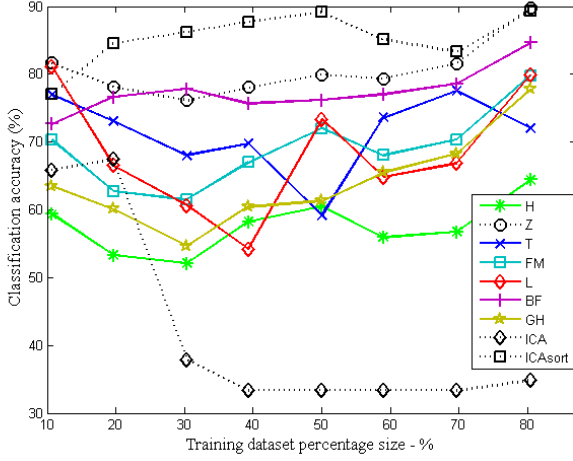


Fig. 5: Average classification accuracy for different feature descriptors, using k-NN and euclidean distance.

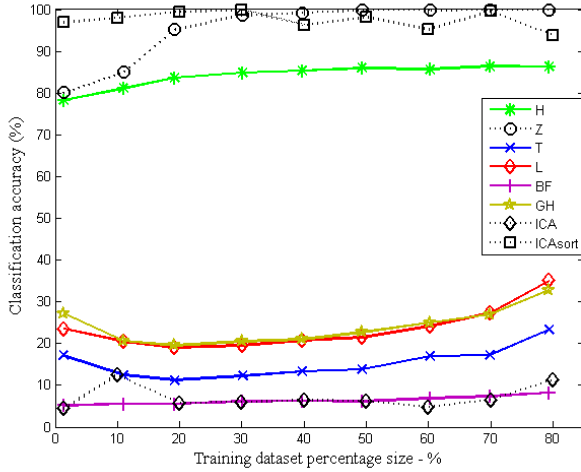


Fig. 6: Average classification accuracy for different feature descriptors, using k-NN and euclidean distance. Obs: Fourier-Mellin omitted for time saving.

a) Zernike moments perform well because they preserve almost all the image information in a few coefficients. The orthogonality of other methods like, for example, the Legendre polynomials has a negative effect when the image is discretized, leading to numerical errors in the calculated moments.

b) Bessel-Fourier has a comparable performance only for the dataset C, which is low-resolution and small database. Furthermore, it seems to be little influenced by the size of the training dataset.

b) Tchebichef moment as feature descriptor leads to irregular classification accuracy, with good performance being achieved in small-sized and large-sized training datasets. However, for datasets A and B it performs poorly.

c) For what concerns the computational efforts issues, we highlight the fact that Fourier-Mellin moments demand more

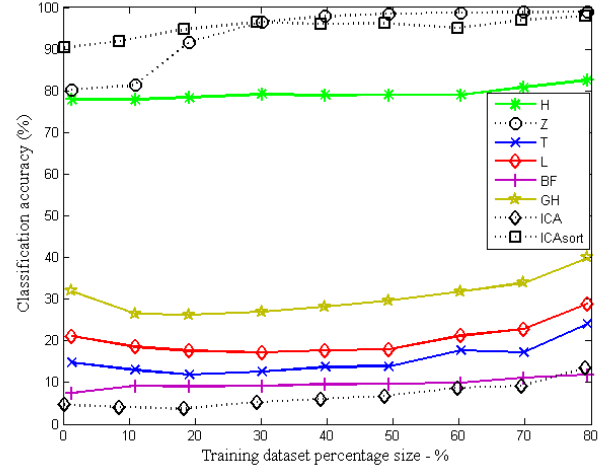


Fig. 7: Average classification accuracy for different feature descriptors, using k-NN and euclidean distance. Obs: Fourier-Mellin omitted for time saving.

TABLE I: Average elapsed times, in seconds, for the various processing steps relative to dataset C.

Dataset C		
Descriptor	Feature Extraction	Training and classify
H	0.7486	0.0013
Z	0.5782	0.0014
L	0.4580	0.0023
FM	18.49	0.0030
T	0.4485	0.0019
BF	0.5114	0.0009
GH	0.4493	0.0008
ICA	0.6558	0.0011
ICAsort	0.4811	0.0009

resources than others. Its huge running time therefore limits the use in any kind of on-line industrial process. This can be assigned to the large number of summations in the equations, usually solved with iterative loops at the computational level.

d) ICA estimation achieved bad performance in all the experiments carried out. We associate this to the fact it is inherently affected by rotation of the images.

e) Finally, an alternative to this limitation is the *ICAsort* approach, which presented top performance when used as feature descriptor of large datasets even for training sets of reduced size.

Partial conclusion is that feature descriptors that are invariant to rotations in the image plane can be easily constructed using Zernike moments (performed good in most scenarios), but *ICAsort* is a promising alternative.

From the results presented so far, the feature extraction based on Zernike moments as well as on ICA and *ICAsort* have been chosen for evaluating the performance of classifiers, shown next.

B. Performance of classifiers

In this study of different approaches for classification, we adopted the following scenario:

a) SVM is implemented with polynomial kernel $d=1$

b) ANN-MLP has sigmoid (tanh) as activation function, and the number of neurons in the hidden and output layers equal 7 and 5, respectively; convergence is ensured by the backpropagation training algorithm.

In Figure 8 we plot the mean accuracy for different combinations of descriptor/classifier as the training dataset size is increased. In this study, only images from dataset C were used. Some comments can be drawn from that figure:

a) using ICA as feature extractor makes the classification very dependent on the training dataset size; a significant sensibility is observed for all the classification approaches.

b) the same comment applies for the *ICAsort* descriptor when SVM or neural networks are used as classifiers. The only good exception is the k-NN approach, which shows nearly regular accuracy in the range of training dataset size studied.

c) Also the feature descriptor based on the Zernike moments seems to be less sensitive to the training set when k-NN is used. For what concerns the SVM and neural network based classifiers, however, there is a positive trend along the training dataset size, and the approach based on Zernike moments + SVM classifier revealed to be superior in the present analysis.

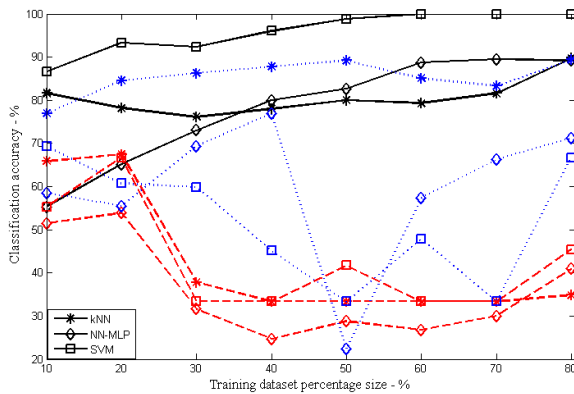


Fig. 8: Performance of classifiers for different alternatives of feature extraction.

VI. CONCLUSION

This paper reported a comparative study between seven invariant moments (Hu, Zernike, Legendre, Fourier-Mellin, Tchebichef, Bessel-Fourier and Gaussian-Hermite) and Independent Component Analysis as feature descriptors of images from different databases. Furthermore, a short comparison between three approaches for classification was made, namely the k-NN classifier, a Neural-Network based classifier and a Support Vector Machine.

The study of the feature extraction step revealed that Zernike moments and *ICAsort* are good candidates for feature description, with a slight advantage of *ICAsort* when k-NN is adopted as classifier.

The study of the classifier, in turn, revealed the superiority of the SVM when the Zernike moments are used as feature descriptor.

To sum up, according to the present study we may state that whenever the feature extraction comes from the *ICAsort* algorithm, the k-NN should be preferred as classifier. Should the feature extraction be performed with Zernike moments, so a Support Vector Machine classifier is recommended instead.

ACKNOWLEDGEMENT

Authors acknowledge CAPES and FUNCAP (PP1-0033-00032.01.00/10) for financial support. Authors also thank Rockwell Automation do Brasil, for the support through the Scientific and Technical Cooperation Agreement with the Federal University of Ceará. Authors finally thank NUTEC, for administrative facilities.

REFERENCES

- [1] E. B. Corrochano, *Handbook of Geometric Computing Applications in Pattern Recognition, Computer Vision, Neural Computing and Robotics*. Heidelberg: Springer, 2010.
- [2] E. N. Malamas, E. G. M. Petrakis, M. Zervakis, L. Petit, and J. D. Legat, "A survey on industrial vision systems, applications and tools," *Image and Vision Computing*, vol. 21, 2003.
- [3] D. Sankowski and J. Nowakowski, *Computer Vision in Robotics and Industrial Applications*. Singapore: World Scientific, 2014, vol. 3.
- [4] M. A. Selver, O. Akay, F. Alim, S. Bardakci, and M. Olmez, "An automated industrial conveyor belt system using image processing and hierarchical clustering for classifying marble slabs," *Robotics and Computer-Integrated Manufacturing*, vol. 27, 2011.
- [5] J. Flusser, T. Suk, and B. Zitová, *Moments and Moment Invariants in Pattern Recognition*. Chichester: John Wiley & Sons.
- [6] M. A. A. Lima, A. S. Cerqueira, D. V. Coury, and C. A. Duque, "A novel method for power quality multiple disturbance decomposition based on independent component analysis," *International Journal of Electrical Power & Energy Systems*, vol. 42, no. 1, pp. 593–604, 2012.
- [7] T.-Y. Lin and S.-H. Chiu, "Using independent component analysis and network dea to improve bank performance evaluation," *Economic Modelling*, vol. 32, pp. 608–616, 2013.
- [8] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Components Analysis*. Canada: John Wiley & Sons, Inc., 2001.
- [9] M. K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on information theory*, 1962.
- [10] D. Sridhar and I. V. M. Krishna, "Face recognition using tchebichef moments," *International Journal of Information & Network Security*, vol. 4, pp. 243–254, 2012.
- [11] M. R. Teague, "Image analysis via the general theory of moments," *Journal Optical Society American*, vol. 70, no. 8, pp. 920–930, 1980.
- [12] R. Mukundan, S. H. Ong, and P. A. Lee, "Image analysis by tchebichef moments," *IEEE Transactions on Image Processing*, vol. 10, no. 9, pp. 1357–1364, 2001.
- [13] B. Xiao, J. F. Ma, and X. Wang, "Image analysis by bessel-fourier moments," *Pattern Recognition*, vol. 43, 2010.
- [14] B. Yang and M. Dai, "Image analysis by gaussian-hermite moments," *Signal Processing*, vol. 91, 2011.
- [15] L. Fan, F. Long, D. Zhang, X. Guo, and X. Wu, "Applications of independent component analysis to image feature extraction," in *Proceedings of the Second International Conference on Image and Graphics*, vol. 4875, 2002, pp. 471–476.
- [16] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Waltham: Elsevier, 2006.
- [17] A. H. Kulkarni and S. A. Urabinahatti, "Performance comparison of three different classifiers for hci using hand gestures," *International Journal of Modern Engineering Research*, vol. 2, no. 4, pp. 2857–2861, 2012.
- [18] R. E. Maleki, A. Rezaei, and B. M. Bidgoli, "Comparison of classification methods based on the type of attributes and sample size," *JCIT*, vol. 4, no. 3, pp. 94–102, 2009.
- [19] A. Safia and D. He, "New brodatz-based image databases for grayscale color and multiband texture analysis," *ISRN Machine Vision*, no. id 876386, 2013.

Point cloud partitioning approach for ICP improvement

Nicolas S. Pereira, Cinthya R. Carvalho, George A. P. Thé

Depto. De Engenharia de Teleinformática

Universidade Federal do Ceará

Fortaleza, Brazil

Email: nicolassilva.ti@gmail.com, george.the@ufc.br

Abstract—In 3D reconstruction applications, an important issue is the matching of point clouds corresponding to different perspectives of a given object in a scene. Traditionally, this problem is solved by the use of the Iterative Closest point (ICP) algorithm. In view of improving the efficiency of this technique, in this paper we propose a preprocessing step which works on the raw point cloud. Additionally, we propose a metrics to evaluate the outcome of the ICP algorithm. Our experiments have been carried out on artificial as well as on real depth maps acquired from a time-of-flight sensor, and revealed that our cloud partitioning approach makes the ICP algorithm to run 25 times faster, at least.

Keywords—component; Iterative Closest Point; point cloud registration; point sampling

I. INTRODUCTION

In the context of industrial automation, computer vision is playing an important role as a provider of innovative and efficient solutions for optimizing shop floor processes [1]. As an example, collision avoidance [2] is a hot research topic because it opens up the way to operation even in non-structured scenarios. To name a few, in [3], for example, the harmonic and safe coexistence of humans and machines is investigated, and in [4], is proposed a framework for avoiding moving obstacles during visual navigation with a wheeled mobile robot. Particularly for industry, safe human-machine interaction is a goal, because it may speed up manufacturing processes, eventually leading to revision of approved protocols and regulations for shop floor operation.

In collision avoidance, detecting and identifying an obstacle is the first challenge. Indeed, videosurveillance systems for non-structured environments usually requires occlusions to undergo some treatment; it is an important issue, and several solutions have been discussed in the literature [5]. Among those, data fusion of images from different camera perspectives is very popular, and consists essentially on making partial descriptions of a scene to merge into a whole representation of it, for example, a 3D model for the surface of an object.

In the literature, this task has been first solved with the Iterative Closest Point algorithm [6]. The goal of this algorithm is obtain the transformation able to minimize the distance between two datasets, for example, two point clouds of a given application, allowing for integration of images acquired from different camera position and orientation, as illustrated in Fig. 1.

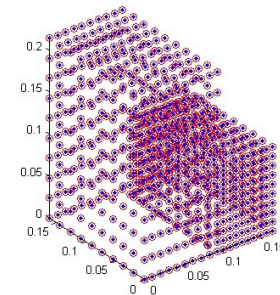
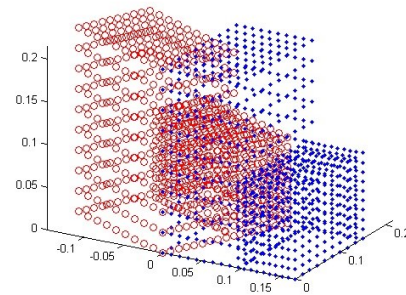


Figure 1. Example of the first and last step of the ICP algorithm.

State-of-the-art of the ICP algorithm includes the paper of [7], which proposed a protocol that allows a comparison between ICP variants, as well as the very recent approach of [8], which proposed a new ICP variant and compares it to the Generalized ICP (GICP) [9]. According to [10], prior to running ICP algorithm, the point cloud has to be analyzed and conveniently preprocessed to efficient operation.

In line with that, in the present paper a new approach for input data selection prioring to ICP algorithm is proposed, which consists on a partitioning of a given point cloud into smaller clouds, hereafter named sub-clouds. As results will reveal, this approach leads to significant improvements on ICP execution time. Furthermore, for very large point clouds, the efficiency of the proposed approached makes it a better choice than the traditional ICP.

For evaluating the technique, other than performing no prior selection of the input points, we implemented two additional approaches, namely the random points selection and vector quantization from an unsupervised self-organization map (SOM) neural network.

This work is organized as follows: in Section II the methodology is discussed and the different datasets are described. In Section III is reported the results for the image fusion of the different datasets, and the performance of the points selection approaches. Finally, in Section IV, conclusions are drawn.

II. MATERIALS AND METHODS

To study the image fusion problem, a very simple experimental setup containing only one object-of-interest (OOI) on a scene was designed. In Fig. 2, a photograph of the scene with the OOI in evidence is shown. Different perspectives of the OOI are then acquired from a time-of-flight 3D camera (from ifm electronics) whose output is a 3D depth map (a point cloud) of the environment. The aim is to investigate how different approaches for pre-selecting the point clouds lead to efficient execution of the ICP algorithm.



Figure 2. Object of interest.

To accomplish with that, two scenarios were studied: in the first one, the OOI shown in Fig. 2 is represented by an artificial point cloud, whereas in the second one we work on the experimentally acquired point cloud itself. The goal on using an artificial dataset is that of getting a better visualization of the different 3D perspectives, since the sensor available is a noisy and very low-resolution device.

A. Artificial point cloud

In Fig. 3, the artificial point cloud representing the OOI is shown. Part a) shows the cloud having 847 points, whereas in b) it is illustrated a denser point cloud representation (to mimic a high-resolution acquisition device).

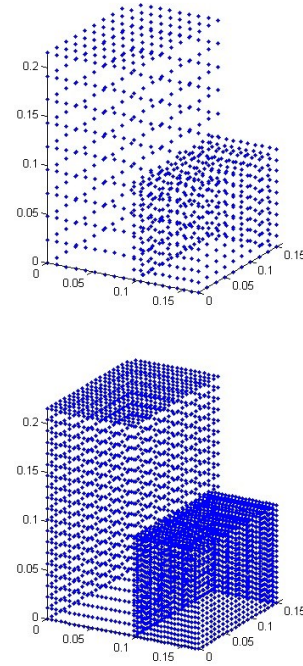


Figure 3. a) Low density artificial cloud point with 847 points b) Denser artificial point cloud with 3687 points.

The point clouds of Fig. 3 were then sampled according to four approaches:

- a) no sampling: every point of the artificial cloud was selected to undergo the ICP algorithm
- b) random sampling
- c) sampling from SOM-based vector quantization
- d) sampling from partitioning cloud approach (authors' proposal)

Approach of b), which is a homogeneous sampling consists on taking a specific amount of points as representative of the whole cloud. The approach of c) instead takes into account the spatial distribution of the data and how they are organized; a number of regions is chosen in advance and the centroids of each region is obtained after convergence of an unsupervised SOM-based neural network.

B. Proposed technique

In the point cloud partitioning approach proposed in this paper, both datasets required to perform the ICP algorithm are divided in k subsets where each subset have the same amount of points. Fig. 4 shows an example of a cloud point after being partitioned where k equals to 4. For better visualization, it was given an offset between the sub-clouds.

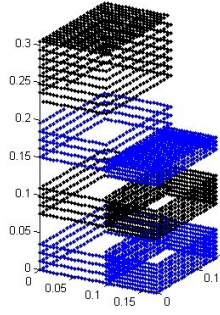


Figure 4. Proposed technique with 4 sub-clouds.

After partitioning, each k -th sub-cloud pair, composed by the k -th sub-cloud of each dataset, is utilized as an input for the ICP algorithm.

Consider that the number of points in each complete cloud is N_x and N_p . The cost for the closest point computation is $O(N_x N_p)$. Utilizing the algorithm, the same process of closest point computation is now $kO(N_x N_p/k^2)$, which can be seen that the bigger the k , smaller the cost. The closest point calculation is selected as reference because it is the most expensive step of the ICP algorithm.

C. Adding new perspective

In order to perform point cloud registering from ICP algorithm, at least two clouds are required. Therefore, taking the artificial cloud of Figure 3b as a reference, we applied a known rotation operation, thus producing the point cloud of Fig. 5.

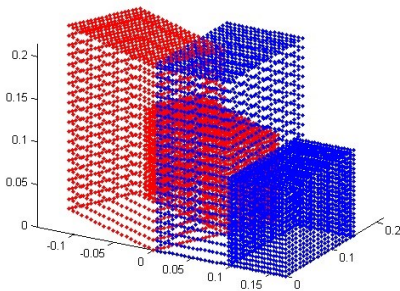


Figure 5. Point clouds used as input for the ICP algorithm.

For a given artificial dataset and its corresponding rotated perspective, we adopted (1) as the metrics to evaluate the performance of the different point selection approaches:

$$d(k) = I - R_R * R_k \quad (1)$$

Where $d(k)$ gives, for each k -th sub-cloud the distance between the identity matrix and the product of R_R and R_k matrices. These matrices represent, respectively, the known rotation operator and the output of the ICP algorithm having the k -th sub-cloud as input.

Whenever d equals zero, matrices R_R and R_k represent opposite rotations and in such case we can state the ICP algorithm was able to find, from the data of the k -th sub-cloud only, the known rotation applied to the whole ensemble.

D. Experimental point cloud

In addition to the study based on artificial point clouds, we also inputs real data as acquired from the TOF sensor previously described. Here again we work with 2 datasets, described in the following:

1) One perspective only

After acquiring the point cloud from the sensor for a given perspective view of the OOI, we subject the whole point cloud to a known rotation operation in order to produce a second dataset (this is similar to the procedure of Section II.C). The ICP algorithm can therefore be runned and the metrics of (1) is used again.

2) Two acquired perspectives

In the last scenario the data available comes from acquisitions of two different (arbitrary and unknown) viewpoints. After acquisition, there is a preprocessing for extracting the OOI from the scene before entering the points selection step and the ICP algorithm itself. Note that the second perspective of the OOI may be regarded as a rotation transformation in the point cloud corresponding to the first perspective, but it is not known previously and the metrics of (1) fails.

To overcome that, we propose to compare the matrices $R_{i,k}$ (which are in turn the outputs of the ICP algorithm for each k -th sub-cloud) to the matrix R_O , which is the output of the ICP algorithm when the whole point cloud is used instead. Equation (2) synthetizes this calculation, where the index i indicates the adopted points selection approach:

$$d_i(k) = R_O - R_{i,k} \quad (2)$$

III. RESULTS

This section presents the performance of the different points selection approaches in terms of elapsed time from the instant just after the segmentation of the OOI in the scene until the ICP algorithm termination. A few details about the different point selection approaches are given below:

a) in the random sampling approach, we investigate the cases in which: i) 50% of the original data is used for point cloud registration and ii) 70% of the original data is used.

b) two SOM neural networks have been designed; they have 125 neurons (5x5x5 grid) and 512 neurons (8x8x8 grid), respectively.

c) in the point cloud partitioning approach, the input data to the ICP is divided into smaller point clouds; in the present study, the number of sub-clouds ranges from 2 to 100.

A. Artificial point clouds

The performance of the ICP algorithm for different points selection approach is reported for both datasets of Figures 3a and 3b in the following tables.

TABLE I. RESULTS FOR THE FIRST DATASET

	Selection approach					
	ICP with all points	Random		SOM		Partitioning
		50%	70%	[5,5,5]	[8,8,8]	
Metric	0	0.021	0.0070	0.0201	0.0219	0
Time(s)	10.64	12.68	21.49	29.51	168.15	0.42

TABLE II. RESULTS FOR THE SECOND DATASET

	Selection approach					
	ICP with all points	Random		SOM		Partitioning
		50%	70%	[5,5,5]	[8,8,8]	
Metric	0	0.0122	0.0077	0.0594	0.0277	0
Time(s)	233.70	247.05	383.17	115.12	584.53	2.98

These results show that the proposed approach outperforms the other approaches in both running time and efficiency of the point cloud registration. Indeed, when sub-clouds undergo the ICP algorithm, the process is at least twenty five times faster (compare 0.42s to 10.64s, Table 1 and 2.98s to 115.12s, in Table 2). Moreover, the distance between matrices R_R and R_K equals zero only for the proposed technique (see the row labeled as Metrics in the tables).

B. Experimental point clouds

We report first the results of the ICP algorithm for the experiment described in Section II.D.1. The dataset of this case is plotted in Fig. 6, and the results for each points selection approach tested are reported in Table 3. Once again, with the proposed technique ICP algorithm reached superior performance, with only 3.09s elapsed time.

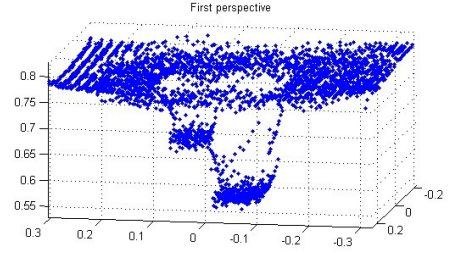


Figure 6. First perspective.

TABLE III. RESULTS FOR THE ONE PERSPECTIVE ALONE CASE

	Selection approach					
	ICP with all points	Random		SOM		Partitioning
		50%	70%	[5,5,5]	[8,8,8]	
Metric	0	0.9046	0.9045	0.0380	0.0031	0
Time(s)	529.14	74.91	422.72	104.81	501.13	3.09

Moving to the experiment described in Section II.D.2, in Fig. 7 it is plotted the second point cloud, which is an arbitrary and unknown perspective of the OOI to be discovered by the ICP algorithm.

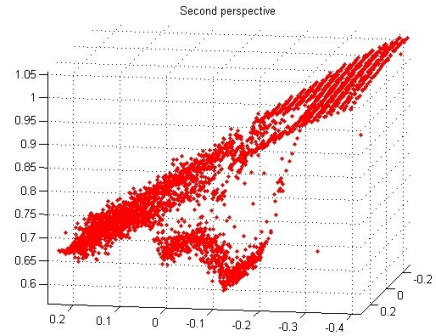


Figure 7. Second perspective.

TABLE IV. RESULTS FOR THE DUAL PERSPECTIVE CASE

	Selection approach					
	ICP with all points	Random		SOM		Partitioning
		50%	70%	[5,5,5]	[8,8,8]	
Metric	-	0.04	0.05	0.35	0.19	0.06
Time(s)	16.82	5.10	10.62	23.95	121.66	0.19

In Table 4 the performance results for this case are reported. Once again the proposed technique beats all the other points selection approaches, making the ICP algorithm faster. If compared to the the 50% random sampling approach, the proposed cloud partitioning technique is about thirty times faster and, in comparison to the no-sampling approach, about a hundred times faster. For what concerns the point registration efficiency, although it did not lead to the best registration result (the random sampling outperformed on this), the metrics were on the same magnitude order. For the sake of illustration, in Fig. 8 the results of the point cloud registration using the proposed technique is reported.

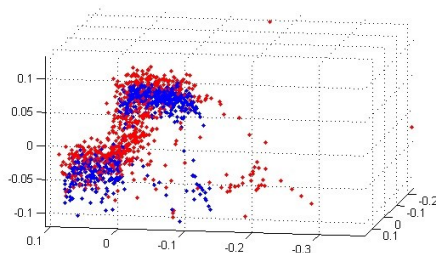


Figure 8. Point cloud registration.

IV. CONCLUSION

In this paper we proposed a point cloud partitioning approach as a preprocessing for the ICP algorithm. Also a metrics has been proposed, and the results revealed that the proposed technique led to better efficiency of the ICP, especially when compared to other point cloud sampling methods, such as random sampling and vector quantization by an unsupervised SOM network.

The main finding is an impressive 25 times faster convergence of the ICP algorithm. In addition to that, the quality of the point cloud registration itself was improved (or kept unchanged, in some scenarios), making, therefore,

this point cloud selection an option for real-time point cloud registration.

ACKNOWLEDGMENT

Authors thank NUTEC and CENTAURO, for administrative facilities.

REFERENCES

- [1] E. N. Malamas, E. G. M. Petrakis, M. Zervakis, L. Petit, and J. D. Legat, "A survey on industrial vision systems, applications and tools," *Image and Vision Computing*, vol. 21, 2003.
- [2] S. Hutchinson, P. Leven, Planning collision-free paths using probabilistic roadmaps, in *Handbook of Geometric Computing: Applications in Pattern Recognition, Computer Vision, Neural Computing and Robotics*. By E. B. Corrochano (org.), chapter 23, Heidelberg, Springer, 2010.
- [3] G. Oriolo, A case study of safe human/robot coexistence, Research Proposal, available at <http://www.dis.uniroma1.it/~labrob/theses/theses.html>, DIAG Robotics Lab, Sapienza University of Rome, 2015.
- [4] A. Cherubini, F. Spindler, F. Chaumette, Autonomous Visual Navigation and Laser-Based Moving Obstacle Avoidance, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, NO. 5, 2014.
- [5] A. S. Ogale, C. Fermüller, Y. Aloimonos, Detecting independent 3D movement, in *Handbook of Geometric Computing: Applications in Pattern Recognition, Computer Vision, Neural Computing and Robotics*. By E. B. Corrochano (org.), chapter 12, Heidelberg, Springer, 2010.
- [6] P.J. Besl, N.D. McKay, A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992.
- [7] F. Pomerleau, F. Colas, R. Siegwart, S. Magnenat, Comparing ICP variants on real-world data sets, Open-source library and experimental protocol, *Auton Robot*, 2013.
- [8] J. Serafin, G. Grisetti, Using Augmented Measurements to Improve the Convergence of ICP, in *Simulation, Modeling, and Programming for Autonomous Robots*. By D. Brugali, pp 566-577, Springer, 2014.
- [9] A. V. Segal, D. Haehnel, S. Thrun, Generalized-ICP. In: *Proc. of Robotics: Science ad Systems (SS) (2009)*
- [10] A. Torsello, E. Rodolà, A. Albarelli, Sampling Relevant Points for Surface Registration, 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, 2011.

On the Algorithm For Reconstruction of Polyhedral Objects From a Single Line Drawing

Shengfeng Qin¹ and Huaiwen Tian²

¹Department of Design, Northumbria University, Newcastle upon Tyne, NE1 8ST, UK

²School of Mechanical Engineering, Southwest Jiaotong University, Chengdu, China

Sheng-feng.qin@northumbria.ac.uk

Abstract— This paper presents a new level-by-level 3D reconstruction method from a single axonometric wireframe drawing with or without hidden edges of a planar object. This method solves a 3D reconstruction problem in stepwise fashion via face propagation, and allows a detailed relative importance study of key regularities and their usages for reconstruction rules establishment, which is the focus of this paper. Based on the proposed reconstruction method, two new trustfulness measures in terms of form and size distortions have been devised to evaluate regularities and their limitations. As a result of adapting the existing concepts of the MSDA [5] and the MSDSM [9] into a stepwise system, two new localized regularities have been developed in terms of the localized minimizing standard deviation of angles (L-MSDA) and the localized minimizing standard deviation of segment magnitudes (L-MSDSM). The proposed method and the identified key regularity have been tested with many cases. A range of weightings for the combination of L-MSAD and L-MSDSM regularities have been identified for practical use. The test results show that (1) the level-by-level reconstruction method is useful to 3D reconstruction, (2) the combined usage of localized key regularities L-MSDA and L-MSDSM, can produce satisfactory results with the proposed method, and (3) the demonstrated success of the combined usage of L-MSDA and L-MSDSM suggests that the concept of combining MSDA and MSDSM should be utilized in any optimization-based reconstruction approaches to improve their accuracy.

Keywords—line drawing; level-by-level reconstruction; key regularity identification; polyhedral objects

I. INTRODUCTION

3D reconstruction of a 2D line drawing is a crucial research topic in the field of artificial intelligence (AI) and computer vision [1]. In the recent years, a new interdisciplinary research area of sketch-based interface and modelling has emerged. Conversion of freehand sketches to line drawings involves sketch segmentation, sketch stroke grouping and line fitting, connection of lines at appropriate junctions and other 2D regularity enhancement. This is regarded as early processing for sketch understanding. We do not discuss this preprocessing and refinement works here, assuming that scribbling and overtracing have been removed and refined so that a resultant clean line drawing consists of lines well connected at vertices [15], and take line drawings with or without hidden lines as the starting-point for our 3D reconstruction work.

In order to create 3D geometric models from 2D line drawings, line drawings need to be represented properly

for different reconstruction approaches. In general, they can be described in graph-theoretic terms either as face/surface connection graph (FCG/SCG [2]) or edge-vertex graph [3]. These two representation models provide boundary information of a 3D object at different levels. Most of optimization-based inflation 3D reconstruction processes [3-4] use edge-vertex graph while others such as a Cubic Corner approach [12] utilize face connection graph. In our research, we use face connection graph representation.

Research work in the reconstruction of 3D objects from single 2D line drawings has been summarized in [4]. 3D reconstruction of polyhedral objects from single parallel projection largely uses optimization-based inflation methods, which add depth to vertices iteratively and test a set of constraints through their compliance functions. In general, a number of image regularities need to be applied with different weightings as constraints in their compliance functions. However, it is very difficult to know how to set up weightings for regularities because we don't know what regularities should be utilized in the optimization process against a variety of line drawings and their relative importance.

Our research here aims to study relative importance of some regularities by the proposed trustfulness measures and then identify what are the key regularities in 3D reconstruction of a 2D line drawing and test their utility. Therefore, at the beginning, accurate 2D line drawings are provided by projecting existing 3D objects from CAD models for trustfulness tests. Also, in order to remove projection ambiguity issues, the projection parameters used in producing 2D drawings are used in our reconstruction process. As mentioned before, we use a face connection graph to represent a 2D drawing and face topology to guide our investigation. In this way, some image-based regularities [8] such as face planarity and line parallelism can be automatically guaranteed because the proposed method uses faces instead of edges to propagate in stepwise fashion. Thus, we can focus on the relative importance study of key regularities as well as their combinations. After key regularities and their usages are identified, they are tested satisfactorily with the proposed level-by-level reconstruction method for many cases. The line drawings for these cases are generated with AutoCAD®. We have developed a Matlab® program to test and evaluate this method. The test results show that this method is able to achieve good reconstruction results for many cases.

The paper is structured as follows. Section 2 discusses the related work and section 3 presents the proposed method in principle. The regularity trustfulness evaluation and key regularity identification are described in section 4, followed by some complex case evaluations in section 5. Finally, discussions on the method and identified key regularities are included in section 6 followed by conclusions in section 7.

II. RELATED WORK

Early work is focused on line labelling of 2D line drawings [1] [5]. Line-labelling attempts to identify each line in the drawing by identifying its corner types. An optimization-based inflation process is then followed to produce 3D objects.

Marill [6] observed that human minds prefer a simple interpretation over a complex one and defined the simplicity as minimizing the difference among angles created between lines at junctions across the reconstructed object. The use of minimizing the standard deviation of the angles (MSDA) in an inflation process produced much higher likelihood of the computer interpretation of the scene matching the human interpretation. [7] extended the method by adding face planarity as another image regularity in optimization. Both can recover a limited range of objects. [8] extended the approach further by using more image regularities into optimization process, and demonstrated the recovery of a wide range of objects, both manifolds and non-manifolds. Brown [9] modified Marill's approach by replacing the angles at vertices by the line segment lengths in his optimization process, minimizing standard deviation of segment magnitude (MSDSM). The use of MSDSM obtained better results in some drawings which cannot be handled by MSDA. This research suggested searching for alternative regularities to reflect the simplicity requirement.

Varley et al [10] tested the necessity of using line labelling and developed an alternative approach to obtain frontal geometry in an inflation process. They found that the cubic corner property in their compliance function was particularly useful. A cubic corner is a junction where three mutually orthogonal planes meet. Perkins established the relationship between a cubic corner in 3D and its 2D projection [11]. In general, optimization-based inflation methods above can produce good results for some cases and bad results for others depending on how to choose regularities and their weights.

Lee and Fang [12] studied a direct method for recovering a 3D object from a 2D drawing (a single parallel projection of the 3D object). In order to deal with imperfection or inaccuracy of 2D sketches, the authors also investigated a hybrid method [13] by incorporating the cubic corner method into an optimisation-based inflation process with six regularities: face planarity, line parallelism, corner orthogonality, skewed face orthogonality, MSDA and isometry.

Yuan et al [16] reviewed the usages of various regularities, conducted an automatic relevance determination study and finally recommended line parallelism, face planarity, skewed facial symmetry,

skewed facial orthogonality and MSDA as an optimal regularity set for general 3D reconstruction. The MSDA regularity is found the most important rule.

It can be seen that different regularities have been used in previous research, but little work has been done to evaluate their relative importance. It is questionable whether using different weights in association with different regularities would make any difference given that we don't know how to set up proper weightings.

III. THE LEVEL-BY-LEVEL RECONSTRUCTION SCHEME

In engineering design, typically, a single line drawing to depict a 3D object is a parallel axonometric projection of the object such as its isometric drawing or oblique drawing with or without hidden lines (Figure 1). An axonometric drawing can provide a general view of the object. In such a drawing, there exist three object-relative perpendicular axes (i , j , and k , or a local coordinate system O -XYZ), none of which is aligned with the 2D drawing x - and y -axes or the perpendicular z -axis added by inflation to 2.5D. Thus, direct reconstruction of an axonometric drawing into its object-relative axes is very attractive to Engineering designers. The identified axis-aligned planes [17] can be used in an inflation process [10], and Qin [18] utilized the projection rules for generating isometric drawing to enhance freehand sketches for conceptual design modelling.

[19] interpreted a 3D object directly from a rough 2D oblique drawing, given knowledge of the directions of the principle axes in the drawing, line labelling and the adjacent graph of the drawing. The limitations of this approach were discussed in detail in [20]. All these methods [8, 12, 13, 19, 20] used the cubic corner approach described by Perkins [11]. For a three-connected vertex, when applying Marill's MSDA rule, the cubic corner is expected in result. It is well known that Marill's MSDA rule is good for some cases and bad for others. But none of them tried the combination with Brown's MSDSM regularity and applied MSDA locally.

When reading such a drawing, as trained from Engineering Graphics courses, we usually start with a face as a key reference and then understand its connected faces level-by-level as a component. Similar to our reading and understanding process of a drawing, recognition-by-components is regarded as a theory of human image understanding [21], thus we propose a level-by-level 3D reconstruction scheme.

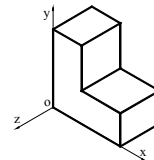


Figure 1 Axonometric projection

This level-by-level 3D reconstruction method doesn't use an inflation process (taking the x - and y -coordinates of a vertex in 2D drawing as its true 3D X - and Y -coordinates and then inflating its Z value) to reconstruct 3D objects from 2D drawings. Instead, it regards the x - and y -coordinates of a vertex in 2D

drawing as the projection of the corresponding 3D point (X, Y, Z). Their relationship is well defined by the theory of axonometric projection. Thus, the x- and y-coordinates of a 2D vertex in the drawing is not the same as its corresponding 3D X and Y coordinates. Therefore, this method cannot be directly compared against inflation-based approaches. According to the theory of axonometric projection, the relationship between a 2D vertex (x, y) in the drawing and its 3D corresponding point (X, Y, Z) is

$$[x \ y \ 0 \ 1] = [X \ Y \ Z \ 1] * T \quad (1)$$

Where

$$T = T_1 T_2 T_3 = \begin{bmatrix} \cos \alpha & \sin \alpha \sin \beta & 0 & 0 \\ 0 & \cos \beta & 0 & 0 \\ \sin \alpha & -\cos \alpha \sin \beta & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2)$$

T_1 represents the transformation matrix of rotating the object α degrees clockwise around the Y axis, T_2 represents the transformation matrix of rotating the object β degrees counterclockwise around the X axis, and T_3 is the transformation matrix for the orthographic projection to XOY plane. Under this transformation and Equation 1, every 3D vertex V (X, Y, Z, 1) of the 3D object gets its 2D drawing vertex v (x, y, 0, 1) by $v = V * T$.

For reconstruction, it is clear that the matrix T is not reversible. But if a constraint is applied stepwise to the reconstruction process that a point to be reconstructed is on a planar face, then the matrix T can be updated for reverse transformation. For generality, a 3D planar face can be described by its face normal (a, b, c) and any 3D point on the face meeting Equation (3)

$$aX + bY + cZ + d = 0 \quad (3)$$

Now, we can use Equation 3 to modify the third column of the T matrix, as a result, we can get a reversible matrix R

$$R = \begin{bmatrix} \cos \alpha & \sin \alpha \sin \beta & a & 0 \\ 0 & \cos \beta & b & 0 \\ \sin \alpha & -\cos \alpha \sin \beta & c & 0 \\ 0 & 0 & d & 1 \end{bmatrix} \quad (4)$$

With the matrix R, if we know any 2D projection (x, y) comes from a 3D point on the planar face, its corresponding 3D coordinates can be obtained by a reverse transformation, that is,

$$V = v * R' \quad (5)$$

Based on the above ideas, the level-by-level 3D reconstruction includes the following steps:

Step 1: Select the first reference face (parent face) and conduct its reconstruction. Currently, the first parent

face must be interactively selected from some faces parallel to datum planes: XOY, XOZ, or YOZ and simply regarded as a datum plane. In this way, we can easily get the reference plane equation in terms of a, b, c, and d parameters for Equation 4. For a given axonometric projection, we have known the transformation parameters such as $\alpha=45^\circ$ and $\beta=35^\circ$. As a result, R can be obtained and from Equation (5), all vertices on the first parent face can be reconstructed.

Step 2: once the first parent face is selected, the corresponding face connection graph (FCG) can be established as a resolvable representation of the object [28] to guide the reconstruction process in due course. Face connection information can be gained from a face finding process [22-26].

Start from the first parent face, its connected faces will be its child faces and each child face will have its own child faces. A parent face connects to each child face via a connecting edge. For instance in Figure 2a, five faces F1, F2, ..., F5 can be identified. If the face F1 is selected as the first parent face, the face connection graph can be built up as shown in Figure 2b. F1 is the parent face for F2 and F3, and in turn, F2 and F3 are child faces for F1. F1 connects F2 and F3 via the connecting edge e1 and e2 respectively. Similarly, F2 is the parent face for F4 and F4 is a child face of F2. Their linkage is the connecting edge e5. F3 has a child face F5 and linked via the edge e6. The relationship between the edges e1 and e2 or e5 and e6, such as sharing a common vertex can also be established during the face finding process.

Step 3: Reconstruct the child faces level-by-level. Once we know the parent faces in 3D and their connections. The connecting edges between the parent faces and their corresponding child faces are known in 3D. Now if we rotate a parent face from the current position about the corresponding connecting edge, a new possible child face passing through a common vertex (x_0, y_0, z_0) can be obtained via its new normal (a', b', c') and $d' = -(a'x_0 + b'y_0 + c'z_0)$ from Equation 3. For example, rotating F1 about the edge e1, a possible face F2 meeting the projection requirement can be found, and the F3 will be determined accordingly based on their connections: sharing the same parent F1 and sharing a common edge. Therefore, we can find one best rotation angle for determining the faces F2 and F3 (actually all faces at the same level). F1 is at Level 0. F2 and F3 are at Level 1 and share the same parent face while F 4 and F5 are at Level 3, not sharing the same parent face but linked by a common edge. Let F2 and F3 have been reconstructed, edges e5 and e6 will be known in 3D. In order to reconstruct the face F4, we can use the same process to rotate the parent face F2 about a known edge e5 to determine a possible solution for F4. Given F4 in 3D, F5 will be obtained from their connection information.

From the above reconstruction process, it is clear that the reconstruction process is iterative to reconstruct child faces from known parent faces in level-by-level fashion. The reconstruction step is repeated until all faces are updated to 3D.

The above reconstruction scheme needs a set of

good regularities to be incorporated in Step 3 to stop a rotation process and give best reconstruction results to faces on a level. Therefore, there is a need to identify best regularities and their usage.

Compared with Grimstead's linear optimization method [15] and other non-linear optimization method [8], the authors prefer this stepwise method because it can provide a good chance to study key regularities and their relative importance, not on the grounds that non-linear optimization is very slow.

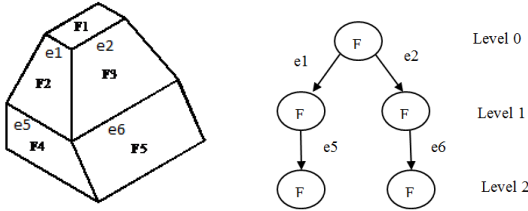


Figure 2 Faces in a drawing and their connection graph

IV. SEARCHING FOR KEY REGULARITIES

From previous research, it is not very clear what are key regularities and how to set up their weightings. In order to search for key regularities, we designed a series of tests to identify key regularities and their weightings.

We used a non-regular pentagonal prism and its variations as 3D truncated pyramid objects to study key regularities. This looks a simple test case but actually it is hard for both original MSDA and MSDSM as clearly stated in [8]. The section of the initial prism is a five-sided polygon with no symmetrical axis but all vertices are on a common circular circumference (see fig 3). The height of the prism is set to 70. Then, we created 9 different variations by keeping the top section and the height unchanged and scaling the bottom section around the center of the reference circle with scaling factors (sf= 3, 2.5, 2, 1.5, 1, 0.8, 0.5, 0.3 and 0.1 respectively). The variations are shown in Figure 4. We built 3D CAD models for each case and then projected them to produce the corresponding 2D line drawing for our tests.

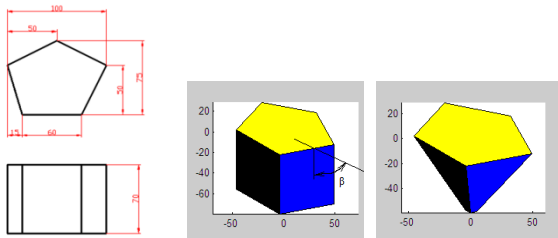


Figure 3 A test prism

When constructing child faces at the same level, note that two child faces share a common edge and the parent face and a child face share a connecting edge (See Fig 5a), therefore, all child faces at the same level will be constructed in the same iteration. For each level, it is intent to interpret all child faces as simple as possible. According to this principle, we tested the following four regularities to stop a search process: (1) minimising child face areas, (2) minimising length of all shared common edges, (3) minimizing localized Standard Deviation of the Angles (L-MSDA), that is, all angles between all pairs of

lines meeting at junctions associated with the parent faces must be similar, and (4) minimizing localized Standard Deviation of Segment Magnitudes (L-MSDSM), that is, all edges meeting at junctions associated with the parent faces must be similar.

A. Localized MSDA regularity (L-MSDA)

As Varley and Martin commented in [27], the minimum standard deviation of angles (MSDA) proposed by Marill [6] "is a property of the object as a whole, not a local property, it cannot be incorporated in a linear system approach. It is not ideal even for the optimization approach which adjusts a single vertex at a time since the MSDA for the entire object must be recalculated after each adjustment."

Here, we interpret the MSDA locally and loosely to form a localized MSDA (L-MSDA) regularity, that is, all angles between all pairs of lines meeting at junctions on faces associated with the parent faces must be similar. In other words, all evaluable angles at the current reconstruction level must be similar or with the minimum standard deviation.

As shown in Figure 5b, at the current reconstruction level, the parent face is P, having three child faces C1, C2 and C3. The interior angles of the parent is already known and fixed. Therefore, all interior angles A_i ($i=1,2,\dots,K$) on child faces such as A1 to A12 are evaluable at this current level. When rotating a parent face such as C1 of \square degree around its 3D connecting edge such as AB to the parent, all these interior angles on child faces are changed accordingly. Thus, the localized standard deviation of (evaluable) angles is given by δ_{\square}

$$\delta_{\square} = \sqrt{\frac{1}{K} \sum_{i=1}^K (A_i - A_m)^2} \quad (6)$$

Where A_m is the average of all interior angles of child faces. Finding the corresponding angle \square ranged from 0 to 180 degree with the minimum of δ_{\square} will be a solution for the current level searching.

Note that there are two differences between the proposed L-MSDA and the MSDA [6]. First, L-MSDA only involves local evaluable angles at the current reconstruction level, not referencing backwards to known angles from all previous construction levels and forwards to anything unknown at next levels. Second, it is applied to only interiors angles of child faces, not to all angles at junctions on the parent faces. This relaxes the regulation for local application with respect to the fact that parents are born before their children.

B. Localized MSDSM regularity(L-MSDSM)

For the minimum standard deviation of segment magnitudes (MSDSM) regularity proposed by Brown in [9], similar to Marill's MSDA [6], it is a global property of the object, not a local one. For our level-by-level reconstruction method, the MSDSM regularity has also been localized, that is, all edges meeting at junctions associated with the parent faces must be similar. Because the connecting edges between the parent face(s) and child faces are already known for this current reconstruction level, we only take the common edges into consideration.

That is, if there are N common edges (see Figure 5a), the standard deviation of segment magnitudes is represented as σ

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (L_i - L_m)^2} \quad (7)$$

Where L_i is the length of the i th common edge, and L_m is the average length of all common edges.

Note that L-MSDSM is locally applied to common edges on the child faces at current reconstruction level while existing connecting edges are not utilized. This provides a loose bonding between parents' segment magnitudes and their children's and can help prevent failures from objects built from segments of varying lengths [9].

C. Evaluation of regularities

A searching process starts with an already reconstructed parent face and rotates it around a connecting edge shared by the parent face and its child face (see Fig 5a). The rotation angle is ranged from 0 to 180 degree and increased one degree per step. After a rotation, all faces on the same level can be constructed based on their connections. After that, the four regularities discussed previously are applied separately to check whether the corresponding regularity is met. If so, the searching process stops and the current rotation angle α is then used to reconstruct all child faces at the current level. In this way, applying each regularity to a level-by-level searching process results in a 3D object. Finally, the reconstructed 3D objects from different regularities are compared against the corresponding test object to measure their trustfulness (usefulness) of regularity. Based on our current knowledge, previous research paid little attention to measuring truthfulness of regularities. Here we define and test the truthfulness of a regularity by examining, how its reconstructed results are close to their originals. The trustfulness of a regularity is defined and measured as percentages of form distortion and size distortion respectively. The form distortion is calculated as a relative angle error Δ_β . For each test case, the ground true value of the angle β (see Figure 3) is known, giving $\Delta_\beta = \text{abs}(\alpha - \beta) / \beta$. Similarly, the size distortion is calculated as a relative area error Δ_a between the true surface area (a_0) and the reconstructed surface area (a_c), that is, $\Delta_a = \text{abs}(a_c - a_0) / a_0$. Note that trustfulness of regularity is used to evaluate relative importance of a single regularity or a combined usage, for key regularity and their usage selections. After that, it is not used in the reconstruction process.

After the testing, it was found that the first two regularities are not very useful because they produce quite big form distortion (for this reason, their results are not given here). But the latter two are quite good. Their results for each case are shown in Fig 4. In general, L-MSDA regularity is very good to reconstruct objects when β is close to 90 degree. But when β is out of a 90 degree range (90 ± 15 degree), it will give a big form distortion. For example, when $\beta=57$ in case 2, the $\Delta_\beta=39\%$, while $\beta=49$ in case 1, the $\Delta_\beta=67\%$. This means that L-MSDA regularity is quite sensitive to form

variations. When the true angle is far from 90 degree, the calculated angles with L-MSDA tend to be much bigger than the true value.

Therefore, the two key regularities are identified as (1) L-MSDA and (2) L-MSDSM. But they need to work together with proper weightings as their typical usage.

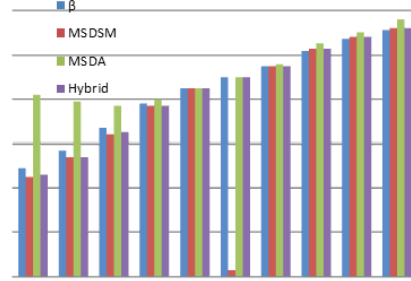


Figure 4 Angular error comparison for each case

V. TEST AND RESULT ANALYSIS

We have implemented our level-by-level reconstruction method in Matlab® with the identified key regularities: L-MSDA and L-MSDSM and tested their typical usage as: 50% MSDA+50% MSDSM for our cases. The reconstruction results are satisfactory. Fig 5 shows some successful examples while Fig 6 gives examples of challenges cases such as a transition part.

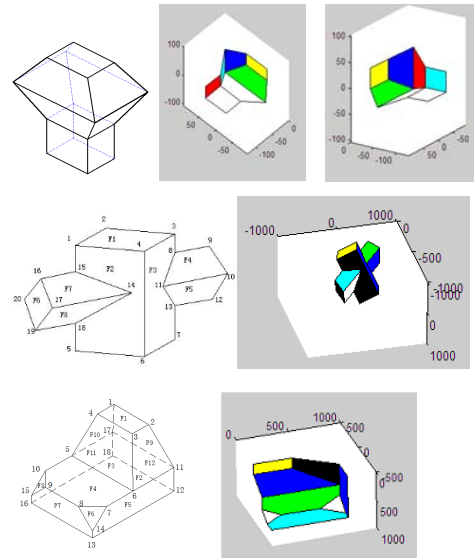


Figure 5 Examples of tested cases

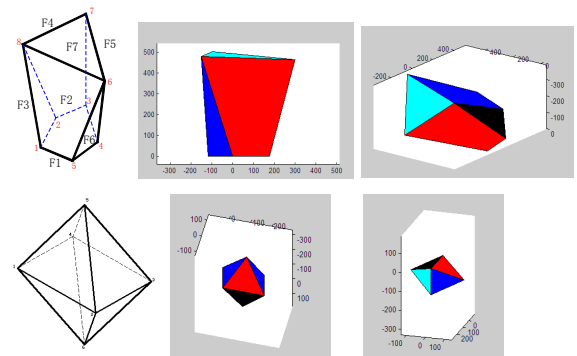


Figure 6 Challenging cases

For this case in Fig 6, if the face F7 (top face) is selected as the parent face, then its child faces will not share a common edge between them. But if the bottom face F1 is chosen as the parent face, then it is just a normal case. But in general, our method is not good at dealing with this type of parts. For the second case in Fig 6, we cannot find a real face for the parent. When we used a 'virtual' face in the middle, we can reconstruct it.

VI. DISCUSSION AND CONCLUSIONS

Sugihara [28] discussed resolvable representation of polyhedra and proved that any polyhedron homeomorphic to a sphere has a resolvable representation, in which small numerical errors do not violate the symbolic part of the presentation. A resolvable representation is a step-by-step reconstruction sequence which satisfies the above three conditions. Based on these conditions, all of our test examples are resulted from a resolvable representation.

The proposed new level-by-level 3D reconstruction method solves a reconstruction problem of polyhedral objects in stepwise fashion, thus it provides a good way to study key regularities and their usages by examining their trustfulness measurements in terms of form and size distortions of reconstructed objects. From our study, it is found that the concepts of localized regularities L-MSDA and L-MSDSM can be incorporated in a stepwise way.

With our method, the parallelism of two parallel lines in 2D on the same plane will be kept because a plane is determined before lines are reconstructed reversely on it. All lines in 2D parallel to the lines on the first parent face will be kept in parallel in 3D without the use of parallelism rule in reconstruction. Thus, for this method, a good 2D tidy-up process is required to deal with a rough 2D sketch. Afterwards, there is no need to incorporate parallelism as regularity in 3D reconstruction. And furthermore, we need to have a face graph generator to support this level-by-level reconstruction method.

REFERENCES

- [1] MB. Clowes, "On seeing things. Artificial Intelligence," Vol 2, pp.79-116, 1972.
- [2] T. Kanade, "A Theory of Origami World," Artificial Intelligence, Vol. 13, pp. 279-311, 1980.
- [3] P. Varley and R. Martin, "The junction catalogue for labelling line drawings of polyhedra with tetrahedral vertices," Int. J. Shape Modeling, Vol.7, pp. 23-44, 2001.
- [4] P. Company, A. Piquer, M. Contero, and F. Naya, "A survey on geometrical reconstruction as a core technology to sketch-based modeling," J. Computers & Graphics 2005; 29:892-904.
- [5] DA. Huffman, "Impossible objects as nonsense sentences," J. Machine Intelligence, Vol. 6, pp. 295-323, 1971.
- [6] T. Marill, "Emulating the human interpretation of line-drawings as three-dimensional objects," Int. J. Computer Vision, Vol 6, pp. 147-161, 1991.
- [7] YG. Leclerc, and MA. Fischler, "an optimization-based approach to the interpretation of single line drawings as 3D wireframes," Int. J. Computer Vision, Vol. 9 (2), pp. 113-36, 1992.
- [8] H. Lipson and M. Shpitalni, "Optimization-based reconstruction of a 3D object from a single freehand line drawing," J. Computer Aided Design, Vol. 28, pp. 651-663, 1996.
- [9] EW. Brown, "Why we see three-dimensional objects: another approach," 2004, (access: 1/10/2014).
- [10] P. Varley, R. Martin, and H. Suzuki, "Frontal geometry from sketches of engineering objects: is line labelling necessary?" J. Computer Aided Design, Vol.37(12), pp. 1285-1307, 2005.
- [11] D. Perkins, "Cubic corners. Quarterly progress report 89," MIT Research Laboratory of Electronics. pp. 207-14, 1968.
- [12] YT. Lee and F. Fang, "3D reconstruction of polyhedral objects from single parallel projections using cubic corner," J. Computer-Aided Design, Vol.43, pp.1025-1034, 2011.
- [13] YT. Lee and F. Fang, "A new hybrid method for 3D object recovery from 2D drawings and its validation against the cubic corner method and the optimisation-based method," J. Computer-Aided Design, Vol. 44, pp.1090-1102, 2012.
- [14] JZ. Liu, LL. Cao, ZG. Li and XO. Tang, "Plane-based optimization for 3D object reconstruction from single line drawings," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.30, pp:315-327, 2008.
- [15] IJ. Grimstead, "Interactive sketch input of boundary representation solid models", PhD Thesis, Cardiff University, 1997.
- [16] S. Yuan, LY. Tsui and S. Jie, "Regularity selection for effective 3D object reconstruction from a single line drawing," J. Pattern Recognition Letters, Vol. 29, pp.1486-1495, 2008.
- [17] P. Varley, R. Martin and H. Suzuki, "Progress in detection of axis-aligned planes to aid in interpreting line drawings of engineering objects," in ed. T. Igarashi and JA. Jorge, Sketch-Based Interfaces and Modelling, Eurographics Symposium Proceedings, 99-108, 2005.
- [18] SF. Qin, DK. Wright and IN. Jordanov, "From on-line sketching to 2D and 3D geometry: A system based on fuzzy knowledge," J. Computer Aided Design, Vol. 32, pp. 851-866, 2000.
- [19] D. Lamb and A. Bandopadhyay, "Interpreting a 3D object from a rough 2D line drawing," Proc. of Visualization'90, pp.59-66, 1990.
- [20] P. Company, J. Conesa and N. Aleixos, "Axonometric inflation in line drawings reconstruction, a Technical Reports," 2001. (<http://www.regeo.uji.es/publicaciones/regeo01.pdf>, accessed 08/12/2014)
- [21] I. Biederman, "Recognition-by-components: a theory of human image understanding," J. Psychological Review, Vol.94 (2), pp. 115-147, 1987.
- [22] JZ. Liu and YT. Lee, "A graph-based method for face identification from a single 2D line drawing," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 23(10), pp. 1106-19, 2001.
- [23] JZ. Liu, YT. Lee and W. Cham, "Identifying faces in a 2D line drawing representing a manifold object," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 24(12), pp. 1579-93, 2002.
- [24] P. Varley and P. Company, "A new algorithm for finding faces in wireframes," J. Computer-Aided Design, Vol. 42, pp. 279-309, 2010.
- [25] MC. Leong, YT. Lee and F. Fang, "A search-and-validate method for face identification from single line drawings," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.35(11), pp.2576-2591, 2013.
- [26] F. Fang and YT. Lee, "Efficient decomposition of line drawings of connected manifolds without face identification," J. Computer-Aided Design, Vol. 51, pp. 18-30, 2014.
- [27] P. Varley and R. Martin, "Estimating depth from line drawings," SM'02, June 17-21, 2002.
- [28] K. Sugihara, "Resolvable representations of polyhedra," Discrete and Computational Geometry, Vol. 21(2), pp. 243-255, 1999.

Morphology Element Research on Chinese Small-Sized Liquor Bottle Design

Ye Zhang^{1,2}, Huanzhi Lou¹, Hui Yu²

¹School of Architecture and Art Design,
Beijing Jiaotong University,
Beijing 100044, China
531892226@qq.com

²School of Creative Technologies,
University of Portsmouth
Portsmouth PO1 2DJ, United Kingdom
hui.yu@port.ac.uk

Abstract—Based on Kansei engineering methods and performed user testing, this work analyzes small-sized liquor bottle of the Chinese liquor market. Firstly, perform the cluster analysis and multidimensional scaling (MDS) analysis to obtain 26 representative sample bottles. Secondly, According to the analysis results of an expert group, the shape of the bottle is divided into six feature categories and 36 morphological unit categories. And then, the factor analysis table is built according to sample bottles. After that, the level of user preferences on bottle shape is evaluated by the image scale method over a number of consumers. By investigating associations between the user preference values and the form elements table from the values of the statistics table, researchers have found a preliminary result that some elemental characteristics indeed have a greater impact on consumer preferences. Finally, Pearson product-moment correlation coefficient inspection method is used to verify the conclusion mentioned above. Based on the correlation some elements are the key factors that affect consumer preferences in Chinese liquor market, while some are not preferable elements by customers. Those non-preferable elements should be considered to avoid in design. This study provides important reference and theoretical support for Chinese liquor bottle design and brand development.

Keywords—component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Liquor is one of the most important categories in food industry. With the development of Chinese economy in recent years and changes in social values, the traditional attribution of liquor as gift value is replaced by form of ownership. Due to poor sales in the 1 kg Pack liquor gifts in recent years, the survival pressure of liquor enterprises makes them focus more on young consumer groups. They investigated the status of liquor market, and then, started to promote 100-250ML products. Since 2012, all major companies rolled out over more than 200 small-sized liquor brands. Small-sized liquor will be an important development direction of liquor products and brands in the future. The shape design of liquor bottle is one of the most crucial and urgent problems. The product image plays an important role in the consumer's preference and choice of the products [1]. An appropriate bottle design can effectively deepen people's impressions of the product, and thereby enhance the consumer's desire to buy the product. Consumer-oriented Kansei Engineering

has been developed as a methodology that transforms a consumer's feeling or image about a product into the design elements of the products [2]. With Kansei research method, the proposed research attempts to look for the association rules between bottle-type form characteristic elements and consumer preferences.

Based on the principles of consumer-oriented, first the consumer's interests are tested in bottle-type and willingness of purchase under the same conditions. Second, further analysis is taken on liquor bottle design and classifies language elements from common design patterns with expert group. Third, Pearson product-moment correlation coefficient is used to study the relevancy between the form factor and user preferences. Finally, this paper describes the research process and methods in detail.

II. USER TESTING

In this research, the methodology is based on Kansei Engineering. Kansei Engineering has been applied successfully in the product design field in order to explore the relationship between the feeling (perception of the product image) of the consumers and the design elements of the product [3-7]. Perceptual studies design from collection to analysis is divided into four stages. Firstly, three groups of test subjects are established. Personnel of each testing group cooperate according to their functions. Second, the final testing samples are established by market survey group and the Panel of experts. Third, the Panel splits the samples and product elements table. Fourth, an online questionnaire of consumer preferences is achieved through the scale method. By the most suitable adjective pairs, Intention multidimensional method is used to quantize the level of users' psychological preference, combining with the design result of pictures [8]. Finally, we analyse user preference testing data obtained previously using the Pearson product moment correlation coefficient and decide the key issues to the design of small-sized liquor regarding to morphological elements. The details of each step are described in the rest of this paper.

A. Establish experimental group

Participants in the study are mainly divided into three parts. Group-A has ten people, which are seven men

and three women, with average age of 27.4 years old. They have drinking experience, but not alcoholics. They are assigned to collect 100-250ml small-sized bottles that existed in the market. They need to visit various shopping malls and buy as many samples of goods as possible. And they will remove the outer packages and logos of the bottle body pattern information of these samples and label them uniformly in order to prevent the final styling elements from the impact of graphic design, brand, color, and other non-bottle-type interference in perceptual studies process. Group-B is the expert group of five members with average age of 35.5 years old. There are two men and three women in this group. They have six years' experience in product design and brand planning, and they will classify main elements of key morphological characteristics of the typical liquor bottle based on experience and morphological analysis. Group discussion is used to determine the final 26 pieces of representative pictures for user testing (Figure 1). Group-C consists of the subjects of perceptual studies. This research uses a network data collection with a total number of 90 copies of questionnaires issued. The number of effective feedback copies is 79 with 88% valid rate. The age of subjects ranges from 18 to 55 years old. Male 48.15%, female 51.85%, 18-23 years old people accounted for 25.93%, 23-28 years old subjects accounted for 39.51%, 28-33 years old participants accounted for 17.28%, and only 1.23% for the age over 38. According to the feedbacks, we can conclude that the subject can represent the new generation, and they have drinking experiences, but are not alcoholics. By using the

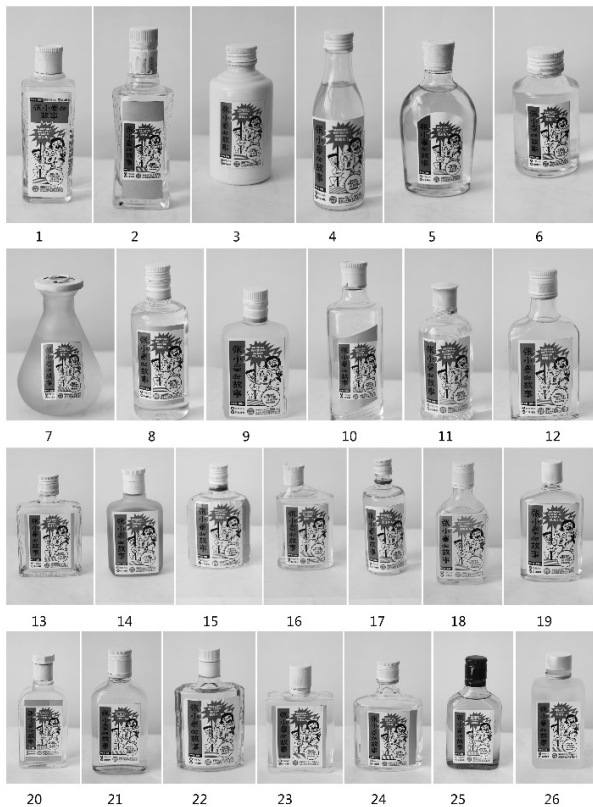


Fig.1 26 Test samples

SD method over these subjects, we can achieve the bottle intent scale scores, and get the statistical data with respect to test scores of 79 valid subjects over 26 sample bottles. After calculating the mean value and the standard deviation, these values can be seen as one of the basic data source for further analysis of user intentions.



































B. Image Samples to Determine

This study is aimed at small-sized alcoholic commodity whose volume of bottles is 100ml to 250ml. The market research staff (Group-B) visited all large, medium and small-sized markets and collected 65 samples with different brand and styles. These products are available to market after 2013. They can represent the majority of small-sized alcoholic commodities in current market. After collecting enough samples, without any pattern information such as packages and logos on the bottle body, we use uniform consistent external stickers and take pictures with white background (Fig.1). Through a computer screen with Kawakida Jirou Method^[8] based on similarity, the samples are divided into different groups. Then, based on the method of multiple scales (MDS) population analysis [9], 26 representative products are obtained finally. Then according to 26 of the elected sample, 5 members of the Group of experts produced morphological analysis tables.

C. Form Factors division

According to the 26 representative samples, based on morphological analysis, panelists (Group-B) describe the form elements of the samples based on their own experience and domains. First, they induce constituent elements of the bottle through several revisions and reviews. They delete duplicate and non-essential factor and retain important components of the bottles. Based on the characteristics of the product form, a bottle consists of four parts: caps, bottleneck, bottle shoulder, and bottle body. The different parts are subdivided into their morphological components, which follows the principle of non-redundant and no shortage of items. The sample is divided into six morphological characteristic categories (Table 1), where caps, based on the ratio of the width and height, can be divided into four classes. There are six kinds of forms in the shape of the bottle cap. Bottle necks can be divided into five categories. Bottle shoulder can be divided into eight categories. Based on the scale of the bottle body, it can be divided into seven categories. According to the top view, five most important features can be extracted. Then, the panel will be matched the 39 kinds of form feature elements to the 26 test bottle samples. Each bottle has its own shape feature element code. For example, as for the sample No. 8, the representative elements are $X_1 = 4$, $X_2 = 2$, $X_3 = 4$, $X_4 = 3$, $X_5 = 4$, $X_6 = 2$. These form the important element data for further analysis of the feature elements of the bottles.

TABLE 1 ELEMENTS OF MORPHOLOGY

Elements	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Type 8	Type 9
Cap Ratio (X_1)	 1 : 0.5	 1 : 1	 1:1.5	 1:2					
Caps Form (X_2)	 X_{2-1}	 X_{2-2}	 X_{2-3}	 X_{2-4}	 X_{2-5}	 X_{2-6}			
Bottleneck (X_3)	None	 Long Chamfer	 Short Chamfer	 Oblique Chamfer	 Curve Link				
Bottle Shoulder (X_4)	None	 Line	 Arc	 Semicircle	 Square - Chamfer	 Trapezoidal	 Arc - Trapezoidal	 Step Form	 Irregular
Bottle Body (X_5)	 Line 1:1	 Line 1:1.2	 Line 1:1.5	 Line 1:2	 Arc	 Trapezoidal	 Drop		
Top Form (X_6)	 Square Chamfer	 Circle	 Oval	 Semicircle + Square	 Other				

D. Investigation of Consumer Preferences

The research process uses an intuitive way. First, it assumes that all samples are in the same condition to assess the degree of consumer preference for a product shape and willingness to purchase. We use the 7-scores intention scales with score one meaning absolutely dislike and seven as a favorite. Dislike-Like (D-L) is used to indicate the degree of user preferences. In the same way not want to buy-Want to buy (N-W) is used to indicate the degree of expectation to purchase. Subjects are asked to fill the form with the score of their feelings to the 26 sample. And then, we compute mean and standard deviation values according to the statistics. After discussion from the panel of experts, we use the first set of questions (D-L) for study. In everyday life, the degree of preference to a product deeply influences the willingness to purchase it, and also can be seen as an

important reference to explore what product form factors are more important to consumer preferences. Table 2 shows that the first line is the column name, X_1 to X_6 are morphological characteristic symbols. Line 3 to 28 show the 26 sample elemental codes, the degree of preferences, standard deviation and the mean values of expectation purchase.

E. Preliminary Analysis

Through a questionnaire investigation, value of the consumer preferences is got for each product. It can be seen from the results that the highest average value is 4.24 from the sample No.7. The second is No.9, whose average is 3.987. The lowest average value is 2.924 of the samples No.15, and the second lowest average value is 2.937 of No.24. By comparative analysis of the sample scores sorting and distribution of elements types, following rules can be taken.

TABLE 2 USER INVESTIGATION DATA

Sample no.	X1	X2	X3	X4	X5	X6	D-L value		N-W
							Average	Standard deviation	Average
1	3	2	4	2	6	1	3.684	2.066	3.506
2	4	3	4	2	5	2	3.620	2.096	3.430
3	1	2	3	8	3	2	3.051	1.954	2.873
4	2	2	5	1	4	2	3.418	2.048	3.342
5	3	4	2	4	6	3	3.684	1.885	3.430
6	1	2	1	6	3	2	3.532	1.973	3.380
7	1	1	2	1	7	2	4.241	2.260	4.038
8	4	2	4	3	4	2	3.266	1.708	3.101
9	3	3	1	6	2	1	3.987	2.016	3.873
10	3	2	2	3	4	4	3.443	1.824	3.342
11	2	5	3	5	3	2	3.481	1.846	3.405
12	2	2	2	7	2	1	3.532	1.920	3.316
13	2	2	3	6	1	1	2.987	1.891	2.962
14	3	2	3	3	2	1	3.608	1.793	3.557
15	3	2	3	3	1	3	2.924	1.810	2.810
16	2	2	1	9	2	4	3.076	2.093	2.975
17	3	2	3	3	4	2	3.241	1.726	3.203
18	3	2	5	7	3	4	3.291	1.848	3.000
19	2	1	3	3	2	3	3.633	1.988	3.443
20	3	6	3	4	2	1	3.722	1.881	3.481
21	3	2	3	7	3	1	3.684	1.871	3.456
22	3	3	2	2	2	5	3.013	1.829	2.924
23	3	1	1	2	1	1	3.266	1.886	3.114
24	2	2	4	3	1	3	2.937	1.727	2.785
25	3	2	4	7	2	4	3.506	2.075	3.392
26	2	1	3	6	3	1	3.937	2.090	3.722

- Cap ratio is not the key factor that determines consumer preference.
- Cap form greatly affects the consumers' preferences. Obviously, consumers dislike X2 whose top half is a short knob. However, this kind of cap is the mainstream product in the market.
- In a whole view, the bottleneck cannot attract consumer concerns, but what is worth of attention is that X3-5 bottlenecks get negative feedback from consumers. Perhaps this is where the designers should note.
- Bottle shoulder is an important part of the bottles. From data analysis, we found that the 7th sample without bottle shoulder get the highest score. In the case where the samples have the bottle shoulder, the statistics show that consumers dislike the asymmetry bottle shoulder. They prefer Semicircle (X4-4) and arc trapezoid (X4-7), but dislike have a curved shape (X4-3) of the bottle.
- Bottle body is the largest and most important visual component. It is clear from the data. We found that consumers prefer those bottle bodies whose ratio between length and width is 1:1.2 (X5-2). It can be seen through the analysis that consumers dislike 1:1 square bottle (X5-1) and 1:2 lanky linear bottle (X5-4). Consumers are not concerned about whether the bottles are straight or have some curvatures. In addition, the teardrop-shaped (X5-7) bottle gets the highest score. Such bottle is so rare on the market that we only found one sample. Researchers are not optimistic about such teardrop-shaped bottle. Because of the too small sample results, it cannot come to a conclusion. But with the results of consumer testing, more research is needed to understand the value of such bottle shape.
- Top view features determine the sense of thickness of a bottle, and reflects overall relationship between the front and side outline. As is obviously shown from the data results,

consumer prefer squared bottle with a small-sized fillet (X6-1), rather than a straight line + semicircle type (X6-4). For cylindrical, oval cylindrical ones, consumers express that it does not matter whatever. This shows that the current consumers prefer clean, simple designs with some details of lines, and like conventional flat-square bottle most.

Our aim is to further analyze several relationships based on Pearson product moment.

III. PEARSON CORRELATION COEFFICIENT ANALYSIS

Using Pearson correlation coefficient analysis, the relationship between the presented bottle styles and subject ratings has been computed by the following equation:

The proportion of the element styles was taken where the correlation was statistically significant at the significance level $p < 0.05$. Fig.2(a) and (b) demonstrates the statistical significance rating number of the 1st and 7th level rating respectively.

$$r_{xy} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}}$$

In general, the Pearson correlation coefficient analysis has shown the significant styles correlated with consumers' ratings. The style of each element can be obtained according to those coefficients. As it can be seen from Fig. 2, there are clearly different subcategories

The x-axis represents the subcategories of the design style and the y-axis is the significance rating number. In Fig. 2 (a), there are a few subcategories rated above 4, which are taken as the most related design styles based on the 26 given images. Fig. 2 (b) clearly illustrates the significant styles for each element.

IV. CONCLUSION

Currently, Chinese small-sized liquor market gradually booming. Using which bottle shape is a key issue that troubled companies and designers. Based on the test of consumer psychology preferences and Kansei engineering methods, this research focuses on the small-sized liquor bottle shape and design elements. The study comes into final conclusion:

1) Statistical analysis based on subject's rating can provide clues of important elements and styles in liquor bottle design. For example, bottle shoulder is an important design element while the ratio 1:1.2 (X5-2) for the length to width of the bottle body is preferable from the survey data.

2) Pearson correlation coefficient between the subject rating and sample style data can clearly show the preferable design styles for each element of the bottle. This provides the designer the significant design style from the point of view of subject.

These conclusions will provide an important reference in the development of Chinese liquor bottle design, especially in the choice of morphological language elements. In the future, more examinations on some other components will be taken that may affect consumer preferences, such as labels shapes, brands, words, illustrations, color and so on. Therefore, based on consumer preferences, a complete system can be built up for the research of visual design strategy of small-sized liquor. This method will provide important theoretical support for bottle product.

ACKNOWLEDGMENT

This project was supported by "the Fundamental Research Funds for the Central Universities" (2014JBW001, Beijing Jiaotong University, China). The authors also thank all the subjects in china for their participation and assistance in the experimental study.

REFERENCES

- [1] Chuang, MC, Chang, CC, Hsu, SH, "Perceptual elements underlying user preferences toward product form of mobile phones," International Journal of Industrial Ergonomics, vol.27, pp. 247–258, 2001.
- [2] Nagamachi.M., "Kansei engineering: a new ergonomics consumeroriented technology for product development," International Journal of Industrial Ergonomics , vol.15, pp. 3–11,1995.
- [3] Ishihara, S, Ishihara, K, Nagamachi, M, "An automatic builder for a kansei engineering expert system using self-organizing neural net- works," International Journal of Industrial Ergonomics , vol. 15, pp. 25–37,1995.
- [4] Lin, YC, Lai, . HH, Yeh, CH, "Consumer-oriented product form design based on fuzzy logic: a case study of mobile phones," International Journal of Industrial Ergonomics , vol.37, pp. 531–543,2007.

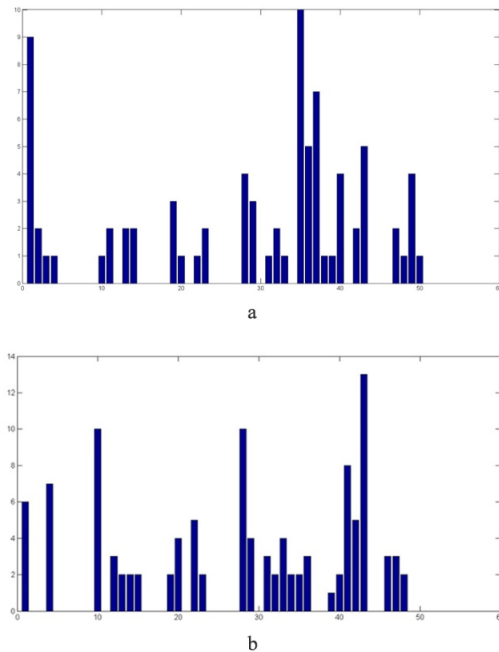


Fig. 2. Statistically significant design style (a) statistical significance rating number of the 1st level rating. (b) Statistical significance rating number of the 7th level rating rated as the significant design styles for these two rating levels.

- [5] Schutte. S, Eklund. J, "Design of rocker switches for work-vehicles-an application of Kansei Engineering," *Applied Ergonomics*, vol. 36, pp. 557–567, 2005.
- [6] Zhai. LY, Khoo. LP, Zhong. ZW, "A rough set based decision support approach to improving consumer affective satisfaction in product design," *International Journal of Industrial Ergonomics*, vol. 39, pp. 295–302, 2009.
- [7] Yang-Cheng Lin, Chung-Hsing Yeh, Chun-Chun Wei, "How will the use of graphics affect visual aesthetics? A user-centered approach for web page design," *International Journal of Human-Computer Studies* , vol. 71, pp. 217-227, 2013
- [8] Cross.M, *Engineering Design Methods*. Wiley, London. Deng, J.-L, "Control problems of grey system," *System and Control Letters* , vol. 1, pp. 288–294, 1994.
- [9] Hair. J, Anderson. R, Tatham. R, Black. W, "Multivariate Data Analysis," New York: Macmillan Publishing, 1995.

Invited Talk -----Towards Industry 4.0

Paulino Rocher

Manufacturing Technology Centre, UK

Forum Discussion: Design Knowledge Capture, Optimisation & Automation to Advance Industry 4.0

Discussion facilitated by Joo Hock Ang

Sembcorp Marine, Singapore

The Factory of the Future Production System Research

Milan Gregor¹, Jozef Herčko², Patrik Grznár³

CEIT, a.s., Slovakia¹, Department of Industrial Engineering, University of Žilina^{2,3}
Žilina, Slovakia

milan.gregor@ceitgroup.eu¹, jozef.hercko@fstroj.uniza.sk², patrik.grznar@fstroj.uniza.sk³

Abstract—This paper is dealing with a new research and development platform for the development of Factory of the Future Production System, referred to as ZIMS (Žilina Intelligent Manufacturing System). ZIMS is a new, open and collaborative environment, supporting creativity, inventing new solutions and their practical implementation in the form of new innovative products. ZIMS is based on holonic based, what brings many opportunities to develop new intelligent solutions for industry.

Factory of the Future; Žilina Intelligent Manufacturing System; Digital Factory

I. INTRODUCTION

The globalized economy is strongly influenced not only by economic cycles but also a rapid change in customer behavior, which result in turbulences. Business community should continually find new ways to respond to these incentives. One of the effective solutions is the use of reconfigurable manufacturing systems.

In the world runs intensive research into Factory of the Future Production System. The EU has launched extensive research programs dedicated to Factory of the Future and Intelligent Manufacturing Systems (IMS), Smart Manufacturing. The goal of all these efforts is to develop a new production system using advanced technology, which will enable to satisfy demanding customer requirements in the future.

Authors of the paper describe a new research and development platform for the development of Factory of the Future Production System, referred to as ZIMS (Žilina Intelligent Manufacturing System).

II. ŽILINA INTELLIGENT MANUFACTURING SYSTEM

A. ZIMS – New Research and Development Platform

Response to the latest trends in the area of Factory of the Future Production System is the emergence of new research platform – ZIMS. This research platform was created in cooperation of CEIT, a. s., Slovakia (Central European Institut of Technology) spin off the University of Žilina, Technical University of Košice, technological and industrial partners. ZIMS responds to trends with the digital transformation of Industry – or Industry 4.0 Europe [1].

ZIMS uses the most advanced technologies for the design, optimization and operation of Factory of the Future (FoF), especially in the area of: Digital Factory and Digital Engineering, Virtual Engineering, Reverse Engineering, digitization (3D laser scanning), Rapid Prototyping, virtual testing, computer simulation and emulation, etc.

The layout of workplaces in ZIMS (Fig. 1) is represented an area of more than 1000 m² (Fig. 2). As shown in Figure 1, the layout of workplaces in ZIMS was carried out on the basis of the logic of development of innovation. It starts with the idea and its presentation in Virtual Reality, Rapid Prototyping and product testing, design of production processes, configuration of the production system using Digital Factory Technologies It finishes with the practical realization of the product in the production system.



Figure 1. Layout of workplaces in ZIMS.

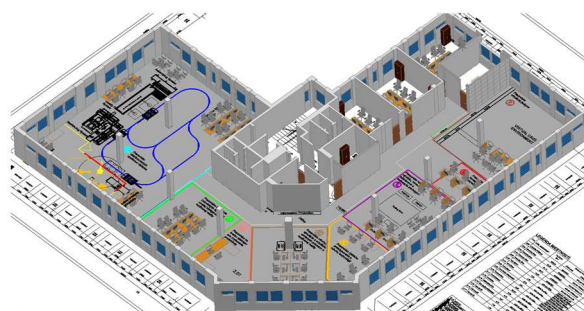


Figure 2. Concept of CPS in ZIMS.

ZIMS is a new, open and collaborative environment, supporting creativity, inventing new solutions and their practical implementation in the form of new innovative products.

This environment fully supports experimentation with new, unknown issue and search of non-traditional approaches to solving existing problems. ZIMS also serves as an incubator latest technology.

B. ZIMS – Cyber-Physical System

ZIMS, as a system is built in three different worlds: the real, the virtual and the digital (Fig. 3), where their interface is created a Cyber-Physical System (CPS) ensuring a direct integration of virtual, digital and real world [2].

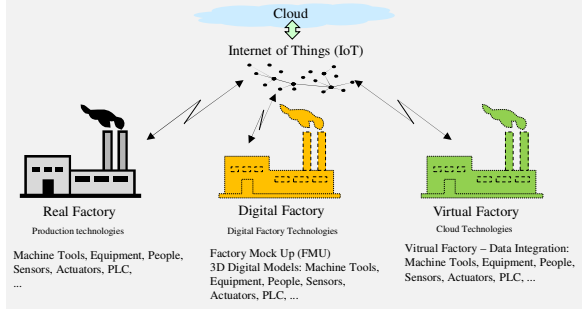


Figure 3. Concept of CPS in ZIMS.

Smart Factory – is built as an agile system that is able to adapt rapidly to changing customer requirements. Intelligent features, automation, and robotics, reconfiguration, automatic control, simulation and emulation technologies are used to create rapid change.

Virtual Factory – virtualization technology and data integration are used to represent the dynamics of real enterprise. Virtual Factory represents cyber feature of real enterprise and virtual representation of all its elements. It uses data from sensors, actuators, video and audio information, biometric data, etc. In real time creates a virtual image of functioning of enterprise.

Digital Factory – digitalization and digital technologies are used to integration of all activities within product life cycle and production systems. Digitizing, modeling, simulation and emulation are used to understanding of comprehensive manufacturing processes and creation of new knowledge, which is used for optimization of real production systems. In contrast to virtual factory, digital factory do not use real data, but use data for example from simulation.

C. ZIMS - Holonic Concept

ZIMS represents a pilot project of Intelligent Manufacturing Systems, which is composed of workplaces that communicate with each other through Holon [3]. Holons form comprehensive holarchy.

The strength of holonic organization, or holarchy, is that it enables the construction of very complex system that are nonetheless efficient in the use of resource, highly resilient to disturbance and adaptable to changes in the environment in which they exist [4].

The proposed Holonic concept of manufacturing system is used for control and monitoring of individual activities multi-agent system (MAS). The functioning of holonic systems are based on the use of autonomous ability of agent. The agents are considered to be autonomous

entities of system. Their interactions can be either cooperative or selfish within the defined level of action. Agents receive tasks from higher level of holarchy, but their solutions are carried out autonomously. Intelligent agent is a natural or computing system that is able to perceive their environment and on the basis of the monitoring carried out actions, which fulfilling the global objectives of the system.

Multi-agent systems (MAS) can be considered an elementary part of distributed artificial intelligence, which forms the conceptual framework for modeling of comprehensive systems. MAS is defined as a loosely bound network consisting of researchers of generated tasks. MAS platform represents distribution, autonomy, interaction (i. e. communication), coordination and organization of individual agents.

D. Knowledge environment - learning from the process

Fig. 4 represents the principle of building a knowledge environment that will encourage learning systems of active processes [2]. This approach is validated in research area of technology for the industrial production of large optical single crystals.

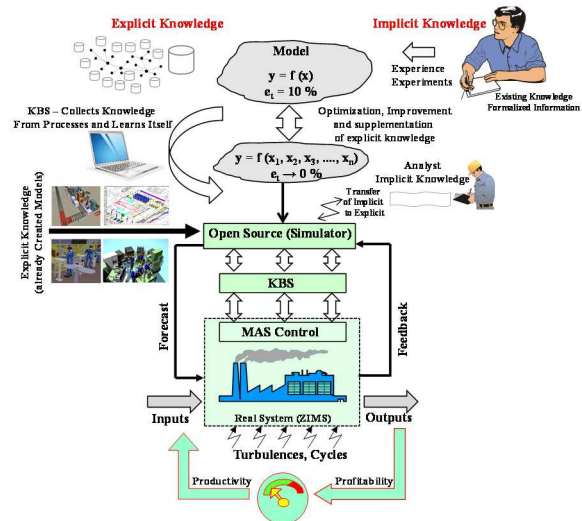


Figure 4. System of learning from the process.

Physical production system in ZIMS is controlled using a multi-agent system (MAS). When experimentation and development management for large-scale production of optical single crystals are used logic control system developed in ZIMS. It is based on a system of learning processes and uses meta-modeling approach. The control system communicates directly with the knowledge system. Knowledge-based systems used for decision support computer simulation (simulator). The simulator performs a set of simulation experiments. From the statistical data are obtained using multiple nonlinear regression analysis generated the desired meta-models. The resulting meta-models are used for making predictions about the future behavior of the controlled system (prediction) and these prediction amenities for Aproximativ production management. Aproximativ (gross) production management, represented by a set of metamodels is cyclically refined, using feedback (data) of real processes and new use of simulation. This creates a closed system of learning process, and its base is built and own knowledge

system. It is integrating the explicit knowledge (for example, models created in the past, the knowledge gained in the past, new explicit theoretical knowledge) and formalized implicit knowledge (conceptual system designers, analysts existing system).

III. INTELLIGENT, MODULAR PRODUCTION CONCEPT

Intelligent Modular Production is one of the workplaces that is built in ZIMS and represents a pilot project for intelligent, modular production solutions.

Intelligent Modular Production is research and experimental workplace, where are developed five subsystems:

- The Intelligent Modular System of Quality Control (InMoSys QC).
- The Intelligent Modular Assembly System (InMoSys AS).
- The Intelligent Modular Automated Guided Vehicle System (InMoSys AGV).
- The Intelligent Modular Robotic Machining System - InMoSys RMS-3.
- The Intelligent Modular Storage System (InMoSys ST) [5].

The principle of Intelligent Modular Production activities can be described simplistically on the basis of the logic of processing of a product.

It starts with the inputs from the input/output storage system InMoSys ST, from which the material is conveyed to the workplace InMoSys RMS-3. There is machined into the desired shape (volume, shape, texture, size, etc.). The processed part is transported by the InMoSys AGV to the quality control workplace InMoSys QC, wherein the control operations are carried out (precision, dimension, size, etc.).

High-quality parts are further transported by InMoSys AGV to assembly workplace InMoSys AS, which are shared with other parts already assembled into the final product. The final products are transported to the input-output storage system InMoSys ST.

IV. SYSTEM INNOVATION OF INTERNAL LOGISTICS

A. The Logistics Towing Units

Reconfigurable manufacturing systems are proposed (of several authors) as a solution to unpredictable fluctuations in market demand and market turbulence [6], [7], [8], [9], [10].

An example of an innovative development using the latest Internet of Things technologies is an autonomous logistics system developed in the framework of research ZIMS [11]. This system uses the logistics towing units Aurora, from the company CEIT (Fig. 5).



Figure 5. The logistics towing units of program Aurora.

These autonomous towing units were developed in the base of the requirements of the automotive industry, in cooperation with Volkswagen, Slovakia. The resulting solution is the Modular Reconfigurable Logistics System.

The following Fig. 6 shows the application of this logistics solution in the automotive industry. As is shown, the system uses automatic identification of towing units position, custom navigation, monitoring and control system, which is integrated to the Production Planning System.



Figure 6. The logistics concept of CEIT.

The following Fig. 7 shows the deployment of current system AURORA in the automotive industry.

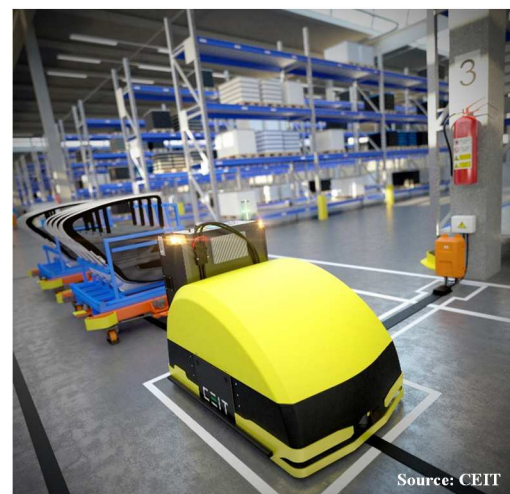


Figure 7. The logistics concept of CEIT.

The development of autonomous mobile robotic control system is implemented in its own development platform Ella® that uses virtual reality and digital models of individual elements of the proposed system. Dynamic verification of the functioning of the system in operating conditions is implemented through computer simulation and emulation (Fig. 8) in the environment of Ella SIM system [12].

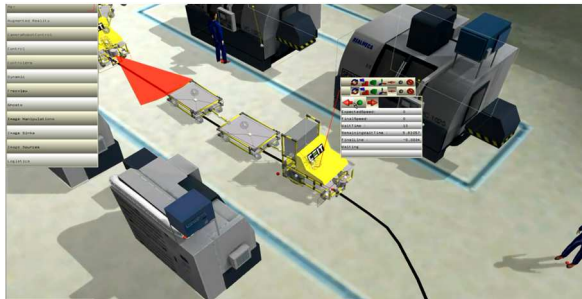


Figure 8. Simulation/emulation environment of Ella SIM system.

The preparation for implementation (Fig. 9) is performed off-line, in the environment of Ella VUP, in the system of virtual commissioning [13]. This approach allows in advance, supported by simulations to undertake predictive studies and optimization [14].

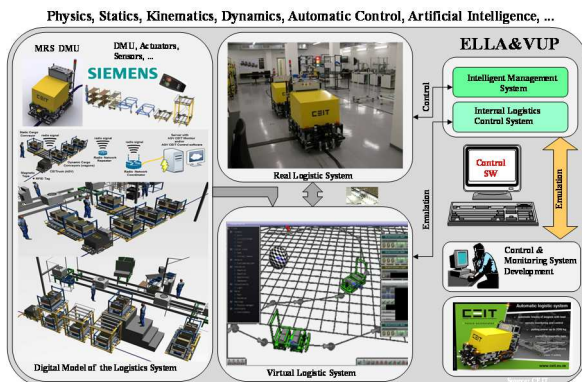


Figure 9. The virtual commissioning Ella VUP.

Digital Factory Technologies enabled the development of the company's own decision-making approach for analyzing potential for developing innovations (Fig. 10). The first step, it is evaluated the innovative concept and technological feasibility of the innovation (it finds out if: Does the innovation work?). In the second step, it is determined the significance of innovation, it is determined the market potential and market interest in innovation (to find out if: Does a customer buy this product?).

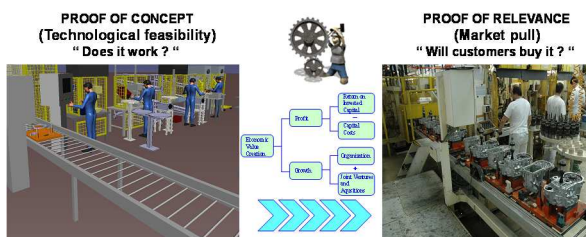


Figure 10. Decision-making on innovation.

One way to facilitate the design of the production system in virtual reality and simulation is to use our own software solution (Fig. 11), referred to as Virtual Design of Manufacturing Systems (VDMS). The trade name of VDMS is CEIT Table [15].



Figure 11. CEIT Table.

CEIT Table represents an integrated solution for support intuitive, team-oriented design of production and logistics systems.

It supports the productivity growth of the design process; it also fully supports the elimination of inefficient decisions in the process of preparation of innovation projects.

The customers have the opportunity to try out the activities of their Factory of the Future Production in the environment of virtual reality and augmented reality (CAVE - Computer Aided Virtual Environment). It was labeled as Adaptive Haptic Virtual Collaborative Development Environment (HVACDE) (Fig. 12).

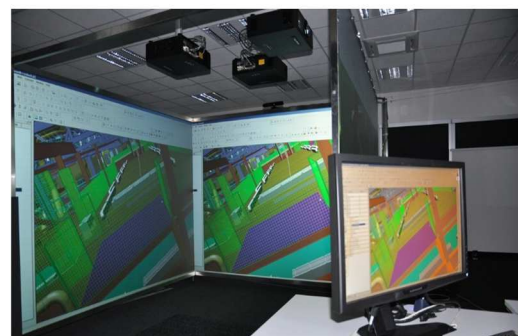
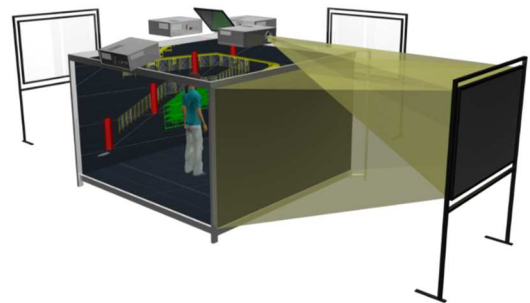


Figure 12. CAVE.

There are also technologies for haptic, special virtual reality headset for 3D dynamic effects (Oculus Rift), or the latest experimental technology for brainwave reading device (EPOC) (Fig. 13).



Figure 13. Oculus Rift and EPOC.

The latest development is oriented to the wider use of Digital Factory Technologies and advanced information and communication technologies [16]. All devices in developing logistics system will automatically monitor through sensors. Their current status (operation, failure, downtime, etc.) will be available to any other element of the production system.

In the development is the solution, in which each product will be carried (as one of the attributes) all the information, which will be required for processing in the base of the current status of the production system.

Intelligent pallets in developing manufacturing system will be equipped with its own processor and will be capable of optimizing of processing of their contents in production. Mobile robots will be fully autonomous and the order for transport will directly receive from the processed products or pallets.

Operation of all elements of the system will be constantly monitored and on the basis of this operation will be carried out predictions of potential fault conditions or interruption of system operation and will be immediately implemented some countermeasures.

An innovative approach to planning, evaluation of audits, deadlines and workshops in order to obtain outputs interactively, with room on-line display of the selected group of workplaces (Fig. 14).

There is possible to control not only the planned date, outputs of the analysis in tables and graphs but also to watch video recording or other desired plan activities.

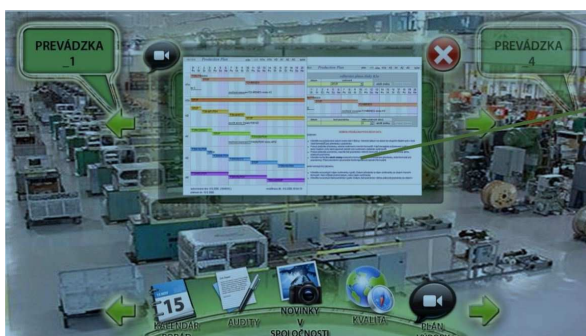


Figure 14. Innovative planning board.

Comprehensive simulation models whose input will be sensors of service consumption (energy, materials, etc.) will enable the optimization in real time.

The whole logistics system will be linked via the Internet of Things to cloud solutions so that all system information will be stored online and will be accessible in the cloud.

Operational interventions will be done after connecting to the internet from an arbitrary point in the world.

The enterprise will not need already undertaking its own IT solutions, servers and data repositories on first use of this solution, but will use Direct Memory Access Databases, which allow enormous acceleration of IT communications in industry.

Using the Data Centric Computing Deep Computing Architecture will run all important calculations at the point of data collection, which will further accelerate real-time communication. [16].

This research is funded by CEIT; spin off the University of Zilina and the Technical University of Košice.

Interface of controlled enterprises with social networks, households, public buildings, transport systems and vehicles is created a new environment – cyberspace.

This cyberspace can no longer function without artificial intelligence. For a description of such comprehensive systems will require special descriptive language. This language was developed in SRN – Unified Service Description Language (USDL).

V. CONCLUSION

Nature is as a comprehensive, self-organized, holonic system. A human is also made up of small, autonomous units - Holon, which create together larger self-organized, comprehensive units and this units form a comprehensive holonic system – a human.

Nature creates biological systems and enables their further development through evolution, towards the highest form of organized mass - Intelligent Systems.

Biological systems represent the most effective and efficient production systems that humanity knows. These systems serve scientists as role models in creating "artificial" mechanisms that imitate nature, for the production of new products.

Learning from Nature has become one of the most important resources for further development of humanity.

Biological systems represent the most effective and efficient production systems that humanity knows. These systems serve scientists as role models in creating "artificial" mechanisms that imitate nature, for the production of new products.

The Advanced Factory of the Future Production Systems may come increasingly closer to optimized biological systems by evolution and will use the latest scientific breakthroughs in artificial intelligence, nanotechnology and information and communication technology (ICT).

Nowadays, research teams are developing the Advanced Factory of the Future Production Systems in Zilina.

ACKNOWLEDGEMENT

This paper is the part of research supported by project VEGA 1/1146/12.

REFERENCES

- [1] T. Bauernhansl, M. ten Hompel, and B. Vogel-Heuser, (Hrsg.), "Industrie 4.0 in Produktion, Automatisierung und Logistik, Anwendung, Technologien, Migration," Wiesbaden: Springer Fachmedien, 2014, ISBN 978-3-658-04681-1, p. 634.
- [2] M. Gregor and Š. Medvecký, "CEIT 2030. CEIT - Technologické trendy do roku 2030," Žilina: CEIT, 2015, CEIT-Š002-03-2015, p. 103.
- [3] P. Marčan, "Holonický koncept inteligentných systémov," Produktivita a inovácie, Vol. 14, No. 6, 2013, ISSN 1339-2271, pp. 44-48.
- [4] V. Botti, and A. Giret, "Anemone: A multi-agent methodology for holonic manufacturing systems," London: Springer-Verlag, ISBN 978-1-84800-309-5, 2008, p. 214.
- [5] J. Rofár, "ZIMS – Žilinský inteligentný výrobný systém," Produktivita a inovácie, Vol. 14, No. 2, 2013, ISSN 1339-2271, pp.17-21.
- [6] A.I. Dashenko, "Reconfigurable Manufacturing Systems and Transformable," 2006.
- [7] E. Westkämper, and E. Zahn, "Wandlungsfähige Produktionsunternehmen. Das Stuttgarter Unternehmensmodell," Berlin: Springer Verlag, ISBN 978-3-540-21889-0, 2009. pp. 321.
- [8] Y. Koren, "The Global Manufacturing Revolution," New Jersey: John Wiley & Sons, 2010, ISBN 978-0-470-58377-7, pp. 399.
- [9] K. Mubarak, "The issues for the implementation of reconfigurable Manufacturing systems in small and medium Manufacturing enterprises," In ARIKA, ISSN 1978-1105, Vol. 4, No. 1, pp. 82-88.
- [10] M. Gregor, and M. Haluška, "Rekonfigurabilita holonického výrobného systému s podporou agentného prístupu," Produktivita a inovácie, Vol. 14, No. 6, 2013, ISSN 1339-2271, pp. 35-38.
- [11] M. Gregor, "ZIMS – Žilina Intelligent Manufacturing System. Nová iniciatíva Žilinskej univerzity a CEIT," Study CEIT-Š001-09-2011, 2011.
- [12] L. Ďurica, "Simulačné, emulačné a vizualizačné procesy v inteligentných výrobných systémoch," Žilina: Žilinská univerzita, 2015, p. 10.
- [13] M. Gregor, T. Michulek, and J. Rofár, "Virtuálne uvedenie do prevádzky – Virtual Commissioning," Produktivita a inovácie. Vol. 14, No. 4, 2013, ISSN 1339-2271, pp. 38-40.
- [14] Liu, Z., Suchold, N. and Diedrich, Ch., "Virtual Commissioning of Automated Systems," In INTECH: proceedings, Kongoli: CC, 2012, ISBN 978-953-51-0685-2, pp. 131-148.
- [15] M. Gregor, Š. Medvecký, and B. Mičjeta, "Žilina Intelligent Manufacturing System (ZIMS)," Žilina: CEIT. Study CEIT-Š001-05-2010, 2010, pp. 50.
- [16] M. Gregor, and M. Gregor, "Ľudský mozog ako počítač," ProIN, No. 16, Vol. 1, 2015, ISSN 1339-2271, pp. 32-40.

Industry 4.0 with Cyber-Physical Integration: A Design and Manufacture Perspective

Alfredo Alan Flores Saldivar¹, Yun Li¹, Wei-neng Chen², Zhi-hui Zhan², Jun Zhang², Leo Yi Chen³

¹ School of Engineering, University of Glasgow, Oakfield Avenue, Glasgow G12 8LT, U.K.

(Email: a.flores-saldivar.1@research.gla.ac.uk; Yun.Li@glasgow.ac.uk)

² School of Advanced Computing, Sun Yat-sen University, Guangzhou, China

³ School of Engineering and Build Environment, Glasgow Caledonian University, Glasgow, U.K.

Abstract— Foreseeing changes in the way companies manufacture products and provide services, future trends are emerging in design and manufacture. Together with growing internet applications and technologies connected through the cloud, a new Industrial Revolution, named “Industry 4.0”, aims to integrate cyber-virtual and cyber-physical systems to aid smart manufacturing, as presented in this paper. Connecting information and physical machinery, this new paradigm relies on how effective and fast connectivity are achieved for Industry 4.0. A new generation of wireless connection, 5G, will help and accelerate this trend. Following analysis of the present cyber-physical integration for the 4th Industrial Revolution, this paper also investigates future methodologies and trends in smart manufacturing, design and innovation.

Keywords- Cyber-physical integration; cyber-physical systems; Industry 4.0; smart manufacturing; networked autonomous production; CAD; CAutoD

I. INTRODUCTION

Design and manufacture are currently moving to a new paradigm, targeting innovation, lower costs, better responses to customer needs, optimal solutions, intelligent systems, and alternatives towards on-demand production. The concept that highlights this significant evolution is “Industry 4.0” (I4), dubbed the “4th Industrial Revolution” [1], with associated concepts of networked embedded systems, cyber-physical systems (CPS), smart factory, Internet of Things (IoT), Internet of Services (IoS) and “Internet+”, to name but a few. All these trends have in common the integration of several features in the same place as a response to challenges of computerized decision making and big data that are proliferated by the internet and cloud computing (CC).

To gauge the development and trends, this paper aims to analyze cyber-physical integration for design and manufacture, and to present a timely survey on Industry 4.0. Section 2 set the scene of Industrial Revolutions (IRs), with cyber-physical systems detailed in Section III. Necessary information and communication technologies (ICT) are analyzed in Section IV. Conclusions are drawn and future agendas are discussed in Section V.

II. INDUSTRY 4.0 – AN EVOLUTIONARY REVOLUTION

A. What Industry 4.0 Is

Ever since the beginning of industrialization, technological advances have led to socio-economic paradigm shifts which are

today termed “industrial revolutions”, i.e., mechanization with steam power for the 1st IR → electrical energy for mass production in the 2nd IR → automated production with electronics and control in the 3rd IR. Today, with advances in digitalisation and the internet, “smart manufacturing” and “smart factories” are becoming a reality, where the manufacturing value chain in the physical world can be integrated with its virtual copy in the cyberspace through CSP and IoT, and then be seamlessly integrated with IoS. Tempted by these future expectations, the term “Industrie 4.0” or “Industry 4.0” was coined *a priori* by the German government promoting their “High-Tech Strategy 2020 Action Plan” in 2013 for a planned “4th industrial revolution” [2]-[4].

The terminologies “Smart Industry” and I4 describe the same technological evolution from the microprocessor embedded manufacturing systems to the emerging CPS, smartly linking (i) demand to (ii) manufacture, (iii) supply, and (iv) services by the internet. Via decentralising intelligence, object networking and independent process management interact with the virtual and real worlds, heralding a crucial new aspect of future industrial production process that integrates the above four processes. In short, I4 represents a paradigm shift from “centralised” to “decentralised” production, a reversal of the logic of production process thus far. The design principles of I4 components are shown in Table 1 [4].

Table 1 Design Principles of I4 Components

Design	CPS	IoT	IoS	Smart Factory
Interoperability	X	X	X	X
Virtualisation	X	-	-	X
Decentralisation	X	-	-	X
Real-Time Capability	-	-	-	X
Service Orientation	-	-	X	-
Modularity	-	-	X	-

B. Importance of the Strategised Industry Revolution

The first three industrial revolutions came about as a result of centralization for production. Now, businesses will establish global networks that incorporate their machinery, warehousing systems and production facilities in the shape of a cyber-physical system, comprising “smart machines”, storage systems and production facilities capable of autonomously exchanging information, triggering actions and controlling each other independently.

A. A. Flores Saldivar is grateful to CONACYT for a Mexican Government scholarship. Y. Li and W.-N. Chen are grateful to the Royal Society and National Science Foundation of China for the support via a Newton Fund.

These form a “smart factory” that allows individual customer requirements to be met, whilst efficiency obtained in automated production is maintained. This means that even one-off items can be manufactured profitably. In Industry 4.0, dynamic business and engineering processes enable last-minute changes to production and offer the ability also to respond flexibly to disruptions and failures. End-to-end transparency is provided over the manufacturing process, also facilitating optimized design and decision-making. Further, Industry 4.0 will result in new ways of creating value and novel business models. In particular, it will provide start-ups and small businesses with the opportunity to develop and provide downstream services. To both developed and developing economies, I4 will reduce factory-floor requirements and help progress of humanity.

III. CYBER-PHYSICAL INTEGRATION

A CPS collaborates computational entities which are in intensive connections with their surrounding physical world and on-going processes, providing and using, at the same time, data-accessing and data-processing services available on the internet [5]. A cyber-physical production system relies on the newest and foreseeable further developments of computer science, ICT, and manufacturing science. Concepts like autonomous cars, robotic surgery, intelligent buildings, and implanted medical devices are just some of practical examples that have already emerged in Research and Developments (R&D) [6].

A. Design of a Cyber-Physical System

Cyber space and virtual systems represented by ICT are now getting integrated with physical control and production systems. This integration is enabling compression of development cycles by reuse of existing methodologies, methods, models, tools and techniques, encapsulated in integrated and customized models and components that can be rapidly used in an innovative or creative design. The unique challenges in CPS integration emerge from the heterogeneity of components and interactions. This heterogeneity drives the need for modelling and analyzing cross-domain interactions among physical and computational and networking domains, which demands deep understanding of the effects of heterogeneous abstraction layers in the design flow [7].

Figure 1 illustrates a well-funded approach to cyber-physical integration to meet design principles, mainly proposed in [8]. It highlights that analysis is a key issue in current CPS developments, integrating various objects, design methods and tools, aspect-oriented development methods and tools, multi-domain physical modelling methods and tools, and formal methods that address different aspects of the development process of CPS. Systems specification, modelling and design method integration involve many aspects of integration at different levels, including:

- Integration of the physical world dimension, communication dimension and computation dimension;
- Integrated object-oriented methodology, multi-domain methodology, aspect-oriented methodology and formal techniques;
- Integration of different design views;

- Integration of the methods used to specify and implement systems requirements;
- Integration of tools that support these methods;
- Integration of physical components and cyber components;
- Integration of different representations;
- Integration of the multiple specification fragments produced by applying these methods and tools; and
- Integration between informal specification methods and formal specification methods.

Model, Methodology and Tool Integration are detailed in the following sections. These aspects help investigate future directions and trends in Industry 4.0.

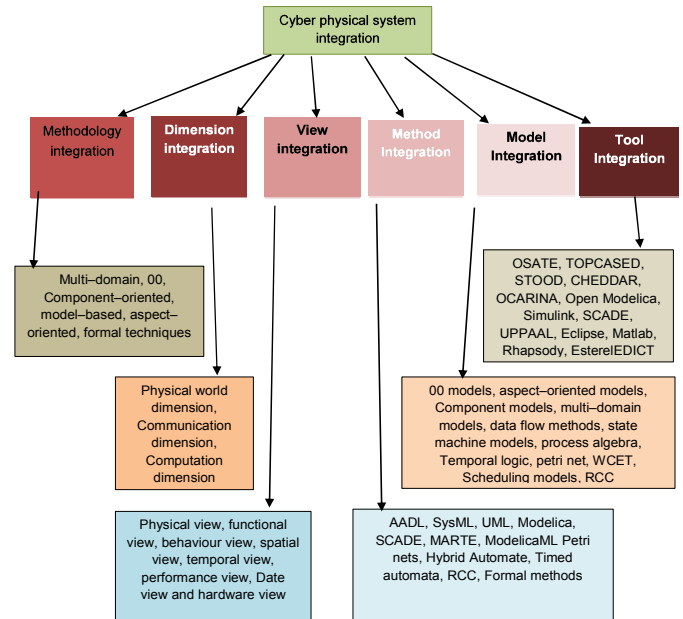


Figure 1. Integrated Approach to develop CPS.

B. Model Integration for Manufacturing-Aware Design Flow

As discussed, it is important to develop methodologies that integrate models, techniques, and tools that can be used in a design customized within its models and components. Components and models in a CPS are heterogeneous, spanning multiple disciplines (physical – thermal, mechanical, electrical, fluid,... and cyber – software, computing, cloud...). These require multiple models to represent the physical aspects, requirements, architectures, behaviours, spatial-temporal constraints, and interfaces, at multiple levels of abstractions [8].

Model and component-based design have been recognized as key technologies for radically changing productivity with CPS. Model-based design uses formal and sufficiently complete models, processes, their environments, and their interactions. The goal of a model-based design is “correct-by-construction”, where properties of the synthesized models of the designed system predict the properties of the implemented or manufactured system with sufficient accuracy [9]. As a result, an integrated tool suite called OpenMETA has been developed to provide a manufacturing-aware design flow, which covers both cyber and physical design aspects. A new integration

model for the OpenMETA suite was proposed in [9], as shown in Figure 2. Basically, the design flow is implemented as a multi-model composition/synthesis process that incrementally shapes and refines the design space using formal, manipulated models. It includes analysis and testing steps to validate and verify requirements and to guide the design process to achieve least complexity, and therefore the least risky and least expensive solutions. For example, the Adaptive Vehicle Make (AVM) [10] project, funded by the Defense Advanced Research Project Agency (DARPA), has constructed a fully integrated model and component-based design flow for the make process of a complex CPS.

- 3) Design Space Models (DSM) that define structural and architectural variabilities;
- 4) Test Bench Models (TBM) representing environment inputs, composed system models connected to a range of testing and verification tools for key performance parameters; and
- 5) Parametric Exploration Models (PEM) for specifying regions in the design space to be used for optimization and models for complex analysis flows producing results such as Probabilistic Certification of Correctness (PCC).

The first emphasis is placed on the development of a model integration language, so as to address heterogeneity to cover all relevant views of multi-physics and cyber domains and to achieve compositionality. Heterogeneity of the multi-physics, multi-abstraction and multi-fidelity design space, and the need for rapidly evolving/updating design flows require the use of a rich set of modelling languages usually influenced/determined by existing and emerging model-based design, verification and simulation technologies and tools. Consequently, the language suite and the related infrastructure cannot be static; it will continuously evolve. Then the second development is methodology integration in this framework [9]

C. Method Integration

To integrate modelling languages for CPS environments, mathematical models in this sense can bring together abstractions that are imported from individual languages and required for modelling cross-domain interactions. Proposed in [9], a language called CyPhyML is constructed as a lightweight, evolvable, composable integration language that is frequently updated and morphed. While these DSMLs may be individually quite complex (e.g., Modelica, Simulink, SystemC, etc...) CyPhyML is relatively simple and easily evolvable. This “semantic interface” between CyPhyML and the domain specific modeling languages (DSML) shown in Figure 3 is formally defined, evolved as needed, and verified for essential properties (such as well-formedness and consistency) using the methods and tools of formal meta-modelling. By design, CyPhyML is moving in the opposite direction to unified system design languages, such as SysML or AADL. Its goal is specificity as opposed to generality, and heavy weight standardization is replaced by layered language architecture and specification of explicit semantics.

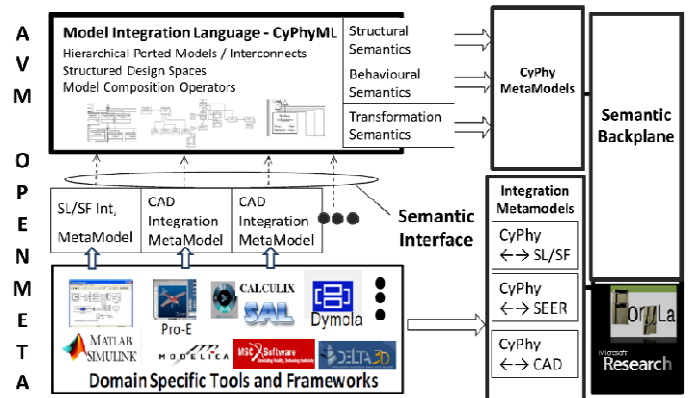


Figure 3. Method Integration Framework.

Figure 2. Model Integration: OpenMETA framework.

The main procedures of this design flow sketch the following phases:

- 1) Currently combinatorial or future intelligent design space exploration and architecture evaluation;
- 2) Behavioural design space exploration by progressively deepening from qualitative discrete behaviours to precisely formulated relational abstractions and to quantitative multi-physics, lumped parameter hybrid dynamic models using both deterministic and probabilistic approaches;
- 3) Geometric/structural design space exploration coupled with physics-based nonlinear finite element analysis of thermal, mechanical and mobility properties; and
- 4) Cyber design space exploration (both hardware and software) integrated with system dynamics.

As highlighted in Figure 2, these elements reflect the drive train challenge for the AVM development, emphasizing the 3D/CAD tools and finite element analysis for verifying blast protection and hydrodynamic requirements remaining on the basic structure of the integration architecture. In synthesis, the modelling functions of the OpenMETA design flow are built on the following model types [9]:

- 1) AVM Component Models (ACM) with standard, composable interfaces;
- 2) Design Models (DM) that describe component architectures and related constraints;

integration. The key is to define the structural and behavioural semantics of the CyPhy model integration language using formal meta-modelling, a tool support formal framework for updating the CyPhy metamodels and verifying its overall consistency and completeness as the modeling languages evolving. In this case, the meta-modelling tool FORMULA from Microsoft Research can be used. This tool is sufficient for defining mathematically modelling domains, transformations across domains, as well as constraints over domains and transformations, since they are algebraic data types (ADTs) and constraint logic programming (CLP) based semantics.

In Figure 2, it is observed as part of the Model Integration that a large suite of modelling languages and tools for multi-physics, multi-abstraction and multi-fidelity modelling are included; OpenModelica, Dymola, Bond Graphs, Simulink/Stateflow, STEP, ESMOL and many others software that are useful for analysis. At the end CyPhyML model integration language provides the integration across this heterogeneous modelling space and the FORMULA - based Semantic Backplane provides the semantic integration for all OpenMETA composition tools [10] The next step in this case, will be to outline Tool Integration according to Development 3, in which execution integration is also provided as a whole platform.

D. Tool Integration

Using the same approaching to [9-11], in which the Tool Integration Framework of the OpenMETA incorporates a network of model transformations that include models for individual tools and integrate model-based design flows, model-transformations are used for the following roles:

- 1) **Packaging:** Models are translated into a different syntactic form without changing their semantics. Taking the development, AVM Component Models and AVM Design Models are translated into standard Design Data Packages (Figure 2, .ACM and .ADM files) for consumption by a variety of design analysis, manufacturability analysis and repository tools.
- 2) **Composition:** Model- and component-based technologies are based on composing different design artefacts (such as DAE-s for representing lumped parameter dynamics as Modelica equations, input models for verification tools, CAD models of component assemblies, design space models, and many others) from appropriate models of components and component architectures.
- 3) **Virtual prototyping:** Several test and verification methods (such as Probabilistic Certificate of Correctness – PCC) require test benches that embed a virtual prototype of the designed system executing a mission scenario in some environment (as defined in the requirement documents). It is found that distributed, multi-model simulation platforms are the most scalable solution for these tests. The author selected the High Level Architecture (HLA) as the distributed simulation platform and integrated FMI Co-Simulation components with HLA.
- 4) **Analysis flow:** Parametric explorations of designs (PET), such as analyzing effects of structural parameters (e.g. length of vehicle) on vehicle performance, or deriving PCC for performance properties frequently require complex analysis flows

that include a number of intermediate stages. Automating design space explorations require that Python files controlling the execution of these flows on the Multidisciplinary Design Analysis and Optimization (OpenMDAO6) platform (that are currently use in OpenMETA) are auto-generated from the test bench and parametric exploration models (Figure 2).

It is highlighted through this development, that both multi-physical and computation modelling present advantages choosing the specification of composition, i.e. for physical interactions; power flow oriented modelling (Modelica, Simscape or Bond Graph modelling languages) requires spending time typing for expressing and enforcing connectivity constraints achieving safe modelling of multi-physics interactions.

The OpenMeta model and tool integration technology needs and infrastructure for creating and executing complex analysis flows. Based on “software-as-a-service” aspect of this development, it allows end users (individuals, research groups, and large companies) to repositories, analytic services and design tools to lower the costs, and excluding the high costs of acquiring and maintaining desktop engineering tools. In Figure 4 is presented the platform for executing the part of tool integration, according to [9].

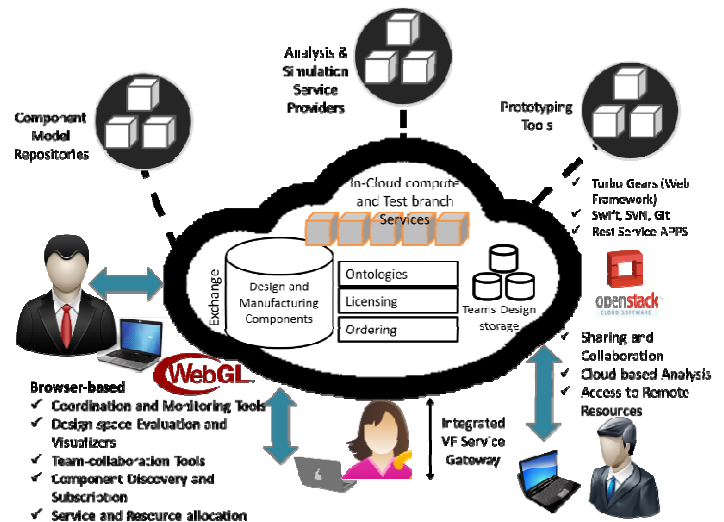


Figure 1. Tool Integration Framework.

The key fundamentals addressed in this platform are:

- 1) Resource elasticity and service staging, to address scalability and system wide optimization.
- 2) Managing the evolution of data, data presentations and their use by service and service integration over time.

With these fundamentals, it is clear that another matter need to be addressed for this platform, as shown in Figure 4. The evolution of data is key to better results and performance of this development.

IV. ICT INFRASTRUCTURE FOR INDUSTRY 4.0

A. Decentralized Computing

One of the main barriers to Industry 4.0 is decentralized systems, as described in [12], for allocating hardware and software resources to individual workstation or desired

location, which have been designed and operated with limited knowledge of the complete system. However, decentralized computing is now a trend in business environments. As there is no central decision maker, the information flow stays mostly local. Viewed as part of the control system, decentralization forms a self-organizing emergent system, with self-adaption, self-management, and self-diagnosis, keys in autonomic computing as well as Industry 4.0. This allows control of more complex systems, although global optimal performance cannot be guaranteed as firmly as with global controllers [12].

For implementing Smart Manufacturing schemes, ICT known as unified communications and integration of telecommunications, computers, necessary software, storage, and all those interfaces that allow users within the systems to access, store, transmit and manipulate information, must be addressed as well. As the number of users increases and the cloud computing grows in scale, decentralised cloud solution can play an importance role of infrastructure [13]. Decentralised intelligence can keep information and communication between the system components. By simulation and virtual design, manufacturing can be improved using optimization and control tools to set system scenarios. However, distributed optimization, owing to explosion in size and complexity of modern datasets, comprises the importance of solving problems and analysis.

Networked production leading to what Industry 4.0 is aiming to achieve, constantly collects Big Data sets, those data sets possess characteristics of being extremely large, high-dimensional and data stored or collected in a distributed manner. All those characteristics can also be noted as part of the machine learning process. As a result, it has become of central importance to develop algorithms that are both rich enough to capture the complexity of modern data, and scalable enough to process huge datasets in a parallelized or fully decentralized fashion [14].

B. Cloud Computing

Through resources virtualization, cloud computing provides infrastructure, platform and software as services. “Cloud computing is a model for enabling ubiquitous, convenient, on demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction” [6]. This cloud model is composed of five essential characteristics, three service models, and four deployment models for optimal scheduling of resources [15].

For an I4 environment, private clouds, where data are restricted to the usage of a company, are imperative as a service linking between the private networks and the public ones, with a well-funded policy of data sharing and data accessing. The linkage between cloud computing and automation systems is strong. Nowadays, developments on both sides are getting involved, so as to achieve higher efficiency with less effort. With cloud computing, CPS are being optimized to reach manufacturing by the press of one button [16]. The gathering of knowledge and perspectives with the proposal of architecture for Global Information are showed in Figure 5.

Facing this paradigm, the integration of CPS, IoT, IoS and CC leads to a transformation, more than a revolution. Those subjects already mentioned before comprise the trend and need to implement Industry 4.0, which gives the way of reorganizing the modes of production using existing tools and placing

reliance on networks. With the mergence between the Internet and factories, Industry 4.0 is characterized by the constant communication and linkage among production, supply chain, tools and workstations via the Internet and virtual networks.

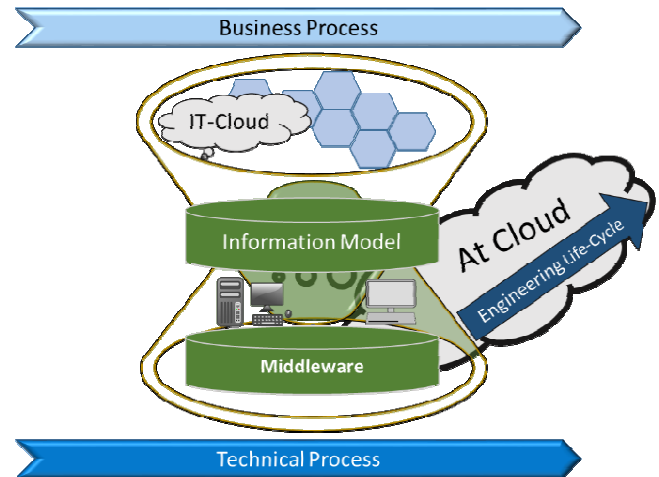


Figure 2. Architecture for Global Information with use of Cloud Computing.

C. Model-Based Integration

Considering Integration as part of the whole-system automation, it comes another topic of interest: “embedded-automation systems”, those which refer to control systems encompassing both domains (control-dominated and data-dominated) and which can be designed using one unified approach. This approach was proposed in [17], where the control-dominated parts deal with asynchronous inputs from an external environment, while the data-dominated parts process these events by calling appropriate functions.

While trying to reduce IT costs, improve efficiency of sharing processes and enhance scalability inside companies’ data and applications integration, there are certain levels of integrations known as: presentation level integration, business process integration (Service Oriented Architecture), data integration and communications-level integration. Depending on a company’s particular needs, communication can be either synchronous, asynchronous or a combination of both. In synchronous communication, a sender application sends a request to a receiver application and must wait for a reply before it can continue with its processing. Then, in asynchronous communication, a sender application sends a message to a receiver application and continues its processing before receiving a response. Asynchronous communication allows the loose coupling of applications, eliminating the need for connection management. This results in an applications integration solution that is more flexible, agile, and scalable - essential attributes denoting what Industry 4.0 seeks for. For designing an integration solution, asynchronous communication offers a number of advantages over synchronous communication, especially when it comes to services in a Service Oriented Architecture. In synchronous communication patterns, timeouts are more common when an application has to wait for responses from several other applications. This means that the availability of services increases since individual processes are not blocked out as frequently due to waiting on other sub-processes to complete [17].

At present, a model-based integration approach is in its infancy and requires significant future research efforts. Many researchers agree that modelling from the CPS is a sizeable obstacle for companies that handle big data and obtain any profitable analysis for prediction. It has been suggested to tackle uncertainties within the data analysis. Tool integration and support from model-based systems and rapid construction of domain-specific tool chains are also suggested from present research [16].

D. Virtual Prototyping with Computer-Automated Design

Utilizing Evolutionary Computation, CAutoD accelerates and optimizes the tedious process of trial-and-error by reversing a design problem into a simulation problem, then automating such digital prototyping by intelligent search via biological-inspired machine learning [18]. Experimental research in order to validate scientific results of the theoretical work is also what researchers suggest. Validation and implementation of this approaches help with a fast rhythm of acquiring knowledge and developments. What is trending now will not be the same in a few more years' time. When launching projects like smart manufacturing and Industry 4.0, companies should stay one step ahead and put efforts on innovative resources for advanced results.

V. DISCUSSION AND CONCLUSION

As stated so far, there exist challenges and future directions when tackling the subject of Industry 4.0, as argued in [3]. These include general reluctance to change by stakeholders, threat of redundancy of corporate IT departments and a lack of adequate skill-sets to expedite the march towards the 4th Industrial Revolution.

Many other trends have developed for Smart Manufacturing, not only in Germany with Industry 4.0, but also in the United States such as the Smart Manufacturing Leadership Coalition (SMLC). What SMLC presents is the infusion of intelligence that transforms the way industries conceptualize, design and operate the manufacturing enterprise [19]. Both perspectives agree on what challenges have to overcome in order to achieve what they pursue, such as analysis of big data-information, interoperability and scalability, among others.

So far, smart manufacturing approaches, analysis, virtualization and the new tendencies like the Industry 4.0 and big data studies have been studied. Summarizing the related work and developments leads to focus on the aspects facing Industry 4.0, such as methodologies that integrate collaborative systems. In this case, researchers suggest that a well-funded methodology that integrates CPS, cloud computing, virtual designs and real-time analysis is key to achieving innovation and a high productivity, because the system at the end becomes self-aware and self-predictive among other properties that are suitable for future research.

REFERENCES

- [1] Lee, J., H.-A. Kao and S. Yang (2014). "Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment." *Procedia CIRP* 16(0): 3-8.)
- [2] Lasi, H., P. Fettke, H.-G. Kemper, T. Feld and M. Hoffmann (2014). "Industry 4.0." *Business & Information Systems Engineering* 6(2): 239-242.
- [3] Kagermann, H., W. Wahlster and J. Helbig (2013). "Recommendations for implementing the strategic initiative INDUSTRIE 4.0." ACATECH NATIONAL ACADEMY OF SCIENCE AND ENGINEERING(3).R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
- [4] MacDougall, W. (2014). "INDUSTRIE 4.0 SMART MANUFACTURING FOR THE FUTURE." *MECHANICAL & ELECTRONIC TECHNOLOGIES, GERMANY TRADE & INVEST*(4): 40.
- [5] Monostori, L. (2014). "Cyber-physical Production Systems: Roots, Expectations and R&D Challenges." *Procedia CIRP* 17(5): 9-13.
- [6] Technology, N. I. o. S. a. (2013). "Strategic R&D Oportunities for 21st Century Cyber-Physical Systems." *Foundations for Innovation in Cyber-Physical Systems Workshop*(6): 32.
- [7] Sztipanovits, J., X. Koutsoukos, G. Karsai, N. Kottenstette, P. Antsaklis, V. Gupta, B. Goodwine, J. Baras and W. Shige (2012). "Toward a Science of Cyber–Physical System Integration." *Proceedings of the IEEE* 100(7): 29-44.
- [8] Lichen, L. (2015). *Model Integration and Model Transformation Approach for Multi-Paradigm Cyber Physical System Development*. Progress in Systems Engineering, H. Selvaraj, D. Zydek and G. Chmaj, Springer International Publishing. 330: 629-635.
- [9] Sztipanovits, J., T. Bapty, S. Neema, L. Howard and E. Jackson (2014). *OpenMETA: A Model- and Component-Based Design Tool Chain for Cyber-Physical Systems*. From Programs to Systems. The Systems perspective in Computing, S. Bensalem, Y. Lakhneck and A. Legay, Springer Berlin Heidelberg. 8415: 235-248.
- [10] Lattmann, Z., A. Nagel, J. Scott, K. Smyth, C. vanBuskirk, J. Porter, S. Neema, T. Bapty, J. Sztipanovits, J. Ceisel and D. Mavris (2012). *Towards Automated Evaluation of Vehicle Dynamics in System-Level Designs*. ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. A. S. M. ENGINEERS: 1131-1141.
- [11] Eremenko, P. (2011). *Philosophical Underpinnings of Adaptive Vehicle Make*. DARPA-BAA-12-15. Appendix 1 (December 5, 2011), Arlington, VA.
- [12] De Wolf, T. and T. Holvoet (2003). *Towards autonomic computing: agent-based modelling, dynamical systems analysis, and decentralised control*. Industrial Informatics, 2003. INDIN 2003. Proceedings. IEEE International Conference on.
- [13] Jun, C., W. Xing, Z. Shilin, Z. Wu and N. Yanping (2012). *A Decentralized Approach for Implementing Identity Management in Cloud Computing*. Cloud and Green Computing (CGC), 2012 Second International Conference on.
- [14] Boyd, S., N. Parikh, E. Chu, B. Peleato and J. Eckstein (2010). "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers." *Foundations and Trends® in Machine Learning* 3(13): 1-122.
- [15] Andreadis, G., G. Fourtounis and K.-D. Bouzakis (2015). "Collaborative design in the era of cloud computing." *Advances in Engineering Software* 81(11): 66-72.
- [16] Givehchi, O., H. Trsek and J. Jasperneite (2013). *Cloud computing for industrial automation systems- A comprehensive overview*. Emerging Technologies & Factory Automation (ETFA), 2013 IEEE 18th Conference on.
- [17] Yoong, L., P. Roop, Z. Bhatti and M. Y. Kuo (2015). *Introduction to Synchronous Programming Using Esterel. Model-Driven Design Using IEC 61499*, Springer International Publishing: 35-64.
- [18] Bukata, B. B. and L. Yun (2011). *Reviewing DSTATCOM for smart distribution grid applications in solving power quality problems*. Automation and Computing (ICAC), 2011 17th International Conference on.
- [19] Swink, D. (2014). *Smart Manufacturing Leadership Coalition (SMLC)*. OSI-SOFT, SAN LEANDRO, CA.

Invited Talk ----- Supercomputing for Industry 4.0

Richard Martin

ARCHIE-WeSt High-Performance Computing Centre, UK

Determination of the material constants of creep damage constitutive equations using Matlab optimization procedure

Javier Zamorano Igual
School of Computing and Engineering
University of Huddersfield
Huddersfield, United Kingdom
U1373763@unimail.hud.ac.uk

Dr Qiang Xu
School of Computing and Engineering
University of Huddersfield
Huddersfield, United Kingdom
Q.Xu2@hud.ac.uk ORCID iD [0000-0002-5903-9781](https://orcid.org/0000-0002-5903-9781)

Abstract— Creep damage constitutive equations based in continuum damage mechanics are characterized by their complexity due to the coupled form of the multi-damage state variables over a wide range of stresses. Thus, the determination of the material constants involved in these equations requires the application of an optimization technique. A new objective function was designed where the errors between the predicted and experimental normalized deformation and lifetime were used in conjunction of the minimal nonlinear least square method from Matlab. Its use is simpler, more compact, and less uncertain and is able to obtain an accurate solution for a sample material (0.5Cr 0.5Mo 0.25V ferritic steel) at the range of 560-590°C. The specific experimental data, the material constants, and all the factors needed are provided as a comparison with the existent investigation of this material. Future works should aim at to further establish the reliability and user-friendliness of the method.

Keywords: Creep constitutive equations, ferritic steel, material constants, optimization, Matlab

I. INTRODUCTION

Ferritic steel alloys are extensive utilized for a welded steam pipes in the assembly of power plant components operating under a critical conditions where the creep deformation and possible failure are significant in the design factors requirements such as strain histories, damage field evolution and lifetimes. Continuum damage mechanics describes the creep behavior using physically based creep damage constitutive equations [1, 2]. These equations are developing into more elaborated because new state variables are introduced to describe more accurately the deformation and the damage mechanisms [3, 4]. The accurate determination of the material constants involve in constitutive equations utilizing the experimental data for a range of temperatures and stresses is a challenging and difficult task according to [5, 6].

In the past decades many researchers have investigated this issue, and commonly optimization procedure are utilized to determinate the constants, by applying the minimal least square method to an objective function which compute the errors of simulated and experimental data. Methods were developed for the creep damage [1, 4], and viscoplasticity model [7, 8]. The optimization routines of these approaches need a set of careful chosen starting values in order to achieve global convergence. To solve this problem Lin & Yang [6], and Li, Lin, & Yao [5] developed a global optimization method for superplasticity and creep

damage, respectively, using genetic algorithms, which do not need a good starting value for a correct convergence, whereas, the difficulty to implement the objective function is increased considerably, moreover, a higher understanding of complex program code routines are needed.

Gong, Hyde, Sun, & Hyde [7] developed a simple optimization program for determining the material parameters in the Chaboche unified viscoplasticity model, using Matlab. In this case the optimization routine seeks for the global minimum of the difference between the square sum of the predicted and experimental stresses. Runga-Kutta-Fehlberg algorithm was used to solve the ODE's of the model, and the Matlab optimization toolbox function, '*lsqnonlin*' which implement the Levenberg-Marquardt algorithm for each iteration step, was used to solve the nonlinear least square optimization.

Kowalewsky, Hayhurst, & Dyson [2] generated a satisfactory three-stage procedure to estimate the initial estimation of the material constants of the constitutive equations for an aluminum alloy. This equations can be related to the different parts of the creep curve, then, working out them, it can be found a good enough first guess. Later on, a general optimization process is used to estimate the final values. Similarly, Mustata & Hayhurst [1] developed a methodology for a 0.5Cr 0.5Mo 0.25V ferritic steel. The objective function utilized for the optimization is separated in three parts. First, the strain estimated and compared with the experimental, separating, each stage of the curve with a scaling factor, second, a time term with amplification factor, and third, a penalty function with the minimum strain rates. This objective function is significantly complex, the values of the several scaling factors are not given, resulting in uncertainty in its generic application. Furthermore, both approaches utilized a NAG numerical library in FORTRAN to implement the optimization routine, which is not as easily available as Matlab.

This paper reports the determination of the material constants for a set of creep damage constitutive equations, a similar approach of [7] for the viscoplasticity model. It is featured by the design of new objective function where both the differences of creep strain and the time between experimental and prediction are normalized including a weighting function.

II. OBJECTIVES

The main objective of this paper is to develop a general optimization procedure, using Matlab, to calibrate the material constants of the CDM-based creep constitutive equations for 0.5Cr 0.5Mo 0.25V ferritic steel. The program developed has to be able to reproduce the behavior of the creep mechanics of this material operating at high temperatures.

III. CONSTITUTIVE EQUATIONS

The hardening and softening mechanisms and the initiation and growth damage of the ferritic steel alloy are expressed by CDM-based constitutive equations. The uniaxial from proposed by Dyson, Hayhurst, & Lin [9] for a constant temperature is given by the following set of equations:

$$\frac{d\varepsilon}{dt} = A \sinh \left[\frac{B\sigma(1-H)}{(1-\Phi)(1-\omega)} \right] \quad (1)$$

$$\frac{dH}{dt} = \frac{h}{\sigma} \frac{d\varepsilon}{dt} \left(1 - \frac{H}{H^*} \right) \quad (2)$$

$$\frac{d\Phi}{dt} = \frac{K_c}{3} (1-\Phi)^4 \quad (3)$$

$$\frac{d\omega}{dt} = C \frac{d\varepsilon}{dt} \quad (4)$$

where the state variables represents, Φ , the coarsening of the carbide precipitates, the variable changes from zero to one, ω , the intergranular creep constrained cavitation damage, and also varies from zero (no damage state) to ω_f (failure), and, H , the strain hardening effect, in the beginning, it is zero and increases to a boundary value H^* at steady-state creep. A , B , C , h , H^* and K_c are material constants to be calibrated with the optimization method, ε is the deformation, and σ is the stress applied to the material. The material constant can be related to difference stages of the creep curve [3]: 1) h and H^* describe the primary stage, where is produced the hardening process, 2) A and B model the secondary stage, strain rate remains almost constant, 3) C and K_c describe the last stage of the curve, where are localized the damage mechanisms.

IV. OPTIMIZATION METHOD

The identification of a material constants in the CDM-based creep constitutive equations is a reverse process based on experimental data. A nonlinear least square optimization procedure is adopted. The primary aim is to find the value for the material constants which produce a global minimum of an objective function which basically simulate the difference between the predicted and experimental deformation under different stress levels at the same temperature.

$$f(b) = \sum_{i=1}^m \left[\sum_{j=1}^n \left(\varepsilon(b)_j^{\text{pred}} - \varepsilon_j^{\text{exp}} \right)^2 \right] \quad (5)$$

$b \in R^n$; $LB \leq b \leq UB$ where $f(b)$ is the basic objective function, b is the optimization variable set (a vector of n -dimensional space, R^n), which for this specific case are material constants on the CDM-based creep constitutive

equations, $b = [A, B, H^*, h, K_c, C]^T$, LB and UB are the lower and upper boundaries of b allowed during the calibration, $\varepsilon(b)_j^{\text{pre}}$ and $\varepsilon_j^{\text{exp}}$ are the model predicted total strain and the experimental measured strain, respectively, at a specific time j within the loop of maxim n , i is the specific curve used in the optimization for m number curves with different stress levels.

During the calibration of the boundary constraints has been noticed that for the material at 560°C the upper boundary for the constant A , has to be fixed on $1.00e-9 \text{ h}^{-1}$ for an accurate solution. For the other parameters, it was left a range of variance around them. The values can be seen in the Table 1.

A. Numerical Techniques

The prediction of the creep deformation at specific temperature and stress can be achieved by integrating the set of ODE's for a set of identify material constant vector b . From the set (1) to (4) a first order non-linear system with four differential equations with four variables $x = [\Phi, \omega, H, \varepsilon]^T$ can be identified. Solving the ODE's system by a numerical method such as Runge Kutta-Fehleberg algorithm can be estimated the creep damage characteristics (deformation, lifetime, and rupture strain). The Runge-Kutta-Fehleberg algorithm uses a pair of Runge-Kutta methods to obtain both the computed solution and an estimate of the truncation error [10]. Matlab has a command named as '*ode45*' which implement this algorithm directly, it is needed only to specific a range time, initial values for the variables, and a tolerance for the solution [11].

The nonlinear least square optimization algorithm applied here was used satisfactorily by Gong, Hyde, Sun, & Hyde [7], the Levenberg-Marquadt which in Matlab is implemented in the '*lsqnonlin*' command. This function ask for a vector valued function as input:

$$f(b) = [f_1(b) \ f_2(b) \ \dots \ f_n(b)] \quad (6)$$

where b is a vector of the unknown values to be estimated, and $f_n(b)$ are the vectors of the objective function [11]. The output of this command can be represented mathematically as the following nonlinear least square equation:

$$\min_b \|f(b)\|_2^2 = \min_b [f_1(b)^2 + f_2(b)^2 + \dots + f_n(b)^2] \quad (7)$$

where the variables represent the same as in the previous equation.

B. Experimental Data

Experimental data of the uniaxial creep curves from [1] were digitized and shown in Fig.1, and Fig.2 schematically, and numerically in the Table 2, and Table 3. A lack of data is observed from the experimental tests, thus extra points were interpolated for a curve fitting purpose, which were also shown in the Figures and Tables. The new data is represented as dots, and clearly, it can be seen that, specially, for the 85 MPa curve of the material at 560 °C the rebuilt is needed because the data in the primary and tertiary stage is insufficient.

The experimental lifetimes for the material at 560 °C are 91000, 51900, 31111 hours, for 85, 100, 110 MPa, respectively [1]. For the 590 °C are unknown, thus, they were estimated from the curves being 5100, 2700, 1400 hours, for 100, 110, 120 MPa, respectively.

Table 1. Boundary constraints for the material constants

Boundary Constraints	Material at 560°C		Material at 590°C	
	LB	UB	LB	UB
A (h ⁻¹)	1.00E-10	1.00E-09	1.00E-10	5.00E-09
B (MPa ⁻¹)	1.00E-01	2.00E-01	1.00E-01	2.00E-01
H* (-)	4.00E-01	8.00E-01	3.00E-01	8.00E-01
h (MPa)	1.00E+04	2.00E+05	1.00E+04	3.00E+05
K _c (h ⁻¹)	1.00E-06	3.00E-05	1.00E-06	1.50E-04
C (-)	3	10	2	8

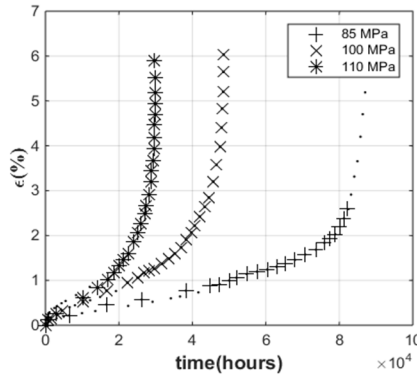


Figure 1. Real experimental data and interpolated (dotted points) for curve fitting purpose of the material at 560°C

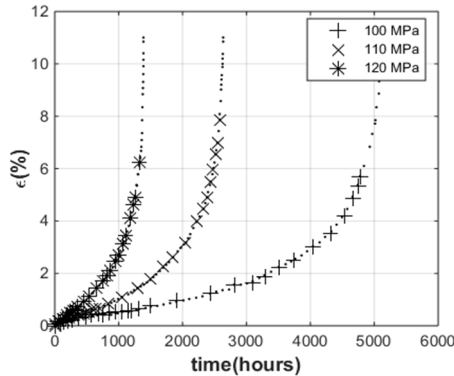


Figure 2. Real experimental data and interpolated (dotted points) for curve fitting purpose of the material at 590°C

C. Initial Guess

How was said in the introduction and according to Kowalewsky, Hayhurst, & Dyson [3] and Mustata & Hayhurst [1] to be successful in the determination of the material constants, it is critical to start with acceptable values for the combined integration/optimization process.

The constants A and B are calculated integrating (1), and applying a linear least square optimization to the variation of the minimum strain rate and the stress. H^* and h are estimated by applying a nonlinear curve fitting for the primary part of the curve. C is calculated by averaging the value of the failure strain, integrating (4) and knowing that $\omega_f=1/3$. Finally, K_c is obtained by applying a similar process to the general optimization, but in this case only is

allowed to vary to the K_c parameter, keeping the remainders constants [3].

The results obtained for the initial values of the constants for 560°C are demonstrated on the Table 4. For 590°C, the only modification in the constants, is $C=2.88$, obtained only accounting the stresses 100, 110, and 120 MPa. Also in the Fig.3, and Fig.4 is illustrated the predicted creep curves using these values for the material constants.

The initial guess for the first case clearly shows a good approximation, whereas, for the second case the approximation diverge considerably from the experimental curve that is due to the initial estimation process is only accurate for a specific temperature. Despite of this divergence in the solution, it will keep the initial values for the optimization process with the intention to check the usefulness of the program to predict the creep mechanic behavior, for different operating temperatures.

D. General Objective Function

The new objective function introduced in this paper to be minimized for the nonlinear least square optimization, is a slightly different to (5), a term to involve lifetime has been introduced following the approach utilized by Kowalewsky, Hayhurst, & Dyson [3], conversely, and it is squared to be part of the least square process. When the predicted range time is longer than the equivalent experimental, some of the simulated data cannot be involved, thus, this term compensates these errors. Furthermore, the strain error is normalized by the failure deformation, therefore, the amplification factor in the time term will have a value in the order of 0 to 1, and due to the normalization, and both terms have the same scale. The value of the weight depends on the level of sensitivity of the creep deformation or lifetime in regard to the parameters to be estimated. When the time and strain have the same relevancy for the optimization, the factor is equal to 1, and when is 0 only the strain errors are accounted. The new function can be expressed as:

$$F_\varepsilon(b) = \sum_{i=1}^m \left[\sum_{j=1}^n \left\{ \left(\frac{\varepsilon(b)_j^{\text{pred}} - \varepsilon_j^{\text{exp}}}{\varepsilon_{fi}^{\text{exp}}} \right)^2 \right\} + w_i \left(\frac{t(b)_{fi}^{\text{pred}} - t_{fi}^{\text{exp}}}{t_{fi}^{\text{exp}}} \right)^2 \right] \quad (8)$$

where the new terms are: $F_\varepsilon(b)$, the new objective function, w_i , a scaling factor for each curve i , $t(b)_{fi}^{\text{pred}}$ and t_{fi}^{exp} denote predicted and experimental lifetime for a specific time and curve, respectively, and $\varepsilon_{fi}^{\text{exp}}$ represents the rupture deformation for a specific stress curve. The second term in the expression is only invoked, when t_f^{pred} is larger than t_f^{exp} . This approach allows to work with values of the same scale, almost guaranteeing an equal contribution in the least square process of each term, and the calibration of the w can be obtained straightforward by using a loop to calculate the optimization solution for $w = [0.1, 0.2, \dots, 1]$.

Table 2. Real experimental data (shaded cells) and new digitized data for curve fitting purpose of material at 560°C

85 Mpa		100 Mpa		110 Mpa	
time (hours)	ε (%)	time (hours)	ε (%)	time (hours)	ε (%)
0	0.000	0	0.000	0	0.000
390	0.051	164	0.027	82	0.113
1014	0.089	906	0.179	150	0.127
3011	0.151	1474	0.165	303	0.179
4961	0.191	2546	0.268	772	0.231
6473	0.217	4504	0.326	1114	0.287
9981	0.268	6581	0.402	2447	0.384
13596	0.316	8042	0.442	3038	0.418
16634	0.461	10240	0.534	4393	0.488
22547	0.430	12184	0.554	5501	0.541
26223	0.570	14901	0.630	8352	0.672
29158	0.521	16630	0.769	10239	0.606
34158	0.597	19218	0.761	12075	0.854
36658	0.639	22366	0.950	14252	0.841
38270	0.769	25070	1.058	16873	1.022
41658	0.730	26954	1.167	18675	1.176
47448	0.913	28101	1.212	20148	1.320
50070	0.995	29575	1.257	21130	1.456
52037	1.058	31705	1.366	22604	1.601
54823	1.140	33589	1.483	23995	1.863
57609	1.194	35227	1.601	25058	2.071
60395	1.239	36947	1.736	25957	2.252
62936	1.302	38256	1.908	27020	2.496
67934	1.456	39730	2.071	27428	2.659
70719	1.574	40630	2.216	28163	2.903
75962	1.836	42021	2.424	28733	3.193
78911	2.026	43330	2.632	28894	3.437
81284	2.379	44475	2.849	29219	3.672
82465	2.587	45700	3.202	29544	3.943
83222	2.910	46598	3.572	29605	4.187
84295	3.312	47494	3.979	29702	4.477
84945	3.653	47981	4.404	29745	4.703
85789	4.201	48058	4.829	29793	4.938
86389	4.701	48382	5.200	29857	5.191
87033	5.189	48459	5.643	30017	5.517
87610	6.017	48536	6.032	30100	5.906

Table 3. Real experimental data (shaded cells) and new digitized data for curve fitting purpose of material at 590°C

100 Mpa		110 Mpa		120 Mpa	
time (hours)	ε (%)	time (hours)	ε (%)	time (hours)	ε (%)
0	0.000	0	0.000	0	0.000
54	0.078	54	0.104	30	0.027
119	0.156	119	0.209	48	0.118
212	0.235	179	0.261	65	0.209
282	0.261	239	0.313	98	0.235
363	0.287	282	0.365	130	0.261
418	0.313	331	0.391	160	0.300
499	0.313	407	0.417	190	0.339
586	0.365	445	0.443	209	0.391
667	0.391	488	0.469	228	0.443
754	0.417	537	0.495	250	0.469
787	0.417	591	0.600	271	0.495
857	0.417	635	0.626	293	0.547
884	0.417	689	0.704	315	0.600
955	0.469	776	0.756	328	0.626
987	0.469	835	0.808	342	0.652
1052	0.521	884	0.873	355	0.665
1080	0.547	933	0.939	369	0.678
1156	0.521	993	1.017	415	0.795
1199	0.547	1052	1.095	461	0.912
1280	0.600	1150	1.199	499	0.991
1318	0.626	1196	1.277	537	1.069
1421	0.678	1242	1.356	597	1.238
1503	0.730	1297	1.434	656	1.408
1606	0.756	1345	1.499	700	1.525
1703	0.808	1394	1.564	743	1.642
1807	0.860	1446	1.655	784	1.773
1904	0.912	1497	1.747	825	1.903
2013	0.965	1543	1.851	868	2.086
2105	1.017	1590	1.955	906	2.242
2213	1.069	1644	2.086	944	2.399
2311	1.121	1698	2.216	968	2.529
2430	1.225	1744	2.320	993	2.659

2528	1.304	1790	2.425	1010	2.757
2626	1.356	1855	2.607	1028	2.855
2729	1.382	1926	2.842	1063	3.050
2805	1.564	1975	2.972	1085	3.220
2913	1.564	2024	3.102	1107	3.389
3011	1.616	2056	3.181	1134	3.598
3098	1.695	2089	3.337	1161	3.806
3201	1.825	2121	3.493	1183	4.041
3282	1.929	2170	3.728	1193	4.178
3385	2.060	2219	3.963	1204	4.315
3505	2.242	2268	4.171	1226	4.588
3613	2.346	2327	4.458	1229	4.595
3700	2.451	2357	4.680	1253	4.875
3765	2.555	2387	4.901	1257	4.970
3863	2.685	2400	5.100	1286	5.345
3949	2.842	2430	5.371	1281	5.372
4042	3.024	2441	5.501	1305	5.775
4145	3.181	2468	5.788	1302	5.789
4231	3.389	2485	6.022	1324	6.205
4329	3.598	2501	6.231	1320	6.114
4410	3.858	2517	6.544	1337	6.439
4497	4.145	2528	6.700	1335	6.394
4579	4.458	2544	6.987	1351	6.674
4660	4.927	2566	7.430	1356	6.831
4741	5.371	2582	7.873	1362	7.091
4779	5.709	2582	8.082	1367	7.352
4839	6.101	2599	8.577	1367	7.847
4931	6.831	2604	9.020	1373	8.343
4969	7.326	2606	9.200	1378	8.838
5013	7.717	2610	9.400	1381	9.073
5018	7.847	2615	9.568	1383	9.307
5034	8.343	2626	9.881	1386	9.607
5072	8.864	2627	10.000	1389	9.907
5078	9.333	2627	10.200	1389	10.155
5094	9.881	2631	10.376	1389	10.402
5094	10.402	2637	10.845	1389	10.845
5099	11.002	2637	11.002	1389	11.002

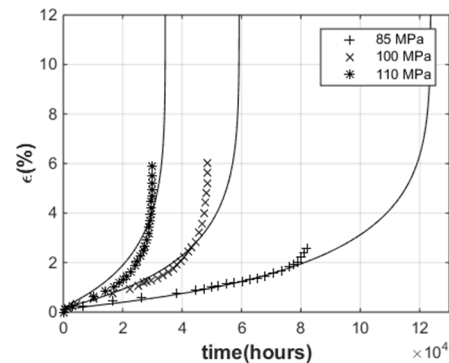


Figure 3. Predicted deformation using the initial estimated values of the material constants for the material at 560°C

Table 4. Initial estimation of the material constants

Initial guess (b_0)	Material at 560°C
A (h^{-1})	1.00E-09
B (MPa^{-1})	1.10E-01
H* (-)	4.26E-01
h (MPa)	5.05E+04
K _c (h^{-1})	6.86E-06
C (-)	4.311

E. Program Development

The program developed in Matlab to obtain the parameters which give the best curve fitting can be divided in four stages.

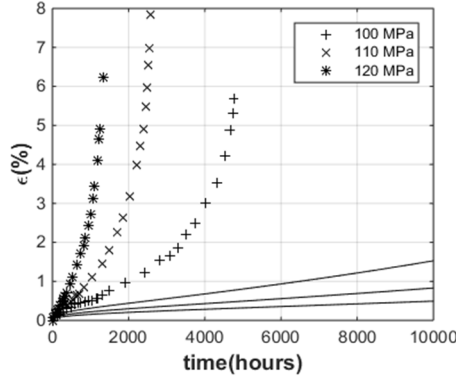


Figure 4. Predicted deformation using the initial estimated values of the material constants for the material at 590°C

First step is to digitize the experimental data and calculate the b_0 , following procedure describe in the sections 4.B and 4.C. Second, the initial conditions are set up, initial values of the state variables x_0 , the tolerances for the optimization solution, and the boundary constraints for the parameters to be optimized. Continuously, it is started a for loop which is utilized to calibrate the value of the time factor w_i , it is decide the number of w tried, N , and the variance on the amplification value Δw , which will vary between 0–1, depending on the different importance of the lifetime in the optimization in each curve. Third, the 'lsqnonlin' iteration process is started, calling the command 'ode45', which is used to integrate the ODE's of the constitutive equations (1)-(4), and predict the strain for each b_k , where k is the specific iteration solution. The simulating range time t_{sim} is specified in the initial condition. The value of b and $F_c(b)$ are obtained and a conditional step comparing with the tolerance says if the optimized solution is achieved. The variable tolerance is identify as λ_1 and function tolerance as λ_2 . Finally, a set of b_p are obtained for the different values of the lifetime factor. The best fitting is achieved by finding the minimal of: normalized residual, error approximation of lifetime, and minimum strain rate, if the experimental values are available. All this process is illustrated at the optimization flow chart at the Fig. 5.

V. RESULTS

A. Ferritic Stainless Steel

The results achieved are demonstrated at the Fig. 6 and Fig. 7, for the material at 560°C and 590°C, respectively. The best values for the lifetime factors are $w = [1, 1, 1]$ and $w = [0.12, 0.12, 1]$ for the temperatures of 560°C and 590°C, respectively. Matlab does not confirm if a global optimum solution has been accomplished but the high accuracy observed in the creep curve behavior, makes think it does, whereas, with the modification of the initial values a slightly difference in the solution is observed.

The values of the optimized constants for both creep curves are illustrated at the Table 5. It can be identify a high difference in the values, specially, for the constants A, and C, which are quadruple and double for the material at 590°C. That confirms the severe dependency on the material constants value in order to represent accurately the creep mechanical behavior.

Table 5. Optimized values for the material constants

Constants	590°C	560°C
A (h^{-1})	4.32E-09	1.00E-09
B (MPa^{-1})	1.26E-01	1.07E-01
H* (-)	4.05E-01	4.52E-01
h (MPa)	1.22E+05	4.98E+04
K_c (h^{-1})	6.84E-05	1.58E-05
C (-)	3.24	6.09

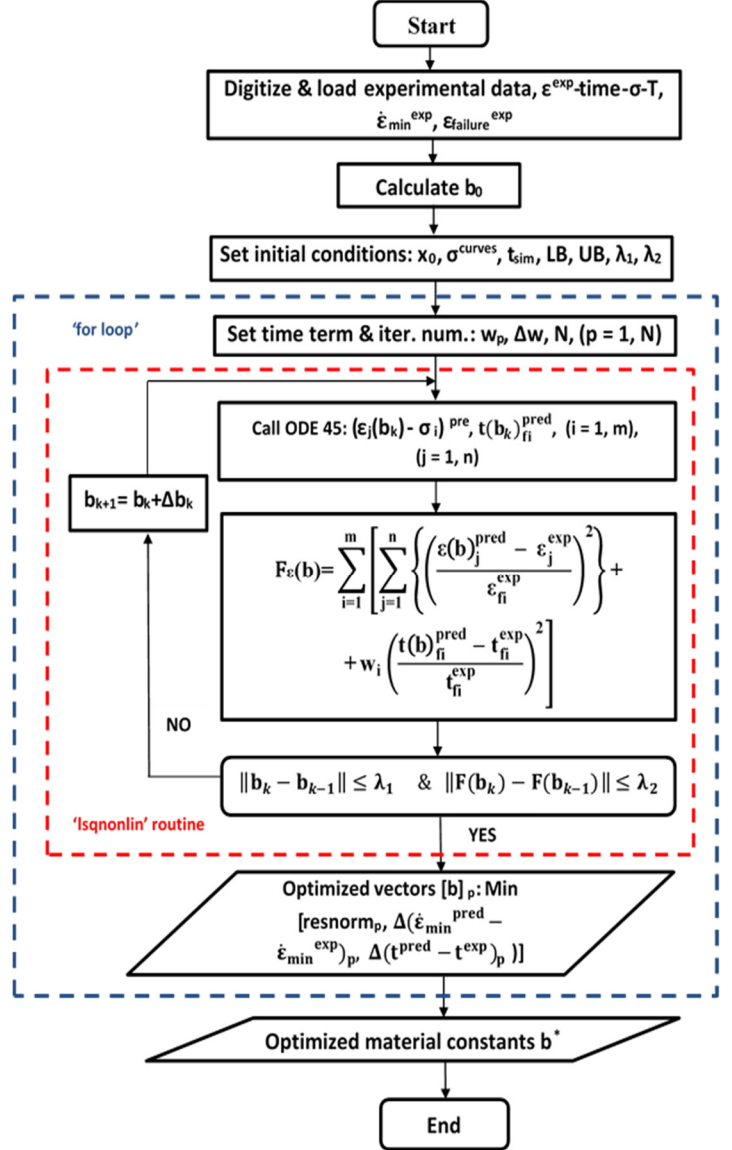


Figure 5. General flow chart of the optimization process

How was said in the section 4.C, the initial estimation for the material at 590°C was not a good guess even that, the Fig. 7 shows a high accuracy in the prediction of the creep damage mechanical behavior, meaning that the optimization routine had ran, and optimized the parameters. Similarly, Fig. 6 illustrates the predicted deformation for the material at 560°C, demonstrating an almost perfect fitting with the experimental data.

B. Comparison with Mustata & Hayhurst Solution

In order to a further validation a comparison with the solution obtained for Mustata & Hayhurst [1] for the same material and conditions has been done.

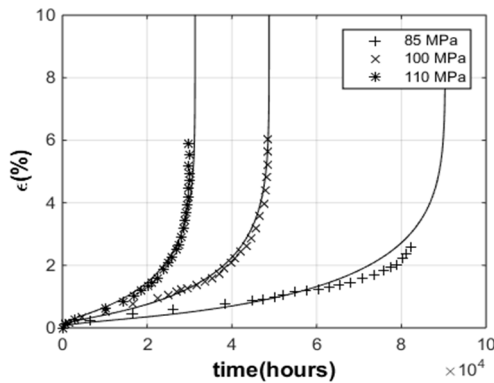


Figure 6. Prediction of mechanical behaviour of material at 560°C determined with the optimized material constants

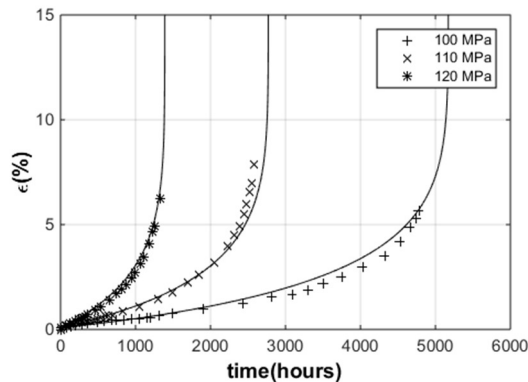


Figure 7. Prediction of mechanical behaviour of material at 590°C determined with the optimized material constants

It is generated the Table 6, and Table 7, which, demonstrates the percentage of error approximation between the predicted and experimental lifetimes and minimum strain rate of both approaches. In respect of the prognostic for the material at 560°C of the minimum rates the error of [1] is nearly 12%, a 5% better, whereas, the error of the lifetime forecasted by this project is a 0.5% better. For the material at 590°C, experimental data for the minimum strain rate is not available, thus, only the error approximation of lifetimes is compared. The average error for this report approach is under 2%, whereas, the other authors approach gives over 6% which is more than the triple of error.

Table 6. Error approximations for lifetimes and minimum creep strain rates for the estimated set of constitutive for material at 560°C

T(560°C)	Mustata & Hayhurst		This project	
Stress (Mpa)	ϵ_{min} (%)	Lifetime (%)	ϵ_{min} (%)	Lifetime (%)
85	3.72	0.13	11.21	0.74
100	16.29	6.54	17.53	6.01
110	14.99	2.34	12.64	0.61
% Average	11.67	3.00	13.79	2.45

Table 7. Error approximations for lifetimes for the estimated set of constitutive parameters for material at 590°C

T(590°C)	Mustata & Hayhurst	This project
Stress (Mpa)	Lifetime (%)	
100	6.47	1.43
110	8.23	2.66
120	4.76	0.68
% Average	6.49	1.59

VI. CONCLUSION AND FUTURE WORKS

To conclude, it can be said that the main objective of this project has been achieved. The optimization program optimizes the material constants for the ferritic steel in the operating temperatures required, and the creep damage mechanical behavior is reproduced with high accuracy.

The objective function is simpler, more compact, less uncertain and at least as accurate as the past papers presented. However, it must say that the accuracy in the results is dependable in the new digitized points for curve fitting purpose. The initial values has been demonstrated to be a key for obtain accurate solution, however, it was showed for the material at 590°C, that, even without a perfect first guesses the program gives an desirable output.

Future works are required to demonstrate the robustness, reliability and usefulness of the optimization program. Regarding to the program implementation, an upgrade of the Matlab code with the aim to be more user friendly is an expect target.

ACKNOWLEDGMENT

This research paper was developed as the final project of the course of MSc Mechanical Engineering Design programme 2014/2015 of the University of Huddersfield. The first author gratefully acknowledges support thought the University of Huddersfield. Specially, he would like express his gratitude to his supervisor Dr Qiang Xu, for his guidance, patience, encouragement, and challenging during this project.

REFERENCES

- [1] Mustata, R., & Hayhurst, D. (2005). "Creep constitutive equations for a 0.5Cr 0.5Mo 0.25V ferritic steel in the temperature range 565°C-675°C". *International Journal of Pressure Vessels and Piping* 82, 363-372.
- [2] Goodall, I., Leckie, F., Ponter, A., & Townley, C. (1979). "The development of high temperature design methods based on reference stress and bounding theorems". *Journal Engineering of Materials and Technology*, 101, 349-355.
- [3] Kowalewsky, Z., Hayhurst, D., & Dyson, B. (1994). "Mechanisms-based creep constitutive equations for an aluminium alloy". *Journal of strain analysis* vol 29 no 4, 309-314.
- [4] Perrin, I., & Hayhurst, D. (1996). "Creep constitutive equations for a 0.5Cr 0.5Mo 0.25V ferritic steel in the temperature range 600°C-675°C". *Journal of strain analysis* vol 31 no 4, 209-314.
- [5] Li, B., Lin, J., & Yao, X. (2002). "A novel evolutionary algorithm for determining unified creep damage constitutive equations". *International Journal of Mechanical Sciences* 44, 987-1002.
- [6] Lin, J., & Yang, J. (1999). "GA-based multiple objective optimisation for determining viscoplastic constitutive equations for superplastic alloys". *International Journal of Plasticity* 15, 1181-1196.
- [7] Gong, Y., Hyde, C., Sun, W., & Hyde, T. (2010). "Determination of material properties in the Chaboche unified viscoplasticity model". *Journal Materials: Design and Application* vol 224, 19-29.
- [8] Saad, A., Hyde, T., Sun, W., Hyde, C., & Tanner, D. (2013). "Characterization of viscoplasticity behaviour of P91 and P92 power plant steels". *International Journal of Pressure Vessels and Piping* 111-112, 246-252.
- [9] Dyson, B., Hayhurst, D., & Lin, J. (1996). "The rigid uni-axial testpiece: creep and fracture predictions using large displacement analysis". *Proc. R. Soc. Lond.*, A452: 655 – 76.
- [10] Gerald, C., & Wheatley, P. (1999). "Applied numerical analysis". London: Addison-Wesley.
- [11] The MathWorks Inc. (2008). *Optimization toolbox™ 4 user's Guide*

The Interpretation of Experimental Observation Date for the Development of Mechanisms based Creep Damage Constitutive Equations for High Chromium Steel

Xin Yang, Zhongyu Lu, and Qiang Xu

Abstract—It is very important to design a safe factor or estimating the remain lifetime for electric power plant components of steam pipes which mostly are manufactured from high chromium steels and work at high temperature and low stress level. The authors will develop the mechanisms based on creep damage constitutive equations for high chromium steel under lows stress in initial stage: (1) Creep cavities were mostly formed attaching with the precipitation of Laves phase or on grain boundary. The Laves phase should play an active role in the nucleation of creep cavities and suggest to explore the function between cavity nucleation and the evolution of Laves phase; (2) The dominant cavity nucleation damage mechanism controls; (3) Brittle intergranular is fracture model; (4) High density number of cavity of crept test high chromium steel at high temperature under low stress could be as fracture criterion.

Index Terms—High chromium steels, cavitation, constitutive equations, low stress.

I. INTRODUCTION

WITH the increasing demand for power supplies and reducing co2 emissions, it is essential to improve the efficiency in energy conversion for chemical and petrochemical plants or power generation systems etc. by higher operating temperature and design stresses. The modified 9Cr-1Mo steels strengthened with addition of Niobium (Nb) and Vanadium (V) or Mo with W have been extensively used in new advanced fossil-fired steam power plants operating at around 873/923K with higher efficiency [1].

The application of computational approach, particularly the development of creep damage constitutive equations is important for industry and academy. In order to design the safe components for power generation industry, particularly in fossil fuel plants and nuclear reactors, it is necessary to estimate their long-term creep behavior such as creep deformation, rupture time by experimental or calculations. On the other hand, obtaining long term creep data is time consuming and cost expensive, thus the long term creep data is limited, and the

extrapolation using the conventional empirical methods may not be reliable [2]. With more accurately predicting the life, components could work for longer time and in a more operationally flexible manner before being retired or fracture [3]. Therefore, it is necessary to develop valid creep damage constitutive equations to accurately predict the long-term creep life time.

During the past several decades, there were many various developed creep damage constitutive equations models for describing creep damage process and predicting the lifetimes of the components operating at high temperature [4-9]. The classic Kachanov-Robotnov (KR) equations (1) (2) are the earliest to use the concept of continuum creep damage mechanics (CMD) in the development of constitutive equations by defining a single empirical mathematical parameter ‘ ω ’ [4]. The damage parameter as denoting creep damage of materials changes between 0 and 1. It is an obvious advantage for computations of creep damage with a single empirical damage parameter formulation. However, it cannot explore a satisfactory definition with single one damage parameter, even if there is more than one physical creep damage mechanism during creep damage process.

$$\dot{\epsilon} = \dot{\epsilon}_0 \left[\frac{\sigma}{\sigma_0(1 - \omega)} \right]^n \quad (1)$$

$$\dot{\omega} = \dot{\omega}_0 \left[\frac{\sigma}{\sigma_0(1 - \omega)} \right]^m \quad (2)$$

Dyson’s physically framework based on continuum damage mechanisms has contained dominant mechanisms reflecting the evolution of materials’ creep state. Dyson’s equations (3) have coupled with the four various creep damage mechanisms of strain hardening, coarsening of the strengthening particles, solid solute depletion, and grain boundary creep cavitation [4]. Comparing with KR equations, the advantage of Dyson’s equations is different microstructural damage parameters based on physical phenomenon could be defined a dimension damage variable for each different creep damage mechanism. However, there are two difficult aspects for practically using Dyson’s constitutive equations with multitude creep damage mechanisms. On the one hand, any given alloy steels may be operated including all, some or none of creep damage mechanisms. Depending on the given alloy steels’ microstructure and operating conditions such as the stress and

X. Yang, Department of Computing and Engineering, University of Huddersfield, Queensgate, HD1 3DH UK, e-mail: Xin.Yang@hud.ac.uk.

Z.Y. Lu, Department of Computing and Engineering, University of Huddersfield, Queensgate, HD1 3DH, UK, ORCID ID 0000-0002-0585-2806, e-mail: j.lu@hud.ac.uk.

Q. Xu, Department of Computing and Engineering, University of Huddersfield, Queensgate, HD1 3DH, UK, ORCID ID 0000-0002-5903-9781, e-mail: Q.Xu2@hud.ac.uk.

temperature level, the development of creep damage constitutive equations based Dyson's equations should be coupled the effective and relevant creep damage. The other point is to identify the effective operating creep damage mechanisms or parameters based on the quantification analysis of creep damage, especially for cavitation.

$$\dot{\epsilon} = \frac{\dot{\epsilon}_0}{(1-D_d)} \sinh \left[\frac{\sigma(1-H)}{\sigma_0(1-D_p)(1-D_N)} \right]$$

$$\dot{H} = \frac{h'}{\sigma} \left(1 - \frac{H}{H^*} \right) \dot{\epsilon}$$

$$\dot{D}_d = C(1 - D_d)^2 \dot{\epsilon} \quad (3)$$

$$\dot{D}_p = \frac{k_p}{3} (1 - D_p)^4$$

$$\dot{D}_n = \frac{k_N}{\epsilon_{fu}} \dot{\epsilon}$$

The development of cavitation equations proposed by Yin has been applied for high chromium steel at high temperature under middle and high stress, the modeling results of creep life time for P92 agreed with experiment data [6]. However, there are no experiment results for high Cr steels agreements with the extension application of Yin's new constitutive equations for low stress [10]. Chen et al [8] developed new creep damage constitutive equations for high chromium steels, Yang has validated the developed constitutive equations working well under middle and high stress, however, the modeling results of creep curve under low stress level is still show high strain at failure [11]. There are three reasonable exploration for these developed constitutive equations could not be extended from high stress range to low: firstly, confirming the typical breakdown of creep in ASTM Grade92 steel, transition from ductile transgranular fracture to brittle intergranular fracture is the major cause of the breakdown, Lee et al has proposed in the research of causes of breakdown of creep strength in 9Cr-1.8W-0.5Mo-VNb steel [12]; secondly, most of cavities are nucleated at coarse Laves phase particles on grain boundaries, Laves phase precipitates and grows during creep exposure, and coarsening of Laves phase particles over critical size triggers the cavity formation and the consequent brittle intergranular fracture [12]; thirdly, creep damage mechanisms and materials parameters of constitutive equations should be changed under different working conditions.

Generally accepted terms of the main cause of creep failure or fracture is the damage process dominated by starting with cavity nucleation, then growth, finally coalescence of cavity leading to rupture [13-16]. The cavitation is significant internal damage mechanism contributing to the ultimate failure of the material, it indicated in the study of the relative significance of various internal creep damage mechanisms on the overall creep damage and lifetime of P91 steel [17]. However, there is still lacking of the relationship of the architecture evolutions of cavitation damage and no clear correlation of final damage with cavity. Recently, the advanced technique of synchrotron X-ray

micro-tomography, at large synchrotron source for characterizing ex-situ creep damage in materials, has been used in investigating on the evolution of cavitation and provided the 3D spatial distribution and cavitation characteristics with increasing creep expose time [18-22]. And quantitative analysis also provided a fundamental understanding of the underlying cavity damage mechanism of creep failure, in order to develop more advanced and accurate damage constitutive equations for estimating the lifetime of high chromium steels. In this paper, author will interpret the experimental observation data for the development of mechanisms based creep damage constitutive equations for high chromium steels.

II. CURRENT CRITICAL EXPERIMENTAL DATA ON CAVITATION DAMAGE

Recent decade years, the evolution process or fracture surface of cavitation of high chromium steel during or after creep expose have been investigated by traditional optical microscope (OM), transmission electron microscopy (TEM), scanning electron microscopy (SEM), backscattering electron(BSE), and synchrotron X-ray micro-tomography. The section will be summarized and analyses critical experiment data on cavitation in order to understanding the underlying cavitation damage mechanisms for developing creep damage constitutive equations for high chromium steels.

A. Cavitation Nucleation Site for High Chromium Steel at High Temperature

After studying of the materials of ASTM Grade 92 steel crept at 550-650°C for up to 63151h, Lee et al (2006) suggested the precipitation and coarsening of Laves phase are responsible for the intergranular fracture by SEM images of cavity formation in specimens crept with showing cavities are attached with the coarsening Laves phase, more precisely, cavities were nucleated at the coarsening precipitation of Laves phase along grain boundaries [12].

A backscatter SEM micrograph of the ruptured specimen of a 12% Cr tempered martensite ferritic steel under the condition of long term creep (823K, 120MPa, 139971h) shows creep cavities appearing a prior austenite grain boundary perpendicular to the direction of applied stress (2009) [23].

Later on (2010), more than 100000h of creep expose under 80 MPa at 600°C for the microstructure of the Grade 91 steel had been studied. According to SEM images from the distance of 7 nm fracture surface, it shows cavities nucleated at boundaries next to the precipitation of Laves phase [24].

Parker (2013) pointed out the formation of Laves phase could influence creep resistant behavior attributing to relatively hard Laves phase can provide preferred sites for nucleation of creep voids [25].

Recently, creep rupture tests of the 9% chromium steel T92 had been done at different high temperature (600°C, 650°C and 700°C) under between 46 and 200 MPa (2014). The BSE micrographs of creep specimen which creep at 700°C for 8232h in the necked section shows cavities nucleation at the interface of the large Laves phases, and $M_{23}C_6$ carbide and Laves phases

provide potential sites for creep cavities nucleation [26].

Zhu et al (2014) performed a creep test of 9.5% Cr chromium steel without addition of C and N at 650°C under different stresses, in order to avoid the influence of other phase precipitations such as $M_{23}C_6$ and MX, and summarized and reported three points [27]: (1) The precipitation of Laves phase mainly on the grain boundary during creep in the alloy studies; (2) There is no influence for the rate of Laves phase precipitation or the rate of growth and coarsening of Laves phase precipitation; (3) Suggest that increasing cavity nucleation triggered by the coarsening of Laves phase is the reason for the more rapid drop in creep rupture strength at lower stresses testing.

Focuses on the characterization of cavities evolution in a P91 steel pipe, creep samples were applied an initial tensile stress of 60MPa for 7000 and 9000h at 650°C. The SEM micrograph of the gauge section of 9000h crept specimen shows that creep cavities appeared along the lath boundaries and in the vicinity of second phase particles such as Laves phase [22].

Based on above critical experimental observations, in general speaking, creep cavities were mostly formed attaching with the precipitation of Laves phase or grain boundary for high chromium steel under low stress. Taking all above viewpoints into account, the Laves phase should play an active role in the nucleation of creep cavities. So it is important to explore quantitative functional relations between the evolution of Laves phase and cavitation damage mechanisms for the development of mechanisms based creep damage constitutive equations.

B. Dominant Cavity Damage Mechanism for High Chromium Steels

According to different applied creep stress level, a change of cavitation damage mechanism will govern creep damage fracture for high chromium steels at a given temperature. This section will discuss the dominant cavity damage mechanism for high chromium steels under high and low stress levels.

1) Dominant Cavity Growth Mechanism for High Chromium Steels under High Stress Level

In more earlier studying, the principle mechanism at high stress levels is the viscoplasticity-assisted ductile rupture mechanism. The damage mechanism begins with void nucleation at the preferential stress concentration areas inside the grains; then void growth it assisted by grain deformation and followed by coalescence of the cavities, the evolution of cavities damage is equivalent to the creep strained cavity constrained growth dominant [4] [16] [28].

Recently, with three dimensional techniques of X-ray micro-tomography, the spatial distribution and 3D characteristics of the creep void with increasing creep expose time for high chromium steel have been provided at high temperature and under a stress range between 120MPa and 180MPa [18]. Based on the experiment data of cavitation volume fraction, the average diameter of voids and the number density of cavitation curves, void volume fraction and the number density both increase, void volume fraction increases more rapidly. It reveals which of void nucleation or growth and

coalescence process dominate. Furthermore, this results mean that the void growth by coalescence is progressively strengthened over nucleation as the stress is reduced in the range of 120-180MPa [18].

2) Dominant Cavity Nucleation Mechanism for High Chromium Steels under Low stress Level

General speaking, the creep deformation under low stress is of diffusional and the void nucleation is controlled by the maximum shear stress; which is in line with the general understanding reported by Miannay, according to Xu's conference paper [29]. With finer resolution of cavity size, Gupta et al [18] has observed and reported the number density and mean size of crept specimen under different stress. It reveals that though at failure, the number density under low stress is much higher than that under high stress, and it is in the order of 2.5; this strongly indicated the significant effect of time and the nucleation is visco-type rather than stress controlled; it is useful and important to obtain the nucleation rate with time under different stress level and collaborative research on this is sought.

Creep test of P91 steel cross weld specimens were performed at 650°C under stress of 66MPa and interrupted at 20%, 40%, 60%, and 80% of the creep damage which is defined as the ratio of interrupted time to the rupture time [30]. The void number density in cross weld creep damage specimens were observed from 20%, 40%, 60% and 80% creep damage process and slowly grow during 60% and quickly increase from 60% to 80%. The microstructures of the 32% creep damage specimen and of the rupture specimen at 6740 h were observed by the optical microscope, in order to understand void initiation and growth state. It shows more creep cavity nucleated than cavitation growth from 32% to rupture time. Although P91 steel weldment related stress redistribution is different from base material, it gives research reference under the lack of experiment data for high chromium steel under low stress level at high temperature.

3) Exploring the Relationship Function between Cavity Nucleation and Laves Phase

There are many literatures about researching the evolution of Laves phase and its influence on the stability of microstructure and creep rupture strength in creep expose for high chromium steels.

Several years ago, experiments were taken for investigating on the growth kinetics of Laves phase in high chromium steels. The results of simulation of growth of the Laves phase model agreed with scanning TEM measurements [31] [32]. In the creep test of modified P911 heat resistant steel at 823K, it shows that Laves phase appeared after a creep strain of 1%, and then the man size of Laves increased from 190 to 265nm with increasing strain to 18% [33]. Recently, the Laves phase process was characterized by SEM images, it found that creep rupture strength started decreasing more rapidly before the Laves phase reached equilibrium; there was no significant effect on the early stages of Laves phase formation by creep stress and strain[34], and the growth kinetics of Laves phase by creep deformation[27].

Secondary phase ($M_{23}C_6$, Laves phase) don't have influence on cavity growth during creep expose [21]. As discussed in

section II part B, the cavity nucleation is dominant damage mechanism for high chromium steel under low stress. Otherwise, section II part A presents the cavity nucleation site is attached to Laves phase. So the void nucleation rate is very important factor to be taken into account with the evolution of Laves phase. The author suggests building the relationship function between cavitation nucleated and the evolution of Laves phase for high chromium under low stress level at high temperature. However, the experiment data of cavity nucleation at the beginning of creep damage is lack in current, more experiments will be done for the further study of the cavity nucleation rate for high chromium steels under low stress in future.

III. FRACTURE CRITERIA

A. Transgranular or Intergranular Model

The transgranular and intergranular phenomenon of fracture experiment rupture samples were obverted after creep tests by SEM. This section will summary and analyses the fracture model for high chromium steels under different stress and temperature based on experimental observation in past 15 years as shown in below Table I.

TABLE I
TRANS-GRANULAR OR INTER-GRANULAR PHENOMENON OF HIGH CHROMIUM STEEL AT HIGH TEMPERATURE UNDER DIFFERENT STRESS

Materials	Temperature	Applied Stress	Fracture Model	Reason of Fracture
9Cr-1Mo Steel (Goyal et al.,2014) [35]	873K (600°C)	150MPa	Ductile transgranular fracture	Resulting from coalescence of microvoids
Grade 91 Steel (Shrestha et al., 2013) [36]	700°C	200MPa	Ductile transgranular Fracture	Resulting from diffusive cavity growth and coalescence
9Cr-1Mo Ferritic Steel in quenched and tempered (Choudhary, 2013) [37]	873K (600°C)	100MPa	Ductile transgranular Fracture	Resulting from micro-void coalescence
9Cr-1Mo Ferritic Steel in quenched and tempered (Choudhary, 2013) [37]	873K (600°C)	60MPa	Ductile transgranular Fracture	Resulting from micro-void coalescence
9Cr-1Mo Steel (Vanaja et al.,2012) [38]	773K (500°C), 823K (550°C), 873K (600°C)	120MPa, 140MPa, 160MPa, 180MPa	Ductile transgranular fracture	Resulting from coalescence of microvoids
Modified 9Cr-1Mo steel (Masse et al.,2012) [28]	625°C	120MPa	Ductile intergranular fracture	Cavities nucleate, Cavity growth,coalescence of small intergranular cavities

Modified 9Cr-1Mo Ferritic Steel (Choudhary et al.,2011) [39]	823K (550°C)	200MPa	Ductile transgranular fracture	Resulting from coalescence of microvoids
9Cr-1.8W-0.5Mo-VNb Steel (P92) (Lee et al., 2006)[12]	550°C	270MPa	Ductile transgranular fracture	Resulting from coalescence of microvoids
9Cr-1.8W-0.5Mo-VNb Steel (P92) (Lee et al., 2006) [12]	650°C	80MPa	Brittle intergranular fracture	Suggesting the precipitation and coarsening of Laves phase are responsible.
9Cr-1Mo Ferritic Steel in quenched and tempered (Choudhary et al., 1999) [40]	793K (520°C)	150MPa	Ductile transgranular fracture	Resulting from void coalescence
9Cr-1Mo Ferritic Steel in quenched and tempered (Choudhary et al.,1999) [40]	873K (600°C)	90MPa	Ductile transgranular fracture	Resulting from void coalescence

From the comparison of Table I, in general accepted view, ductile transgranular model is appropriate for high chromium steels at high temperature under high stress level and resulting from the coalescence of micro-void, meanwhile, brittle intergranular is suitable for at intermediate and low stress. But Choudhary's experiment observation for high chromium steels under 60MPa at 873K (600°C) shows ductile transgranular fracture on surface [37], the results is inconformity with above summarized brittle intergranular model at intermediate and low stress, because the experiment materials of 9Cr-1Mo steel have been quenched and tempered resulting with more stronger strength and ductility. Confirming fracture of creep in ASTM Grade92 steel, transition from ductile transgranular fracture to brittle intergranular fracture is as the major cause of the breakdown [12].

B. Fracture Phenomenon

In early year, creep tests of smooth and notched hollow cylinders were performed under internal pressure and additional axial load, there was the number of creep cavities on the outer surface, the middle of the cross section and the inner surface of the hollow specimens after the end of the experience [41]. Later on, the creep test of an ASME Grade 91 steel has been done for 113431 h under a load of 80MPa, the SEM images were observed a high number density of cavities through the fracture surface of the crept specimen [24]. And high densities of creep cavities of fracture Grade 92 steel specimen surface at high temperature under low stress was also observed by typical optical micrographs [25] [26].

C. Fracture Criterion

Base on section III part B described fracture phenomenon of a density of cavities on fracture surface, the fracture criteria would be built basing on quantifying cavitation and determining the end of creep model end of materials lifetime.

Classic Kachanov-Robitnov Hayhurst (KRH) creep damage constitutive equations developed for low stress Cr alloy. The variable ω represents intergranular cavitation damage and varies from zero and is related to the area fraction of cavitation damage. The maximum value of ω is approximately 1/3 at failure [42].

Petry and Lindet modified new creep damage constitutive equations based on Hayhurst's (KRH) and predicted the creep life time for high chromium steels (P91, P92) [7]. Because of numerical convergence reasons, the computations were stopped when the macroscopic strain reached 10% value. At this point, the strain rate was so high that remain life could be estimated as less than 0.1% of the elapsed time. The time at a 10% strain was taken as the time-rupture t_R . Moreover, it was assumed that the damage varied between 0 at initial state and threshold value D_c which is taken between 0.1 and 0.3. So, the modified model of Petry and Lindet applied the weakest link approach, so the fracture criterion should be $t_R = \text{Min}\{t(\varepsilon = \varepsilon_c = 10\%), t(D = D_c)\}$.

Above methods provide suggestions for fracture criterion of developing novel constitutive equations for high chromium steels under low stress. The critical size of average cavity diameter could not be the fracture criterion; because of the experiment data shows the size and volume fraction of cavitation increase with increasing stress. Based on section III part B of fracture phenomenon, a high number density of cavities for fracture surface of high creep steels at high temperature under low stress could be regarded as fracture criterion through quantified analysis of fracture specimens section.

IV. CONCLUSION AND FURTHER WORK

Based on the interpretation of experiment observation data, the following points will be accepted and applied for the development of mechanisms based creep damage constitutive equations for high chromium steel:

- 1) Creep cavities were mostly formed attaching with the precipitation of Laves phase or grain boundary for high chromium steel under low stress. The Laves phase should play an active role in the nucleation of creep cavities.
- 2) The dominant cavity nucleation mechanism is appropriate for high chromium steels under low stress level. The Laves phase does not influence the creep void growth rate and suggest exploring the function between cavity nucleation and the evolution of Laves phase.
- 3) Brittle intergranular model is appropriate for high chromium steels at high temperature under low stress level.
- 4) High density number of cavity of crept text high chromium steel at high temperature under low stress could be as fracture criterion.

In order to build novel function of cavitation nucleation mechanism based creep damage constitutive equations for high chromium steels under low stress condition, the further work will be done as follows:

- 1) Creep nucleation of high chromium steels at high temperature under low stress should be measured using OM, TEM, SEM or X-ray;
- 2) Quantify analysis experiment data of creep nucleation of high chromium steels at high temperature under low stress, build the relationship function between cavity nucleation and the evolution of Laves phase;
- 3) Quantify the fracture surface of crept specimens; make sure the value of high density number of cavity as fracture criterion.
- 4) Coupling the new function of creep cavitation mechanisms and creep deformation for novel creep damage constitutive equations

ACKNOWLEDGMENT

Thanks for the School of Computing and Engineering, University of Huddersfield for partial scholarship and this work is carried out as part of first author's PhD research.

REFERENCES

- [1] M. Tatomi, M. Tabuchi, "Issues relating to numerical modelling of creep crack growth," *Eng. Frac. Mech.*, vol.77, pp.3043-3052, 2010.
- [2] W. G. Kim, S. H. Kim, C. B. Lee, "Long-term creep characterization of Gr.91 Steel by modified creep constitutive equations," *Met. Mater. Int.*, vol.17, pp.497-504, 2011.
- [3] J.P. Rouse, W. Sun, T.H. Hyde, A. Morris, "Comparative assessment of several creep damage models for use in life prediction," *I. J. Press. Ves. Pip.*, vol.108-109, pp.81-87, 2013.
- [4] B.F. Dyson, "Use of CDM in materials modelling and component creep life prediction," *J. Press. Ves. Tech.*, vol.122, pp.281-296, 2000.
- [5] V. Gaffard, J. Besson and A.F. Gourgues-Lorenzon, "Creep failure model of a tempered martensitic stainless steel integrating multiple deformation and damage mechanisms," *Int. J. Frac.*, vol.133, pp.139-166, 2005.
- [6] Y. Yin, R.G. Faulkner, P.F. Morris, P.D. Clarke, "Modeling and experimental studies of alternative heat treatments in steel 92 to optimize long term stress rupture properties," *Energy Mater.*, vol.3, pp.232-242, 2008.
- [7] C. Petry, G. Lindet, "Modelling creep behavior and failure of 9Cr-0.5Mo-1.9W-VNb Steel," *Int. J. Press. Vess. Pip.*, vol.86, pp.486-494, 2009.
- [8] Y. X. Chen, W. Yan, P. Hu, Y. Y. Shan, K. Yang, "CMD modeling of creep behavior of T/P91 steel under high stress," *ACTA Metallurgica Sinica*, vol.47, pp.1372-1377, 2011.
- [9] M. Basirat, T. Shretha, G.P. Potirniche, I. Charit, K. Rink, "A study of the creep behavior of modified 9Cr-1Mo steel using continuum-damage modeling," *Int. J. Plasticity*, vol.37, pp.95-107, 2012.
- [10] X. Yang, Q. Xu, Z. Lu, "Preliminary review of the influence of cavitation behavior in creep damage constitutive equations," In: *Machinery, Materials Science and Engineering Applications*, Trans. Tech. Publ., 2014, pp.46-51.
- [11] X. Yang, Q. Xu, Z. Lu, "The Development and Validation of the Creep Damage Constitutive Equations for P91 Alloy," In: *Proceedings of the 2013 World Congress in Computer Science and Computer Engineering and Application*, CSREA Press, 2013, pp. 121-127.
- [12] J.S. Lee, H.G. Armaki, K. Maruyama, T. Muraki, and H. Asahi, "Causes of breakdown of creep strength in 9Cr-1.8W-0.5Mo-VNb Steel," *Mater. Sci. Eng. A*, vol.428, pp.270-275, 2006.
- [13] H. T. Yao, F. Z. Xuan, Z. D. Wang, S. T. Tu, "A review of creep analysis and design under multi-axial stress states," *Nucl. Eng. and Des.*, vol.237, pp.1969-1986, 2007.
- [14] V. Sklenicka, K. Kucharova, M. Svoboda, L. Kloc, J. Bursik, and A. Kroupa, "Long-term creep behavior of 9-12% Cr power plant steels," *Mater. Characterization*, vol.51, pp.35-48, 2003.

- [15] C. Westwood, J. Pan, A.D. Crocombe, "Nucleation, growth and coalescence of multiple cavities at a grain-boundary," *European J. Mech. A/Solids*, vol.23, pp.579-597, 2004.
- [16] T. Masse, Y. Lekeail, "Creep Mechanical Behavior of Modified 9Cr1Mo Steel Weldments: Experimental Analysis and Modelling," *Nucl. Eng. Des.*, vol.254, pp.97-110, 2013.
- [17] X. Yang, Q. Xu, Z. Lu, "The relative significance of internal damage mechanisms on the overall creep damage and ultimate failure of P91 steel," 6th International 'HIDA' Conference: Life/Defect Assessment & Failures in High Temperature Plant, Nagasaki, Japan, Dec.2-4 2013.
- [18] C. Gupta, H. Toda, C. Schlacher, Y. Adachi, P. Mayr, C. Sommitsch, K. Uesugi, Y. Suzuki, A. Takeuchi, M. Kobayashi, "Study of creep cavitation behavior in tempered martensitic steel using synchrotron micro-tomography and serial sectioning techniques," *Mater. Sci. Eng. A*, vol.564, pp.525-538, 2013.
- [19] A. Isaac, F. Sket, W. Reimers, B. Camin, G. Sauthoff, A.R. Pyzalla, "In situ 3D quantification of the evolution of creep cavity size, shape, and spatial orientation using synchrotron X-ray tomography," *Mater. Sci. Eng. A*, vol.478, pp.108-118, 2008.
- [20] F. Sket, K. Dzieciol, A. Borbely, A.R. Kaysser-Pyzalla, K. Maile, R. Scheck, "Microtomography investigation of damage in E911 steel after long term creep," *Mater. Sci. Eng. A*, vol.528, pp.103-111, 2010.
- [21] L. Renversade, H. Ruoff, K. Maile, F. Sket, A. Borbely, "Microtomographic assessment of damage in P91 and E911 after long-term creep," *Int. J. Mater. Res.*, vol.105, pp.621-627, 2014.
- [22] S.D. Yadav, B. Sondergerger, B. Sartory, C. Sommitsch, C. Poletti, "Characterisation and quantification of cavities in 9Cr martensitic steel for power plants," *Mater. Sci. Tech.*, vol.31, pp.554-564, 2015.
- [23] A. Aghajan, Ch. Somsen, G. Eggeler, "On the effect of long-term creep on the microstructure of a 12% chromium tempered martensite ferritic steel," *Acta Materialia*, vol.57, pp.5093-5106, 2009.
- [24] C.G. Panait, W. Bendick, A. Fuchsmann, A.-F. Gourgues-Lorenzon, J. Besson, "Study of the microstructure of the Grade 91 steel after 100000h of creep expose at 600°C," *Int. J. Press. Ves. Pip.*, vol.87, pp.326-335, 2010.
- [25] J. Parker, "In-service behavior of creep strength enhanced ferritic steels Grade 91 and Grade 92 – Part 1 parent metal," *Int. J. Press. Ves. Pip.*, vol.101, pp.30-36, 2013.
- [26] M. Nie, J. Zhang, F. Huang, J.W. Liu, X.K. Zhu, Z.L. Chen, L.Z. Ouyang, "Microstructure evolution and life assessment of T92 steel during long-term creep," *J. Alloys and Compounds*, vol.588, pp.348-356, 2014.
- [27] S. Zhu, M. Yang, X.L. Song, S. Tang, Z.D. Xiang, "Characterisation of Laves phase precipitation and its correlation to creep rupture strength of ferritic steels," *Mater. Character.*, vol.98, pp.60-65, 2014.
- [28] T. Masse, Y. Lejeail, "Creep behaviour and failure modelling of modified 9Cr1Mo steel," *Nucl. Eng. Des.*, vol.246, pp.220-232, 2012.
- [29] Q. Xu, Z.Y. Lu, X. Wang, "Damage Modelling: the current state and the last progress on the development of creep damage constitutive equations for high Cr steels," 6th International 'HIDA' Conference: Life/Defect Assessment & Failures in High Temperature Plant, Nagasaki, Japan, Dec.2-4 2013.
- [30] T. Ogata, T. Sakai, M. Yaguchi, "Damage characterization of a P91 steel weldment under uniaxial and multiaxial creep," *Mater. Sci. Eng. A*, vol.510-511, pp. 238-243, 2009.
- [31] O. Prat, J. Garcia, D. Rojas, C. Carrasco, G. Inden, "Investigations on the growth kinetics of Laves phase precipitates in 12% Cr creep-resistant steels: Experimental and DICTRA calculations," *Acta Materialia*, vol.58, pp.6142-6153, 2010.
- [32] O. Prat, J. Garcia, D. Rojas, C. Carrasco, G. Inden, "The role of Laves phase on microstructure evolution and creep strength of novel 9%Cr heat resistant steels," *Intermetallics*, vol.32, pp.362-372, 2013.
- [33] A. Kipelova, A. Belyakov, R. Kaibyshev, "Laves phase evolution in a modified P911 heat resistant steel during creep at 923K," *Mater. Sci. Eng. A*, vol.532, pp.71-77, 2012.
- [34] M. Isik, A. Kostka, G. Eggeler, "On the nucleation of Laves phase particles during high-temperature exposure and creep of tempered martensite ferritic steels," *Acta Materialia*, vol.81, pp.230-240, 2014.
- [35] S. Goyal, K. Laha, C.R. Das, S. Panneerselvi, and M.D. Mathew, "Creep live predication of 9Cr-1Mo steel under multiaxial state of stress," *Metal. Mater. Trans. A*, vol.45A, pp.619-632, 2014.
- [36] T. Shrestha, M. Basirat, I. Charit, G.P. Potirniche, and K.K. Rink, "Creep rupture behavior of grade 91 steel," *Mater. Sci. Eng. A*, vol.565, pp.382-391, 2013.
- [37] B.K. Choudhary, "Tertiary Creep Behavior of 9Cr-1Mo Ferritic Steel," *Sci. Eng. A*, vol.585, pp.1-9, 2013.
- [38] J. Vanaja, K. Laha, R. Mythili, K.S Chandravathi, S.Saroja, and M.D. Mathew, "Creep Deformation and Rupture Behavior of 9Cr-1W-0.2V-0.06Ta Reduced Activation Ferritic-Martensitic Steel," *Mater. Sci. Eng. A*, vol.533, pp.17-25, 2012.
- [39] B. Choudhary, E. Samuel, "Creep Behavior of Modified 9Cr-1Mo Ferritic Steel," *Jnl. Nucl. Mater.*, vol.412, pp.82-89, 2011.
- [40] B. Choudhary, S. Saroja, K. Rao, S. Mannan, "Creep-rupture behavior of forged, thick section 9Cr-1Mo ferritic steel," *Metal. Mater. Trans. A*, vol.30A, pp.2825-2834, 1999.
- [41] M. Rauch, K. Maile, M. Ringel, "Numerical calculation and experimental validation of damage development in 9Cr steels," 30th MPA-Seminar in conjunction with the 9th German-Japanese Seminar Stuttgart, October 6 and 7, 2004.
- [42] I.J. Perrin, D.R. Hayhurst, "Creep constitutive equations for 0.5Cr0.5Mo0.25V ferritic steel in the temperature range 600–675°C," *J. Strain Anal.*, vol.31 (4), pp.299–314, 1996.

Experiment based investigation into micro machinability of Mg based metal matrix composites (MMCs) with nano-sized reinforcements

Xiangyu Teng, Dehong Huo, Wai Leong Eugene Wong

School of Mechanical and Systems Engineering
Newcastle University, Newcastle Upon Tyne, UK
xiangyu.teng@ncl.ac.uk

Manoj Gupta

Department of mechanical Engineering
National University of Singapore

Abstract— As a composite material with combination of low weight and high engineering strength, metal matrix composites (MMC) have been utilized in numerous area such as aerospace, automobile, medical and optics in the past two decades. With the increasing demand on miniaturized products with enhanced mechanical properties, micro milling has been employed to manufacture the MMC component. However it is recognised as a difficult-to-cut materials due to its improved strength. This paper presents the micro machinability of Mg based metal matrix composites reinforced with nano-sized particles. The influence of cutting parameters on the surface morphology and cutting force was studied. The specific cutting energy, cutting force and surface morphology, whilst the minimum chip thickness was characterised with the aim of examining the size effect. The results show that depth of cut and spindle speed have significant effect on the surface roughness. The higher cutting force and worse machined surface were obtained at the small feed per tooth ranging from 0.15 to 0.5 $\mu\text{m}/\text{tooth}$ due to the size effect.

Keywords- micro machining; metal matrix composites; nano reinforcements; surface morphology; cutting forces; size effect

I. INTRODUCTION

With the increasing demands on miniaturized sizes, low weight and superior mechanical properties in numerous areas [1], magnesium (Mg) based metal matrix composite material has become one of the candidates which is widely used in various areas such as aerospace, automobile, electronics and biomedical engineering, due to its biocompatibility, enhanced stiffness, strength, fracture toughness and reduced weight.

The mechanical properties of Mg based composites are improved significantly due to the addition of hard reinforcement materials such as ceramic or metallic particles. Its heterogeneous structure and high abrasiveness particles make it become a difficult-to-cut material. Several attempts on the fabrication of MMC components using EDM and laser machining has been made. For EDM, the results show that the poor conductivity of ceramic particles result in a low material removal rate and the electrode wear was severe which lead to a high cost [2]. In addition, the surface finish is found to be relatively poor and the micro structure was changed due to the high heat generated during the laser machining

[3]. Therefore due to the cost-effective, controllable and high accuracy, conventional machining is promising to be the proper method to manufacture MMC parts compared to other methods. According to the turning experiment on Al/SiC MMC conducted by Manna and Bhattacharayya et al. [4] high spindle speed, low feed depth of cut are recommended for generating better surface finish, and low cutting speed could lead to a fast flank tool wear.

In order to meet the increasing demand for the high precision and miniature components with superior mechanical properties, as an emerging machining process micro machining attracts more attention from both industries and academia. Micro machining is a mechanical material removal process used to machine high accuracy 3D components with dimension or feature size ranging from several microns to several millimetres. Several critical issues in machining process including minimum chip thickness, cutting edge radius effect, ploughing effect become prominent with the decreasing ratio of uncut chip thickness and cutting tool edge radius [5].

It has been reported that MMCs reinforced with small volume fraction of nano-sized particles produced superior results in mechanical properties than those reinforced with micro-sized particles. [6] The outstanding mechanical properties bring a tremendous challenge for micro machining, and the cutting mechanism are not well understood yet due to rigid nano-sized reinforcements dispersed into matrix. To our knowledge there are very few publications in micro machining of nano-MMC which limits the industrial application of such material. Therefore, it is believed that there is a gap existing in the development of micro machining of MMC reinforced with nano-sized composites.

The aim of this paper is to investigate the influence of cutting parameters on the cutting force and surface morphology in micro machining of Mg based MMC reinforced with two types of nano-sized particles namely, titanium (Ti) and titanium diboride (TiB_2), as well as studying the size effect and finding out the minimum chip thickness according to the variation of surface roughness, surface morphology and specific cutting energy with the cutting parameter of feed per tooth. Finally, the experiment results are presented and discussed to deliver a comprehensive understanding on the micro machinability of Mg based MMC.

II. EXPERIMENTAL SET-UP

A. Machine Tool

Micro machining experiments were carried out on an ultra-precision desktop micro machine tool (MTS5R) which is fitted with a high speed spindle with the rotation speed ranging from 20,000rpm to 80,000rpm. This ultra-precision micro machine tool consists of 3 axes (X, Y, Z) with smallest feed of 0.1 μm . The axis movement is given in Fig. 1.

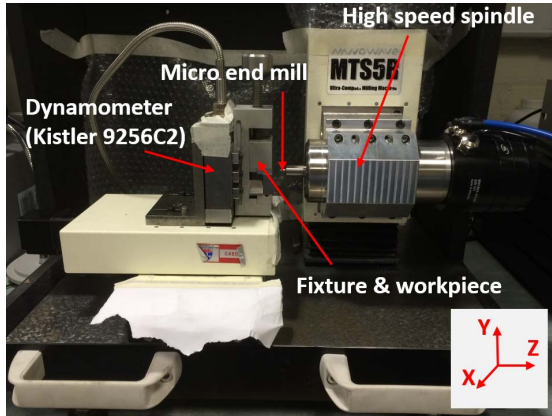


Figure 1. Configuration and machine tool with Kistler dynamometer

B. Workpiece Material

In this research, two specimens including Mg based MMC reinforced with 1.98 Vol.% of nano-sized Ti and TiB_2 were used. Mg particles with a size range from 60 - 300 μm and purity $\geq 98.5\%$ (supplied by Merck, Germany) were used as the matrix material, and Ti and TiB_2 with an average size of 50 nm were used as reinforcements. Pure magnesium powder was blended with the appropriate amount of reinforcements in a RETSCH PM-400 mechanical alloying machine at 200 rpm for 1 h. The homogenized powder mixtures of Mg and reinforcement were then cold compacted at a pressure of 1000 MPa to form billets of 40 mm in height and 35 mm in diameter. Monolithic magnesium was compacted using the same parameters without blending. The compacted billets were sintered using hybrid microwave sintering at 640°C in a 900 W, 2.45 GHz SHARP microwave oven. The sintered billets were then soaked at 400°C for 1 h and subsequently hot extruded at 350°C using an extrusion ratio of 20.25:1 to obtain rods of 8mm in diameter. Further details for the fabrication of the magnesium nanocomposites have been documented in earlier publications [7].

C. Experimental Procedure

Piezoelectric dynamometer (Kistler 9256C2) was used to acquire the cutting force along the X, Y, and Z axis during the machining. The surface roughness of the bottom machined surface was measured using Zygo white light interferometer. In addition, the micrographs for each machined surface were obtained using scanning electron microscope (Hitachi TM3030). 2-flutes UT coated tungsten carbide micro end mills with tool diameter of 1 mm were utilized. Three controlled quantitative factors were used in this experiment, namely, spindle speed n (rpm), feed per tooth f_t ($\mu\text{m}/\text{tooth}$) and depth of cut a_p (μm). Two individual full slot (1mm width and 5mm length) micro milling experiments were conducted

respectively on two types of Mg based MMC. There are two trials carried out in each experiment. Full factorial design was employed in the first set of trial, while 24 slots milling with various cutting conditions (illustrated in Table I) was used to study the effect of cutting parameters on the cutting force and surface roughness. Table II shows the cutting conditions in the second set of trial, there are 16 slots milling with various feed per edge and constant depth of cut and spindle speed with the aim of investigating the specific cutting energy and size effect.

TABLE I. CUTTING CONDITIONS FOR THE 1ST SET OF TRIAL

Cutting parameters	Level. 1	Level. 2	Level. 3	Level. 4
Spindle speed, n (rpm)	20000	40000	60000	N/A
Depth of cut, a_p (μm)	150	300	N/A	N/A
Feed per tooth, f_t (μm)	1	2	3	4

TABLE II. CUTTING CONDITIONS FOR THE 2ND SET OF TRIAL

Exp No.	Feed rate f_z (mm/min)	Spindle speed n (rpm)	Depth of cut a_p (μm)	Feed per tooth f_t ($\mu\text{m}/\text{tooth}$)
1	4	40000	150	0.05
2	8	40000	150	0.1
3	12	40000	150	0.15
4	16	40000	150	0.2
5	24	40000	150	0.3
6	32	40000	150	0.4
7	40	40000	150	0.5
8	64	40000	150	0.8
9	88	40000	150	1.1
10	112	40000	150	1.4
11	136	40000	150	1.7
12	160	40000	150	2
13	240	40000	150	3
14	320	40000	150	4
15	400	40000	150	5
16	480	40000	150	6

III. RESULTS AND DISCUSSIONS

A. Surface Roughness

1) Main effect of cutting parameters on surface roughness

Main effects of cutting parameters on the surface roughness for Mg/TiB₂ and Mg/Ti MMC were plotted in Fig. 2. (a) and (b) respectively. The surface roughness was observed to decrease when increasing the spindle speed from 20,000 to 40,000 rpm, then it increase rapidly when spindle speed increase to 60,000rpm for Mg/TiB₂, whereas an direct proportion relation was found between surface roughness and spindle speed for Mg/Ti. Lower depth of cut should be selected to minimize the surface roughness in machining of Mg/Ti MMC. For machining Mg/TiB₂, however the higher depth of cut results in a lower surface roughness value and this tendency is reverse to micro milling of metallic materials. An increase in feed per tooth from 1 to 3 $\mu\text{m}/\text{tooth}$ results in increase of surface roughness values in both materials, but a decrease was found at feed per tooth from 3 to 4 $\mu\text{m}/\text{tooth}$ for Mg/TiB₂. Over all, the machined surface

quality of Mg/TiB₂ MMC was more superior than that of Mg/Ti MMC in terms of surface roughness value.

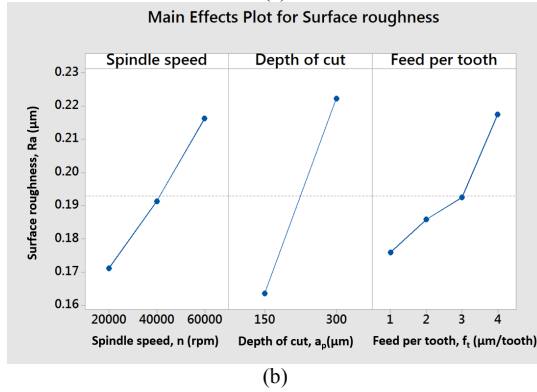
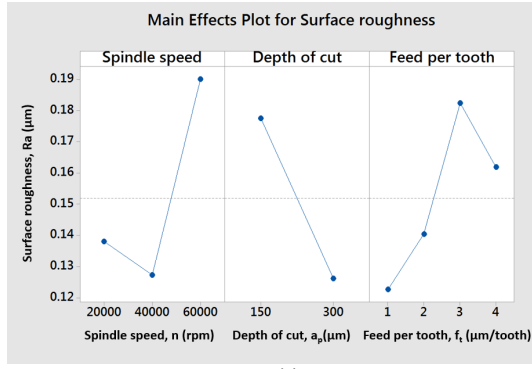


Figure 2. Main effects of cutting parameters on the surface roughness. (a) Mg/TiB₂ MMC. (b) Mg/Ti MMC

2) ANOVA

TABLE III. ANOVA FOR SURFACE ROUGHNESS OF 1ST SET OF TRIAL (MG BASED MMC WITH TiB₂ PARTICLES)

Source	DF	Sum of square (x10 ⁻³)	Mean of square (x10 ⁻³)	F Value	P Value	% Contribution
n	2	18.094	9.047	10.87	0.01	33%
a _p	1	15.794	15.794	18.97	0.005	29%
f _t	3	12.079	4.026	4.84	0.048	22%
n&a _p	2	0.624	0.312	0.37	0.702	1%
n&f _t	6	6.776	1.129	1.36	0.36	12%
f _t &a _p	3	1.397	0.466	0.56	0.661	3%
Error	6	4.994	0.832			
Total	23	45.363				

ANOVA for 1st set of trials in micro machining experiments of two materials was carried out to identify the effect of cutting parameters on the surface roughness. The ANOVA results shows in Table III and Table IV. It can be concluded that the depth of cut and spindle speed have significant influence on surface roughness in both trails. The contribution ratio for the spindle speed and depth of cut is 33% and 29 % for machining of specimen with TiB₂ particles, 19% and 49% for the machining of specimen with Ti particles.

TABLE IV. ANOVA FOR SURFACE ROUGHNESS OF 1ST SET OF TRIAL (MG BASED MMC WITH Ti PARTICLES)

Source	DF	Sum of square (x10 ⁻³)	Mean of square (x10 ⁻³)	F Value	P Value	% Contribution
n	2	8.146	4.073	9.06	0.015	19%
a _p	1	20.709	20.709	46.06	0.001	49%
f _t	3	5.64	1.88	4.18	0.064	13%
n&a _p	2	5.269	2.634	5.86	0.039	12%
n&f _t	6	1.677	0.279	0.62	0.711	4%
f _t &a _p	3	1.225	0.408	0.91	0.491	3%
Error	6	2.698	0.45			
Total	23	45.363				

B. Cutting Force

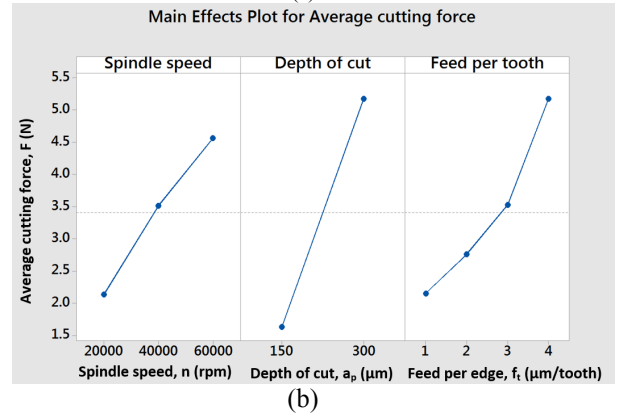
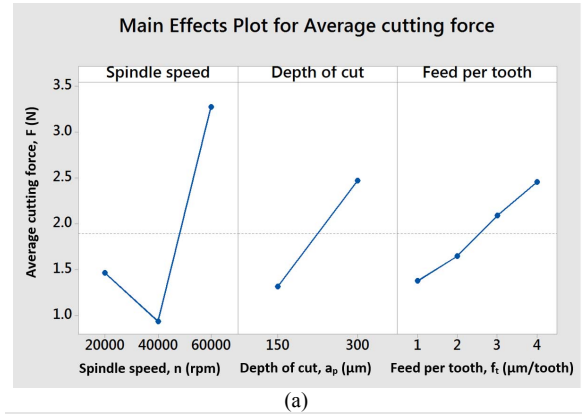


Figure 3. Variation of cutting force with three cutting parameters. (a) Mg/TiB₂ MMC. (b) Mg/Ti MMC

Measurement of cutting force is significant to more effectively understand the machinability and cutting mechanics of materials during machining. The main effect of three cutting parameters on the cutting force in machining of two types materials were illustrated in Fig. 3. Similar to micro milling ductile metallic materials, feed per tooth can be observed to be a most dominating parameter affecting the cutting force in both Mg/TiB₂ and Mg/Ti MMC. It can be concluded that the cutting force for Mg/Ti MMC seems to be more sensitive with the variation of depth of cut and feed per edge compared to Mg/TiB₂.

For Mg/TiB₂ MMC, the cutting force decreases when the spindle speed increase from 20,000rpm to 40,000rpm until a minimum value achieved, and a rapid increase in the cutting force is followed with the increasing of spindle speed from 40,000 rpm to 60,000 rpm. But for Mg/Ti MMC when spindle speed increase from 20,000rpm to 60,000rpm, there is an approximate linear increase in the average cutting force from 2N to 4.5N. Same tendency can be obtained in the variation of surface roughness with spindle speed

Lower level of depth of cut and feed per tooth therefore should be selected to minimize the cutting force while improve the machined surface quality. However, it is important to note that, the feed per tooth should not be smaller than the minimum uncut chip thickness otherwise cutting edge size effect would cause ploughing action governing the cutting process, which would deteriorate surface. Size effect will be introduced in section III C.

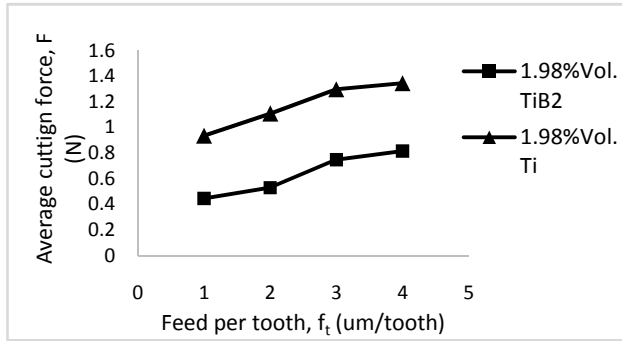


Figure 4. An influence of reinforcement materials on average cutting force

Fig.4. shows the comparison in the variation of cutting force with feed per edge between Mg/TiB₂ and Mg/Ti MMC at constant depth of cut of 150μm, spindle speed of 40,000rpm. The cutting force for the Mg/Ti MMC is much higher than that for Mg/TiB₂.

C. Size effect

1) Mechanism of cutting edge radius size effect

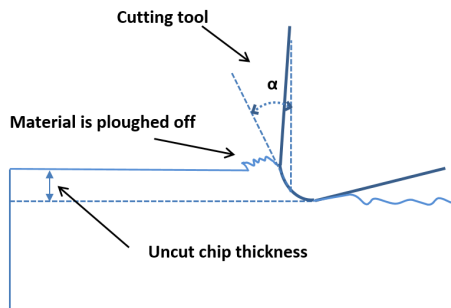


Figure 5. Ploughing action due to size effect

With the decreasing of uncut chip thickness which is dimensionally close to the cutting edge radius of the cutting tool, size effect becomes a dominant factor that leads a transitional regimes associated with intermittent

shearing and ploughing during the machining [6]. The material could be ploughed off, and elastic deformation would become the dominant cutting regime when the uncut chip thickness is below a critical value, chip may not be formed during each tool path which would deteriorate the surface (as illustrated in Fig. 5.) [8]. Essentially, the size effect can lead the higher cutting force, premature tool breakage, and worse machined surface quality. In this study, the investigation of size effect is carried out according to three aspects which are specific cutting energy, cutting force and surface morphology. Additionally, the minimum chip thickness are determined.

2) Specific cutting energy

Specific energy can be defined as the energy consumed in removing a unit volume of material. The specific cutting energy for each materials is calculated through expression (1), where F_x and F_y are the cutting force component in X and Y direction, v_c is the cutting speed in m/min, V_{rem} is removed chip volume, t_c is the cutting time.

$$u_c = \frac{v_c}{V_{rem}} \int_0^{t_c} \sqrt{F_x^2 + F_y^2} dt \quad (1)$$

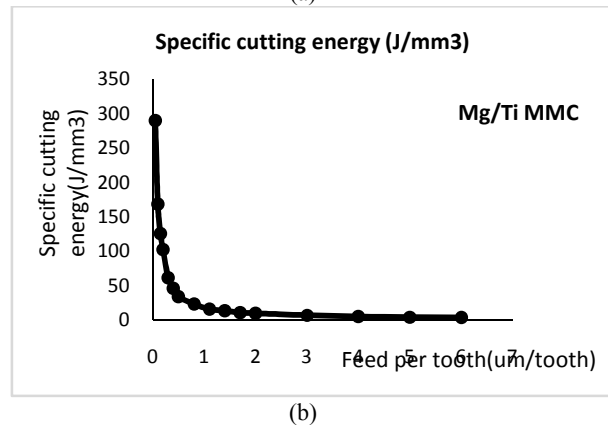
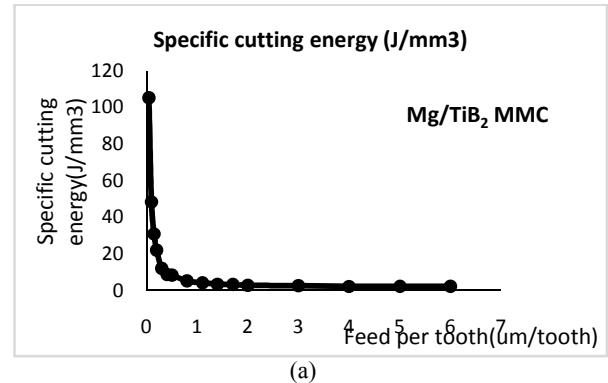


Figure 6. Variation of specific cutting energy with feed per tooth: (a) Mg/TiB₂; (b) Mg/Ti

Fig. 6. shows the variation of the specific cutting energy with feed per tooth for Mg/TiB₂ and Mg/Ti MMC at spindle speed of 40,000rpm, depth of cut of 200μm. The sign of size effect can be observed according to the

non-linear decrease of the specific cutting energy with feed per tooth. As illustrated in Fig. 6, as the ratio of feed per tooth to the cutting edge radius increases, the specific cutting energy decreases rapidly, and transit to a stable value at the feed per tooth between 1 and 2 $\mu\text{m}/\text{tooth}$. The high specific cutting energy at small feed is contributed to that the material undergoes an elastic deformation and partial deformed material fully recovers to its original position. A transition for the specific cutting energy approaching to a stable value is expected and is where the feed per tooth is close to the critical uncut chip thickness. Finally, when the feed per tooth to be larger than the critical chip thickness, the material deforms plastically and shearing become the major cutting action and continuous chips will form. In addition, by comparing Fig. 6. (a) and (b) it can be reported that much more energy is required to cut Mg/Ti MMC than Mg/TiB₂ when the feed per tooth is below the critical value. This leads to a worse machined slot edge for Mg/Ti MMC with more burrs as shown in section C. 4.

3) Variation of cutting force with feed per tooth

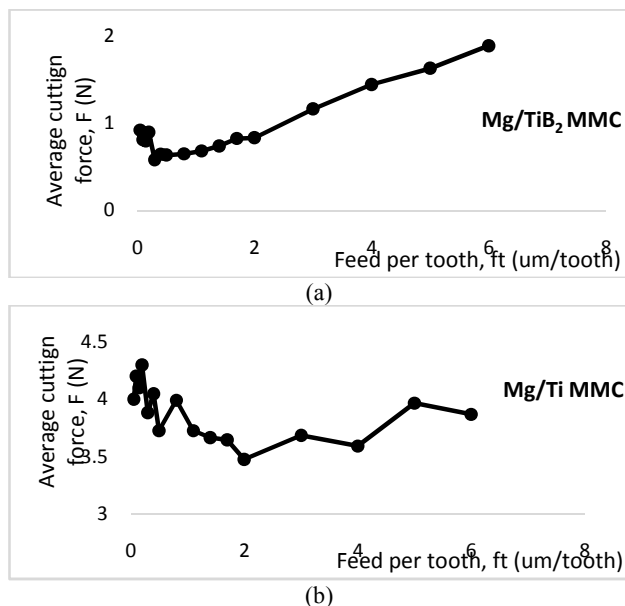


Figure 7. Variation of cutting force with feed per tooth at depth of cut: 200 μm ; spindle speed: 40,000rpm for (a) Mg/TiB₂ MMC. (b) Mg/Ti MMC

The phenomenon of the size effect can be also observed in the cutting force. Fig. 7. shows the variation of cutting force with feed per tooth for Mg/TiB₂ and Mg/Ti MMC respectively. It can be seen from Fig. 7, the cutting force at small feed per tooth is relatively large and it decrease to a minimum value at the feed per edge close to the critical value, then increase with the increase of feed per tooth. It shows good agreement with the previous research that the size effect would leads to a higher cutting force.

4) Machined surface morphology

Before conducting the qualitatively analysis, the specimens were cleaned by an ultrasonic cleaner with acetone with the aim of avoiding oxidation on the machined surface. The phenomenon of size effect can be detected according to the SEM micrographs of the machined surface.

Observation of machined surface in Fig. 8. revealed that for both Mg/TiB₂ and Mg/Ti MMC, various types of surface defects were found. Severe burr formation, worse surface and large area of cracks were observed on the slot edges and machined surfaces at small feed per tooth of 0.15 and 0.3 $\mu\text{m}/\text{tooth}$ (example are Fig. 8. (a)(b)(e)(f) and (k)). However, the amount of burrs on the slots and cracks formed decreases when the feed per tooth increases from 0.3 to 0.8 $\mu\text{m}/\text{tooth}$, and a machined surface without burrs and surface defects was obtained at feed per tooth of 0.8 $\mu\text{m}/\text{tooth}$ (example are Fig. 8. (d)(h) and (j)).

The result shows an excellent agreement with the finding that higher forces incurred in small feed per tooth due to size effect which may result in high value of surface roughness as well as the worse surface quality. It can be seen from Fig. 8. (k) that the material in the same position was cut more than once, which results in overlapping tool marks generated on the surface and indicates that the materials was elastically deformed due to the size effect. Therefore, based all the work done in this paper the minimum chip thickness for cutting Mg/TiB₂ and Mg/Ti MMC can be determined as 0.8 μm and the ratio of minimum chip thickness and cutting edge radius is approximately as 0.53%.

By comparing with two materials, more burrs on slots edge and worse surface quality with evenly distributed crack on the surface were found Mg/Ti MMC. As MMC materials are more brittle than the pure metal, it is very sensitive to impact at the moment which the material and tool come into contact, therefore the uncut chip thickness starts from zero and gradually increase to the maximum, so less burrs were formed at the up milling side according to Fig. 8. (f), (g) and (h). The cutting direction and the tool feed direction is opposite in both up and down milling, therefore the characteristics is different in surface and burr formation in the up milling and down milling [9].

IV. CONCLUSIONS

Micro machining of magnesium based metal matrix composites reinforced with 1.98 Vol.% nano-sized titanium and titanium diboride was performed by micro end milling process using UT coat carbide end mills. The cutting force was analysed and the machined surface were characterised by analysing surface morphology and surface roughness. Also, the size effect was investigated and minimum chip thickness was obtained according to specific cutting energy, cutting force and surface morphology. The following conclusion can be drawn from this work:

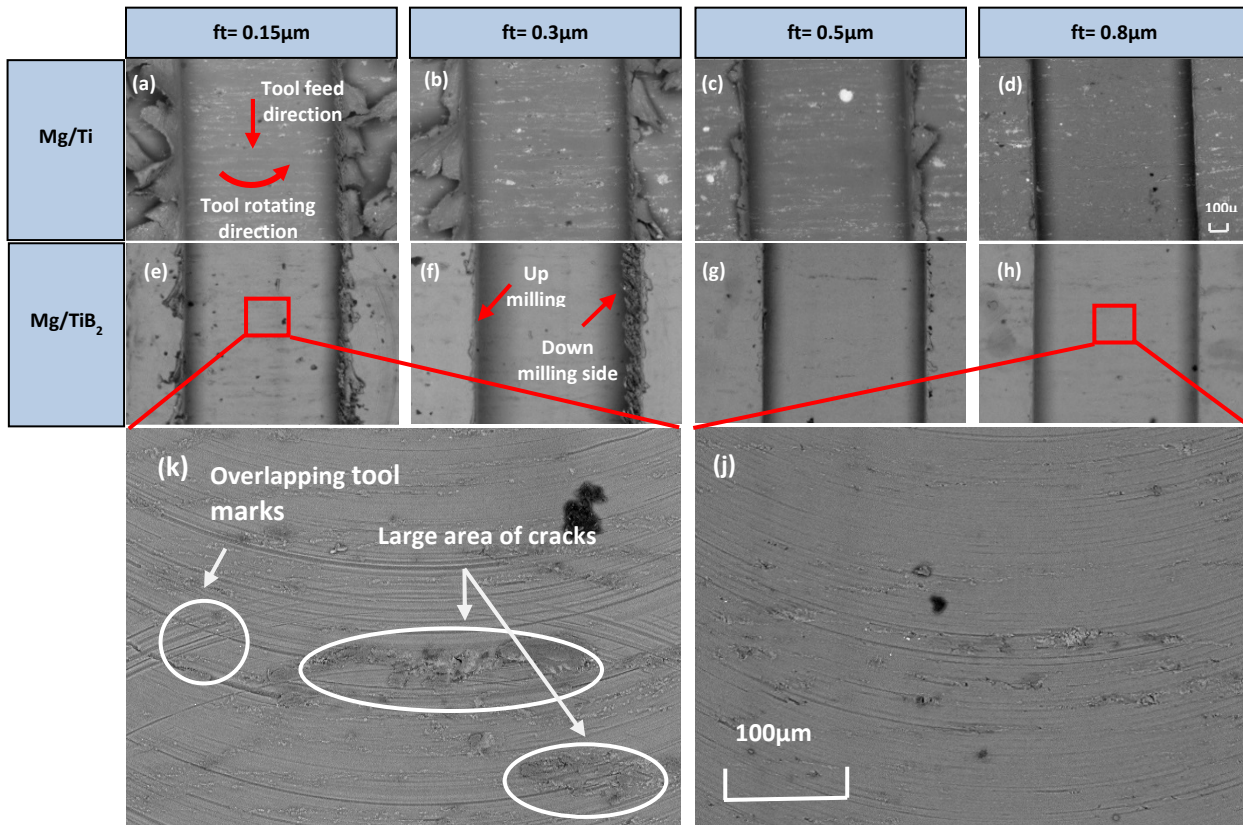


Figure 8. SEM micrographs of micro machined slots for Mg/TiB₂ MMC and Mg/ Ti MMC at different feed per tooth at spindle speed: 40,000rpm, depth of cut: 200μm.

- Based on the ANOVA analysis, the spindle speed and depth of cut have significant influence on surface roughness.
- The cutting force would increase with the depth of cut and feed per edge which is larger than the minimum chip thickness. An approximate 2 times larger cutting force was obtained in the machining of Mg/Ti MMC compared to that for Mg/TiB₂.
- The size effect was found and characterised by studying the specific cutting energy, cutting force and surface morphology, whilst the minimum chip thickness is determined by 0.8μm. Essentially, for machining Mg/Ti & TiB₂ MMC small feed per edge under 0.8 μm/tooth is not recommended.
- The Mg/TiB₂ MMC owns better machinability compared to Mg/Ti MMC in terms of surface morphology.

V. ACKNOWLEDGE

The authors wish to thank EPSRC (EP/M020357/1) for the support of this work. The authors are grateful for Dr. Ganesh Meenashisundaram (NUS) who synthesized the magnesium nanocomposites.

VI. REFERENCE

1. A. V. Beck, F. A. Hughes, The Technology of Magnesium and Its Alloys. F.A. Hughes & co. limited, 1943
2. F. Müller, and J. Monaghan, "Non-conventional machining of particle reinforced metal matrix composite," International Journal of Machine Tools and Manufacture, vol.40, pp. 1351-1366, July 2000.
3. Y. Yao, D. Li, and Z. Yuan, "Mill-grinding machining for particle reinforced aluminum matrix composites," Proceedings of The Seventh International Conference on Progress of Machining Technology, 2004.
4. A. Manna, and B. Bhattacharayya, "A study on machinability of Al/Sic-MMC," Journal of Materials Processing Technology, vol.140, pp. 711-716, September 2003.
5. K. Cheng, and D. Huo, Micro-Cutting: Fundamentals and Applications. Wiley, 2013.
6. X. Liu, R.E. Devor, S.G. Kappor, K.F. Ehmann, "The mechanics of machining at the microscale: assessment of the current state of the science," Journal of Manufacturing Science and Engineering, vol.126(4), pp. 666-678, November 2004
7. W. L. E. Wong, and M. Gupta, "Using microwave energy to synthesize light weight/energy saving magnesium based materials: a review," Technologies, vol.3, pp. 1-18, January 2015.
8. C. Kim, M. Bono, and J. Ni, "Experimental analysis of chip formation in micro-milling," Transactions of the North American Manufacturing Research Institute of SME, vol.30(159), pp. 247-254, March 2002.
9. J Liu, J Li, and C Xu, "Interaction of the cutting tools and the ceramic-reinforced metal matrix composites during micro-machining: a review," CIRP Journal of Manufacturing Science and Technology, vol.7, pp. 55-70, March 2014.

Computational Investigation of Superalloy Persistent Slip Bands Formation

Jianfeng Huang¹, Zhonglai Wang^{1,2}, Yuanxin Luo³, Yun Li⁴, Erfu Yang⁵, Yi Chen¹

¹School of Engineering and Built Environment, Glasgow Caledonian University, Glasgow G4 0BA, U.K.

²School of Mechatronics Engineering, University of Electronic Science and Technology of China, China

³College of Mechanical Engineering, Chongqing University, China

⁴School of Engineering, University of Glasgow, Glasgow G12 8LT, U.K.

⁵Space Mechatronic Systems Technology Laboratory (SMeSTech), Department of Design, Manufacture and Engineering Management, James Weir Building, University of Strathclyde, Glasgow G1 1XJ, U.K.

Abstract— Persistent slip bands (PSB) is the important and typical microstructure generated during fatigue crack initiation. Intensive works have been done to investigate the mechanisms of the formation of persistent slip bands in the past decade. In this paper, a molecular dynamics (MD) simulation associated with embedded atom model (EAM) is applied on the PSBs formation in nickel-base superalloys with different microstructure and temperature under tensile-tensile loadings. Simulation results show that PSBs formed within the γ phase by massive dislocations pile-up and propagation which can penetrate the grain. Also, the temperature will affect the material fatigue performance and blur PSBs appearance. The simulation results are in strong agreement with the experimental test results published before.

Keywords— Persistent Slip Bands; Molecular Dynamics; Superalloys; Computational Simulation

I. INTRODUCTION

Fatigue crack is a kind of mechanical fault which crashed suddenly but without evidence in advance. Since this characteristic of fatigue, it has been the major fault of the industrial product or building structure. The first report of fatigue behavior was in 1837 related to miner chain fracture caused by fatigue. Literally, fatigue is a kind of phenomena of crack propagation in the metal caused by the cyclic load for a long period of time. Fatigue is the most important factor of metal failure because it's unpredictable under working environment. There were many disasters in history caused by the fatigue failure of some working part. For example, the twice Comet airliner disasters occurred in 1954 lead to 56 people dead on board due to fatigue failure [1]. These crashes were caused by a small crack generated by metal fatigue near the radio direction, situated in the front of the cabin roof.

Because of the serious effect of fatigue on the safety, numerous researches have been done since the middle of 19th century. Many theories about metal fatigue have been formalized and new ideas about fatigue life prediction have been published, most important, many advanced instruments have been invented to measure and examine fatigue crack.

As extensive research of fatigue phenomena have been done, it is known that the general scheme of fatigue damage evolution in crystalline materials include three main phases - crack initiation period, crack propagation period, and

finally, the fracture. In many cases, the first phase of fatigue can be the longest period which is nearly 90% of fatigue life [2]. Research indicated that the main reason of fatigue crack initiation is persistent slip band in the material surface. Experiment evidence shows that small markers emerge on the specimen surface which is smooth at first after a cyclic load on the specimen for a period time. With the cyclic load increasing, the density of these markers increases accordingly. And also, these slip marker will occur again at the same area if the specimen continue under the fatigue test after surface polished. This kind of markers are so called the persistent slip band.

Considering the importance of persistent slip bands during the fatigue crack initiation, intensive works have been done to investigate the persistent slip bands formation and evolution. The common knowledge is PSB form after the initial hardening ceases with the dislocation matrix formed. PSBs do not emerge right during the first loading cycle till the matrix is formed. For example, in copper single crystals the PSBs begin to appear when the material hardening stops and the cyclic stress-strain response saturation reaches [3]. And also PSBs form related to the disparate strain ratio distributing in PSBs and matrix. And the shear plastic strain relationship [4] is:

$$\gamma = (1 - f)\gamma_m + f\gamma_{psb} \quad (1)$$

Where γ is the material shear plastic strain and γ_m , γ_{psb} represent the shear plastic strain of matrix and PSBs, f is the volume fraction of PSBs. PSBs are likely form at the range of plastic strain amplitudes of $1 \times 10^{-4} \leq \epsilon_{pa} \leq 1 \times 10^{-3}$ on individual cycled nickel polycrystalline grains having considerable different axial orientations [5].

The persistent slip bands are consist of extrusion and intrusion with ladder-like dislocation wall perpendicular to the primary slip direction. Figure 1 illustrates the typical

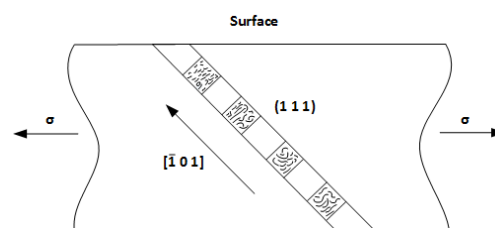


Figure 1. Unit Persistent Slip Band diagram

unit PSB along the $[1\ 0\ 1]$ direction on plane $(1\ 1\ 1)$. The thickness of the slab normally $1\sim 2\mu\text{m}$. The dislocation walls consist majority of edge dislocation dipole and the distance between the dislocation walls is regularly about $1.3\mu\text{m}$ [6].

When cyclic loads applied on the material, dislocations motivation display complex dynamics behaviors. Among the dislocation movement, some dislocations exhibit the movability while others are the inhabitant, the first ones called activator and the later called inhibitor. To evaluate the dislocation, a spatial and temporal based dislocation density function $\rho(x, t)$ provided by [7] [8]. The dislocation density function has two parts, one to describe the activator dislocations density which is the domination. The other part is to describe the inhibitor dislocation density which is a minor partition. The two functions are given below:

$$\frac{\partial \rho_i}{\partial t} = D_i \frac{\partial^2 \rho_i}{\partial x^2} + g(\rho_i) - \beta \rho_i + \gamma \rho_m \rho_i^2 \quad (2)$$

$$\frac{\partial \rho_m}{\partial t} = D_m \frac{\partial^2 \rho_m}{\partial x^2} + \beta \rho_i - \gamma \rho_m \rho_i \quad (3)$$

In the partial differential function, the D is the diffusion coefficient of both mobile and immobile dislocation. It is considered that the gradient of dislocation density of time is contributed by 1) the diffusion of dislocation, mobile and immobile separately 2) the transformation of immobile dislocations to mobile dislocation, and 3) the transformation of mobile dislocations to immobile dislocations. 4) Especially, in the function of immobile dislocation, additional a function of the generation of immobile dislocation by the applied shear stress considered.

Our purpose in this paper is to provide an atomic model to investigate the unit PSB formation mechanism. We will build up the atomic model at section III and introduce our simulation method in section IV. And finally we will analyze the simulation results and give the conclusion.

II. METHODOLOGY

Molecular dynamics is widely applied in chemical, material and biological science because of its special simulation ability on the study of the physical problem in atomic scale by solving a serial of ordinary differential functions, such as classical mechanics motion equations with Newton second law or the classical Hamilton function. Although MD is less precise than other simulation methods based on quantum theory such as density functional theory (DFT), it is popular and well developed since the good performance, and also the precise can be improved by adopting the potential calculated by DFT. The fundamental of MD is to study the macro physical properties by analyzing the discrete particles' trajectory which is calculated by solving these differential equations. Obviously, the molecular dynamics models consist of discrete particles, and besides, include their interactions potential which take action in a given ensemble with defined boundary conditions. It is known from the Newton's second law of motion that $F_i = m_i a_i$. And also the force can be derived from the gradient of potential function $\frac{\partial E_{tot}}{\partial r_i}$ with respect to atomic displacements.

$$F_i = -\frac{\partial E_{tot}}{\partial r_i} \quad (4)$$

Normally, the E_{tot} can be simply represented with a sum of pairwise interactions in which the famous one is Lennard-Jones (LJ) potential.

$$E_{tot} = \sum_i \sum_{j>i} \phi(|r_i - r_j|) \quad (5)$$

While in metal crystal system, the better representation of the energy is the following embedded atom method model potential which is a many-body interaction between atoms.

$$E_{tot} = \frac{1}{2} \sum_{i,j(i \neq j)} \Phi_{S_i S_j}(r_{ij}) + \sum_i F_{S_i}(\bar{\rho}_i) \quad (6)$$

In this EAM potential equation, the first term is the sum of all pair interactions between atoms similar with LJ potential, $\Phi_{S_i S_j}$ is a pair-interaction potential between atoms i and j in which have different chemical sort of S_i and S_j at positions r_i and $r_j = r_i + r_{ij}$. The second term is the sum of embedding energy of all atoms in the system. Function F_{S_i} denotes the energy of atom i , which depends upon the host electron density $\bar{\rho}_i$ at site i induced by all other atoms of the system. Based on this method, in 2009 [9] G.P. Purja Pun and Y. Mishin published a developed Ni-Al system EAM potential which demonstrates a fairly good agreement with experimental and ab initio results for the formation energies of several other compounds of the Ni-Al system.

With the advantage of molecular dynamics simulation, many works have been deployed to research the nature of the material in the atomics level. And many post analysis methodologies have been applied for case study. For example, M.H. Musazadeh and K. Dehghan [10] studied the crack propagation behavior in nanocrystalline Ni which contains different shapes and types of second phases with Pd Cu A and Ag. Their MD simulation found that the short cylindrical shape exhibited a minority effect compared to the long cylinder impurity by analyzing the crack growth rate and strain energy release rate. In 2D plane case, the strain energy release rate G is defined by:

$$G = \frac{EW\varepsilon^2}{2(1-\nu^2)} \quad (7)$$

Where E is Young's modulus and ν is Poisson's ratio, W is the width of simulation box, ε is the applied strain. Similar, in 2012, Paul White researched the fatigue crack growth in aluminum with the MD methods associated with isotropic linear elastic fracture mechanics (LEFMs). In this simulation, a cylinder atomic model with pre-notched crack containing nearly 3 million atoms simulated with three different potential (MFMP99 [14], LEA04 [15], SKCFC11 [16]) potential) under a fatigue load which ratio was $R=0$. The significant difference between those potentials was the measured stable fault energy γ_{SF} which affects the extent of dissociation of dislocation and unstable stacking fault energy γ_{USF} which represents the energy barrier to dislocation nucleation. The simulation and energy analysis revealed that with lower stacking fault energy (SKCFC11), dislocation emission occurred faster than that with others potential, and material with lower unstable stacking fault energy appeared strong stability thus the formation of dislocation became difficult [17].

Also, Po-Hsien Sung and Tei-Chen Chen [11] studied the crack growth and propagation behavior in single crystal

Ni by MD method associated with tight-binding potential. By analysis the von Mises stress and centrosymmetric parameter (CSP) distribution in the nanoribbon single crystal Ni, the partial dislocations slips was observed at the crack tip in the close-packed (1 1 1) plane till material fracture occurring and critical stress in single crystal Ni follows the order $\sigma_{[111]} > \sigma_{[100]} > \sigma_{[110]}$. Kai Zhou et al researched the effect of grain size and shape on mechanical properties of nanocrystalline copper with MD simulation by comparing the some macro properties of the material such as stress and Young's modulus. [12].

Besides the energy release rate analysis, stress - strain analysis and CSP analysis, Mao Wen et al [13] developed a deformation index (DI) during their research of the hydrogen embrittlement phenomena in single crystal Ni with EAM potential. The DI μ_i of atom i is defined as:

$$\mu_i = \max(|\vec{r}_{ij} - \vec{r}_{ij}^0|) / |\mathbf{b}|$$

Where \vec{r}_{ij} is the relative position vector of atom i and j and \vec{r}_{ij}^0 is the vector in reference lattice, \mathbf{b} is the Burgers vector of $1/2[110]$. With the DI analysis, the slip activities and dislocations can be traced and located immediately. Any lattice imperfections, torsion and structure changes can also be displayed clearly.

III. MODELING

To compare the effect of lattice orientation, phase and temperature on PSBs formation in Nickel alloy, three different group cases were studied in this paper. The first group includes two test cases, one with γ phase pure Ni, and another case, which was a benchmark test case, was γ phase associated with γ' phase of Ni_3Al . Both γ and γ' phase were face-centered cubic (FCC) lattice and the lattice constant of γ is 3.52 \AA and γ' was 3.572 \AA [18]. The lattice orientation in the simulation box was along direction $[111]$ $[\bar{1}01]$ $[1\bar{2}1]$ and in the plane of (1 1 1) which was the close package plane and atoms tend to slip. The simulation box was about $520 \text{ \AA} \times 200 \text{ \AA} \times 40 \text{ \AA}$. The unconstrained misfit [23] between the two regions of γ and γ' was 0.015 in the combined phases case. The second group included two test cases all with γ and γ' phase but the orientation was different, one was same as the test case in first group, but another one was along $[100]$ $[010]$ $[001]$ lattice direction. And the test cases in the final group was same as the benchmark test case in the first group but under a different temperature of 600K and 900K separately. Totally, nearly 300 thousand atoms created in all these simulation tests in which the grain size was about 10 nm. This was a very fine grain size expected with high fatigue resistance. Periodic boundary condition was set on the x y z directions which indicated the repeatable atoms topography along the directions. These simulations ran with Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) which is a kind of parallel computation system and visualized by OVITO. During the simulation of this homogenous box, at first, minimization of the system energy was run at the temperature of 300 K. The equilibrium of the system would make the whole system be in a stable condition. After that, a deformation rate of 0.0005 in y direction was applied in the simulation area and uneven velocity range from 0.46 to 0.98 was

applied on the different simulation groups along the y and reverse y direction for 1000 simulation steps separately with step time of 0.001 ps, and total simulation time in this load turn was 2 ps. The whole process repeated several times to apply the cyclic load on test systems.

IV. SIMULATION RESULT

A. Microstructure effect on PSB formation

To quantify the occurrence of defects in this simulation, the microstructure of the material was analyzed by adaptive common neighbor analysis (CNA) and centrosymmetric parameters (CSA) [19]. In single γ phase with the same strain rate and temperature, the dislocations were formed along the strain direction and scatter among the material. From the centrosymmetric parameter picture, the parameter P was greater than 8.3 \AA^2 in an FCC lattice indicated an intrinsic stacking fault [20]. The red particles in Fig. 3 a) shows the intrinsic stacking fault in pure single crystalline nickel generated along the $[\bar{1}01]$ direction in (1 1 1) plane. From Fig. 3 b), CNA analysis in the case indicate the pure FCC structure was only 81.2% while HCP was only 0.5%. That mean few intrinsic stacking fault in this system during deformation. But this dislocation behavior was totally

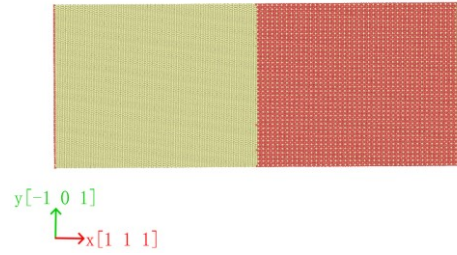


Figure 2. The Simulation Model

different against that occur in γ/γ' phase test case. When investigating the trajectories of particles in γ/γ' phase as

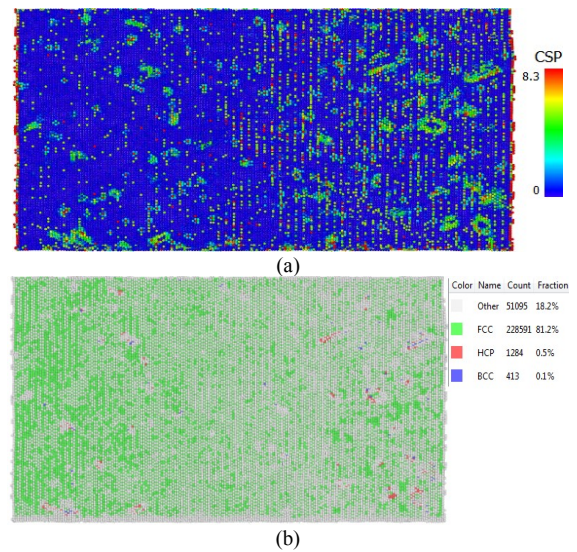


Figure 3 The CSP and CNA contour plot of γ phase pure nickel under fatigue load ratio (min load/max load)=0 after 40ps. a) is CSP display of γ phase pure nickel in which dislocation loop and dislocation slip along $[\bar{1}01]$ direction. b) is CNA display of γ phase pure nickel which shows the FCC structure is 81.2%

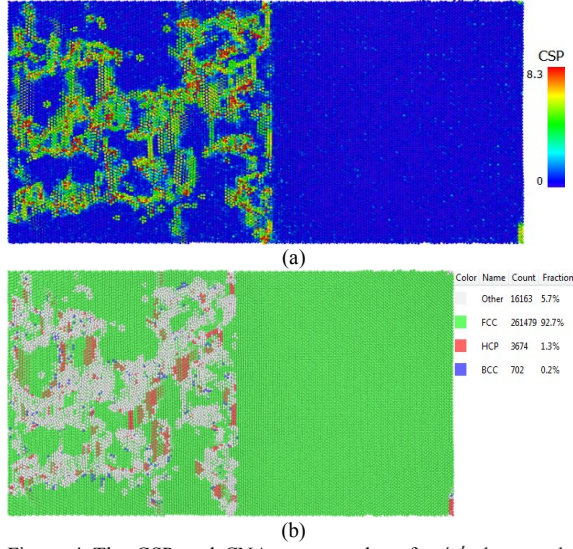


Figure 4 The CSP and CNA contour plot of γ/γ' phase under fatigue load ratio=0 after 40ps. a) is CSP display of γ phase in which dislocation pile-up occurs and dislocation slip along $[1 0 1]$ direction. but few dislocation in γ' phase b) is CNA display of γ and γ' phase which shows the FCC structure is 92.7%

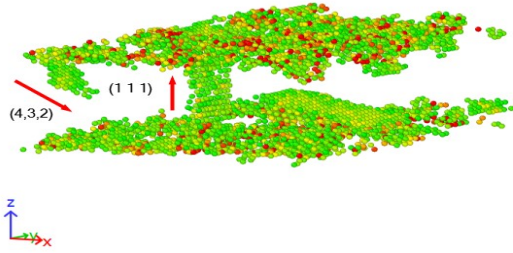


Figure 5 dislocation slip in $(4 3 2)$ and $(1 1 1)$ plane which penetrate the grain in γ phase and nucleated in the boundary of γ/γ' phase

Fig. 4 a) displayed, evidently, the γ phase was the activity basic matrix which likely to absorb energy by dislocations generation in the surface while γ' phase was the strength and stable participation which had less distortion during strain increasing in room temperature. Dislocations were likely to nucleate and slip along the y direction in γ phase. And coupled with the CNA in Fig. 4 b), the atoms which was not belong to FCC structure were nearly 7.3% in the whole system that demonstrate the dislocation volume. At the same time, some dislocations propagate and penetrate into the body by the slip in the plane $(4 3 2)$ and $(1 1 1)$. It was notable that, partial dislocations nucleated in the boundary of γ and γ' phase which demonstrated a clear view of the phase discrimination (Fig. 5).

B. The effect of orientation

When change the lattice parameter from $[1 1 1] [\bar{1} 0 1] [1 \bar{2} 1]$ to $[1 0 0] [0 1 0] [0 0 1]$ of lattice orientation, we found the only difference with the benchmark case was that the dislocation slips along 45 degree directions. By investigating the form process, we discovered, firstly, the intrinsic faults in the surface were introduced by the dislocation slips along the $[1 1 0]$ and $[1 \bar{1} 0]$ direction within the surface, and then, the formation of the dislocation line in these directions. With the cyclic loading continued, dislocation slip began to occur in planes of

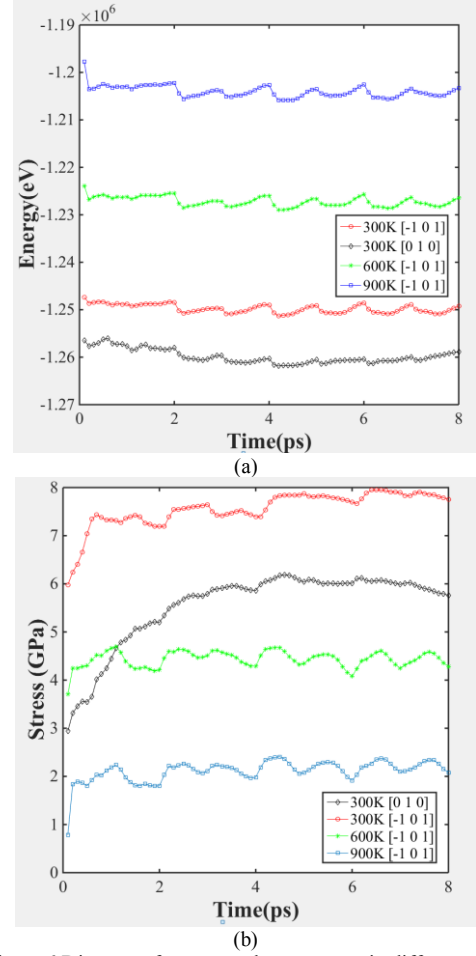


Figure 6 Diagram of energy and mean stress in different case

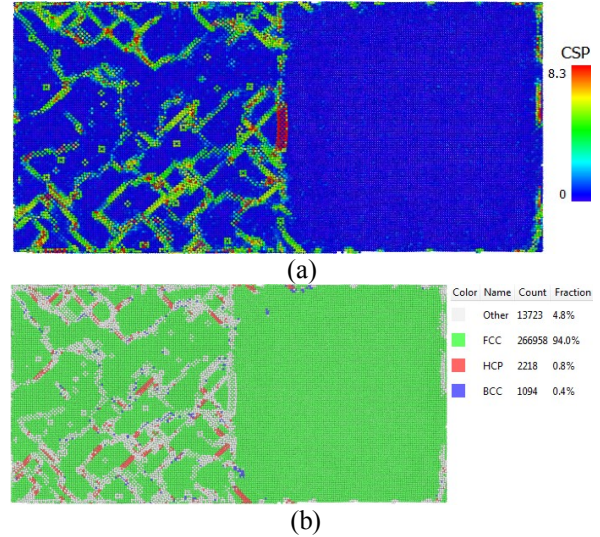


Figure 7 The CSP and CNA contour plot of γ/γ' phase under fatigue load ratio=0 along $[1 0 0]$ directions after 40ps. a) is CSP display of γ/γ' phase in which dislocation slip along 45° directions. But few dislocation in γ' phase b) is CNA display of γ/γ' phase which shows the FCC structure is 94.1%

$(\bar{1} 1 1)$ $(\bar{1} 1 \bar{1})$ and $(1 1 1)$ and infiltrated into the grain center. Once the dislocation slips crossed each other in the body, the intrinsic stacking fault was formed with the cyclic load increased. By comparing the stress (Fig. 6) in the test

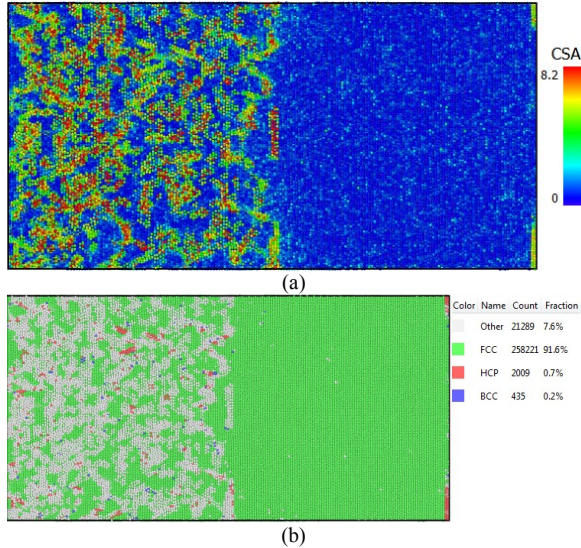


Figure 8 The CSP and CNA contour plot of γ/γ' phase under fatigue load ratio=0 during temperature 600K a) is CSP display of γ/γ' phase b) is CNA display of γ/γ' phase which shows the FCC structure is 91.6%

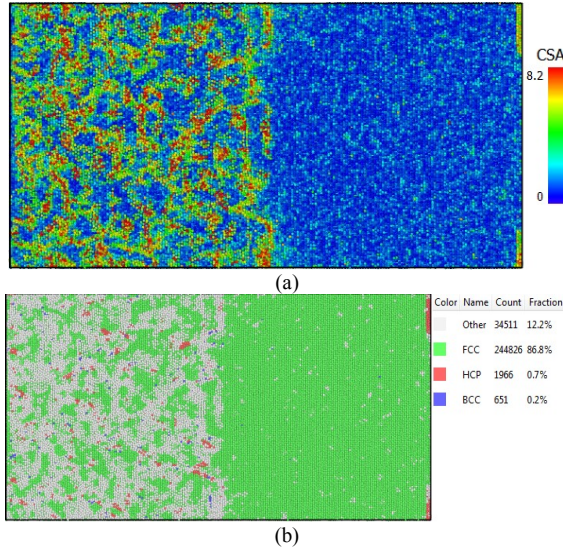


Figure 9 The CSP and CNA contour plot of γ/γ' phase under fatigue load ratio=0 during temperature 900K a) is CSP display of γ/γ' phase b) is CNA display of γ/γ' phase which shows the FCC structure is 86.8%

cases, it seemed that even in the test case with dislocation volume fraction was similar with that in benchmark test case, the mean stress with the same strain was much lower than that in benchmark since there was few dislocations pile-up with the cyclic load applied. Also in the benchmark test case, because of the load applied was perpendicular to the plane of (1 1 1) which was the closest package plane in FCC structure and caused strain along $[\bar{1} 0 1]$ direction, then greater magnitude of stress was needed to generate the same strain [21].

C. Temperature effect

As we know, high temperature fatigue always accompanies with creep behavior which caused by oxidation in the material surface. But, in this simulation, it only reflects the material subsurface microstructure changing. So the oxidation behavior was not included in

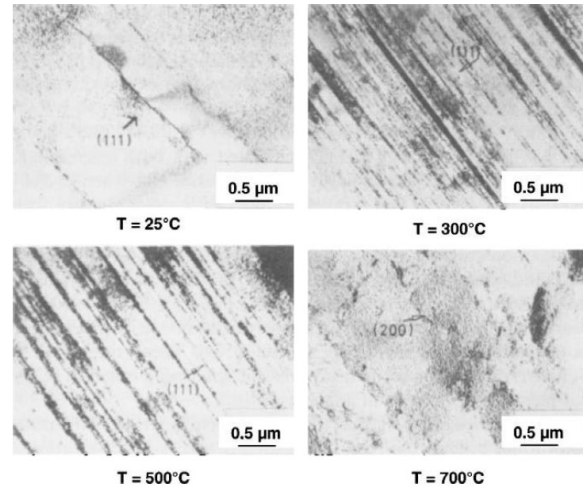


Figure 10 Slip band density as a function of temperature for Nimonic 80A tested at a plastic strain range of 0.15% [22]

this research scope. In this group test cases, with temperature increasing, thermal activation enhanced dislocations motivation, consequently, the dislocations became more homogeneous and dislocation density increased sharply. This could be shown in the last group test cases simulated under the elevated temperature at 600K and 900K separately. With the temperature increasing, the system became more unstable by the mean stress decreased because of system energy increased dramatically as Fig. 6 a) shows. And also some defects of γ' phase which were scattered also observed in these case but not at room temperature. It was notable that, the dislocations and defects in γ phase tend to pile up in the boundary of γ/γ' phase. Since the bonding energy decreases in high temperature which leads to atoms oscillating at a higher frequency and speed up, microstructure distortion became easier than that in room temperature. But at this situation, the intrinsic stacking fault was hard to form since the recovery mechanisms of γ/γ' at very high temperature. From the CSA and CNA graphic (Fig. 9) at 900K, we found the number of crystalline defects had raised up by comparing the pure FCC structure (86.8%), but few of them piled up in the material body. The agreement with the experimental result (Fig. 10) of Nimonic 80A indicates dislocations motivation under high temperature was active very much. And also with the reference from the mean stress graph, it revealed that, to generate such same strain, the stress turned to be much smaller rather than other situations.

V. CONCLUSION

From the MD simulation of Nickel alloy under different situations, following conclusions can be stated and it seems that the simulation results agreed with the experimental results:

1 γ' phase is the strengthening phase in Nickel alloy. During all of the simulation MD case, γ' phase is under few deformation, the PSB ladder like dislocation mostly occur in γ phase. Dislocations tend to pile-up and penetrate into γ phase which contribute to the fatigue crack initiation in Nickel alloy.

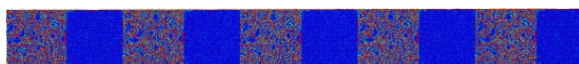


Figure 11 unit PSB formation by repeated boundary condition

2 Dislocation slip is the main activity during PSB formation. The dislocation slip behavior is observed in all test cases. This slip activity is more related to the temperature which cause system bond energy decrease. Once intrinsic stacking fault is formed in a material, dislocations tend to pile up in the close package plane in low temperature. With temperature increased, the dislocation has difficult to pin up because of the recovery mechanisms.

3 A different orientation have the different result of cyclic loading response. Simulation results indicate if the load is perpendicular to close package plane, dislocations will pile up and restrict the dislocation motivation and then enhance the fatigue resistance of the material.

Because of spatial and temporal limitation of MD simulation, it is difficult to simulate the whole PSBs formation mechanism in tiny atomic structure, but with the periodic boundary condition, we have repeated the simulation and it indicates the ladder like structure is formed during fatigue load as on Fig. 11. Our future work will concentrate on the simulation of fatigue crack initiation within multi-crystalline nickel-base alloy and compare with the experimental investigation.

ACKNOWLEDGMENT

The authors would like to acknowledge the scholarship award to the first author provided by the SCHOOL OF ENGINEERING AND BUILT ENVIRONMENT, GLASGOW CALEDONIAN UNIVERSITY. And this work was also supported by National Natural Science Foundation of China (Grant No. 51405044). The authors thank COLLEGE OF MECHANICAL ENGINEERING, CHONGQING UNIVERSITY. Finally, partial results were obtained using the EPSRC funded ARCHIE-WeSt High Performance Computer (www.archie-west.ac.uk). EPSRC grant no. EP/K000586/1. The author thanks the support provided by UNIVERSITY OF STRATHCLYDE.

REFERENCES

- [1] R. J. Atkinson, W. J. Winkworth and G. M. Norris, "Behaviour of Skin Fatigue Cracks at the Corners of Windows in a Comet I Fuselage" No. 3248, HER MAJESTY'S STATIONERY OFFICE, London.
- [2] I. Milne, R.O. Ritchie, B. Karihaloo "Comprehensive Structure Integrity, 4.01 Cyclic Deformation, Crack Initiation, and Low-Cycle Fatigue" vol. 4. Elsevier. pp. 7.
- [3] S. X. Li, Y. Li, G. Y. Li, J. H. Yang, Z. G. Wang and K. Lu "The early stages of fatigue and evolution of persistent slip bands in a copper single crystal" *Phil. Mag. A* vol. 82, issue 5, 2002 pp. 867-883.
- [4] A.T. Winter, "A model for the fatigue of copper at low plastic strain amplitudes" *Phil. Mag.* vol. 30, issue 4, 1974, pp. 719-738.
- [5] C. Buque. "Persistent slip bands in cyclically deformed nickel polycrystals", *International Journal of Fatigue*, vol. 23, issue 6, July 2001, pp. 459-466.
- [6] P. Lukáš, L. Kunz, "Role of persistent slip band in fatigue", *Phil. Mag.* vol. 84, issue 3-5, 2004, pp. 317-330.
- [7] D. Walgraef and E.C. Aifantis, "On the formation and stability of dislocation patterns-III: Three-dimensional considerations", *International Journal of Engineering Science*, vol. 23, issue 12, 1985 pp 1365-1372.
- [8] M.V. Glazov and C. Laird, "Size effects of dislocation patterning in fatigued metals" *Acta Meta. et Mat.*, vol. 43, issue 7, July 1995, pp. 2849-2857.
- [9] G.P. Purja Puna and Y. Mishin "Development of an interatomic potential for the Ni-Al system", *Phil. Mag.*, vol. 89, issue 34-36 2009, pp. 3245-3267.
- [10] M.H. Musazadeh, K. Dehghani "Molecular dynamic simulation of crack propagation in nanocrystalline Ni containing different shapes and types of second phases", *Computational Materials Science*, vol. 50, issue 11, Oct.-Nov. 2011, pp. 3075-3079.
- [11] Po-Hsien Sung, Tei-Chen Chen "Studies of crack growth and propagation of single-crystal nickel by molecular dynamics", *Computational Materials Science*, vol. 102, May 2015, pp. 151-158.
- [12] Kai Zhou, Bin Liu, Yijun Yao, and Kun Zhong "Effects of grain size and shape on mechanical properties of nanocrystalline copper investigated by molecular dynamics", *Materials Science and Engineering: A*, vol. 615, 6 Oct. 2014, pp. 92-97.
- [13] M. Wen, X.J. Xu, Y. Omura, S. Fukuyama, K. Yokogawa "Modeling of hydrogen embrittlement in single crystal Ni", *Computational Materials Science*, vol 30, issues 3-4, August 2004, pp. 202-211.
- [14] Y. Mishin, D. Farkas, M.J. Mehl, D.A. Papaconstantopoulos, "Interatomic potentials for monoatomic metals from experimental data and ab initio calculations". *Phys Rev B* 59, iss. 5, Feb. 1999, pp. 3393-3407.
- [15] X.Y. Liu, F. Ercolessi, J. Adams "Aluminium interatomic potential from density functional theory calculations with improved stacking fault energy", *Model. Simul. Mater. Sci. Eng.* 12, 665, 2004, pp. 665-670.
- [16] H.W. Sheng, M.J. Kramer, A. Cadien, T. Fujita, M.W. Chen, "Highly optimized embedded-atom-method potentials for fourteen fcc metals", *Phys. Rev. B* 83, iss. 13 Apr. 2011, pp. 134118.
- [17] Paul White, "Molecular dynamic modelling of fatigue crack growth in aluminium using LEFM boundary conditions", *International Journal of Fatigue*, vol. 44, Nov. 2012, pp. 141-150.
- [18] S. Boucetta, T. Chihi, B. Ghebouli, M. Fatmi, "First-principles study of the elastic and mechanical properties of Ni3Al under high pressure", *Materials Science-Poland*, vol. 28, No. 1, 2010, pp. 347-355.
- [19] Alexander Stukowski, "Structure identification methods for atomistic simulations of crystalline materials", *Modelling Simul. Mater. Sci. Eng.*, vol. 20, iss. 4, 2012, pp. 45021.
- [20] Helio Tsuzuki, Paulo S. Branicio, José P. Rino, "Structural characterization of deformed crystals by analysis of common atomic neighborhood", *Computer Physics Communications*, vol. 177, issue 6, 15 Sep. 2007, pp. 518-523.
- [21] Mikael Segersäll, Daniel Leidermark, Johan J. Moverare, "Influence of crystal orientation on the thermomechanical fatigue behaviour in a single-crystal superalloy", *Materials Science & Engineering A*, vol. 623, 19 Jan 2015, pp. 68-77.
- [22] Andre Pineau, Stephen D. Antolovich, "High temperature fatigue of nickel-base superalloys – A review with special emphasis on deformation modes and oxidation", *Engineering Failure Analysis*, vol. 16, 2009, pp. 2668-2697.
- [23] G. Brunetti, et. al. "Determination of $\gamma - \gamma'$ lattice misfit in a single-crystal nickel-based superalloy using convergent beam electron diffraction aided by finite element calculations", *Micron*, vol.43, issue 2-3, February 2012, pp. 396-406

Numerical Simulation of Triaxial Tests to Determine the Drucker-Prager Parameters of Silicon

Amir Mir¹ (DMEM), Xichun Luo¹, Amir Siddiq²
1- DMEM, University of Strathclyde, Glasgow, U.K
2- Dept. of Physical Sciences, University of Aberdeen, Aberdeen, U.K
Xichun.Luo@strath.ac.uk

Abstract: Finite element simulation of material response behavior under deformation entails identification of constitutive model parameters to truly expound the material behavior. Silicon is found to be hard and brittle and in the absence of experimental data, it is difficult to obtain constitutive model parameters to simulate the material deformation. In this paper numerical simulation of triaxial compression and triaxial tension tests are performed at different confining pressures and a method was adopted to determine the Drucker Prager parameters of silicon. The method involves extracting the data from stress-strain plots of triaxial compression and tension tests to calibrate the ultimate yield surface and then plotting the data in the meridional (p-t) stress plane.

Keywords- Drucker prager, triaxial test, silicon, finite element

I. INTRODUCTION

Silicon has great importance in opto-electronics, semiconductor, MEMS, space and defense industries due to its superb electro-mechanical properties, great high temperature strength and low thermal expansion. However, with these enviable characteristics it is correspondingly difficult to machine material and brittle fracture is an impediment to high surface quality during machining. High compressive strength and high stiffness of silicon make it hard to explicate its behavior under loading conditions. Numerical simulation has widely been adopted to understand the material response behavior under different conditions in order to avoid costly experimental techniques and trials. Drucker Prager model along with its optimized models have been successfully employed to simulate the material response behavior of pressure-dependent soil, rocks and concrete [1-2]. Experimental uniaxial and triaxial tests are required to obtain the constitutive parameters of materials for different versions of Drucker Prager model. No experimental triaxial compression and tension data is available for silicon in the author's knowledge. Yield's strength of silicon is also a contentious property as it has been reported from 350MPa to 7GPa [3-4]. In the absence of experimental data, parameter optimization techniques can be used to obtain these parameters. However the resultant parameters from optimization techniques are highly dependent on the initial assumption. Another approach is to perform finite element simulation

of experimental work to acquire required material parameters.

II. DRUCKER PRAGER MODEL

Since the von Mises yield criterion imply the dependence of material yielding solely on second deviatoric stress tensor J_2 and is independent of the first stress invariant I_1 , the yielding sensitivity to hydrostatic stress tensor is not incorporated for pressure-sensitive materials. Drucker and Prager in 1952 [5] proposed a model to address the effect of mean (hydrostatic) stress for pressure sensitive materials which von Mises yield criterion failed to address. The proposition acknowledged as Drucker-Prager (DP) model (also known as extended von Mises model).

Drucker Prager (DP) model explicate the material response behavior of granular-like soils, rocks and other alike pressure-dependent constitutive materials. The response behavior of pressure-dependent materials can be expressed in terms of strength increase with increasing pressure. Compressive strength of silicon is higher than its tensile strength [3] and under certain hydrostatic stress, the material is found to behave in ductile mode rather than brittle fracture [6]. This behavior clearly predicts increase in strength of silicon under loading conditions. In order to implement DP model to simulate deformation behavior of silicon, compressive crushing of concrete can be replaced by compressive plasticity of silicon and tensile dilatancy of concrete will be ignored [7].

DP theory in principle is also a modified form of Mohr-Coulomb's theory. The Drucker-Prager (DP) yield criterion is expressed as:

$$f(I_1, J_2) = \alpha I_1 + \sqrt{J_2} - d = 0 \quad (1)$$

Where I_1 is the first invariant of stress tensor and J_2 is the second invariant of the deviatoric stress tensor. α is the pressure sensitivity coefficient and d is known as the cohesion of the material. In DP model, the yield surface is the function of pressure and J_2 .

Since the finite element simulation was carried out in ABAQUS, the DP model representation will be followed as presented in this FEA software. The pressure-dependent linear DP yield function in Abaqus [8] is expressed in three stress invariants and inscribed as

$$f = t - p \tan \beta - c = 0 \quad (2)$$

Where p is the equivalent pressure stress and c is the material parameter known as the cohesion of the material. The term $\tan \beta$ represents the yielding sensitivity to hydrostatic pressure and β itself is the slope of the linear yield surface in meridional p - t stress plane and also known as friction angle of the material. The parameter t is deviatoric effective stress and expressed as

$$t = \frac{1}{2} q \left[1 + \frac{1}{k} - \left(1 - \frac{1}{k} \right) \left(\frac{f}{q} \right)^3 \right] \quad (3)$$

and for uniaxial compression

$$C = (1 - \frac{1}{3} \tan \beta) \sigma_c \quad (4)$$

Where K is the ratio of yield stress in the triaxial tension to triaxial compression, q is von Mises equivalent stress and f is the third invariant of deviatoric stress.

The evolution of equivalent plastic stain can be expounded using flow rule during deformation and provides the plastic strain relevance to stress components. Flow rule is stated in terms of plastic strain rate in the form of following equation

$$d\varepsilon_{ij}^p = d\lambda \frac{\partial f}{\partial \sigma_{ij}} \quad (5)$$

In Abaqus, the flow potential is written in the form as

$$g = t - p \tan \psi \quad (6)$$

Where g is the flow potential and ψ is dilation angle in the p - t plane.

The dilation angle ψ relates to the volumetric strain during plastic deformation and it remains constant during plastic yielding. For $\psi=0$ corresponds no volumetric strain, $\psi>0$ shows volume increase and $\psi<0$ signify reduction in volume. Silicon exhibit volume reductions of 20-25% [9] under loading when endures pressure induced phase transformation correspond to negative dilation angle.

III. TRIAXIAL COMPRESSION AND TENSION TESTS

Triaxial tests are essential to obtain mechanical properties and understand deformation mechanism of solids under pressure conditions. The deformation mode of pressure-dependent materials is highly reliant on the pressure gradient which instigates transformation into various structural phases. The true deformation behavior of pressure-dependent solids can only be understood by performing different types of triaxial test. There are various triaxial tests conducted for geological materials however very few are available for metals and ceramics. Sandia National Laboratories [10] performed uniaxial and triaxial tests under various loading conditions and

confining pressures to understand the behavior of SiC-N in the area of hypervelocity penetration of metal clad armour.

In the triaxial test, strength of the material is found to increase with the increasing confining pressure. The height to diameter ratio, confining pressure and loading rate significantly influence the stress-strain behavior of the material. Drucker Prager model is frequently adopted for rocks and concrete for which crack propagation is prevented under confining pressure in triaxial testing.

It is an established fact that during machining of silicon, the hydrostatic pressure is the governing factor result in increase in strength of material and entails ductile deformation rather than brittle fracture. The brittleness of silicon also disappears under hydrostatic pressure and material shows the ductile behavior. Axial strength of silicon increases significantly with the increasing lateral confining pressure and above certain pressure result in change of brittle behavior into ductile behavior of silicon.

In linear Drucker-Prager model, angle of friction β , flow stress ratio K , and dilation angle ψ are the target parameters to be identified. Triaxial compression and tension tests at different levels of confining pressures are required to obtain these parameters. The guideline of the methodology to determine the parameters is provided in [8].

IV. SIMULATION OF TRIAXIAL TEST

In triaxial test, a general approach is to use cylindrical specimens for balanced pressure. A conventional triaxial test equipment mainly consist of mechanical axial loading system, hydraulic confining pressure system, pressure and volume control modules, strain gauges to measure axial and radial deformation and data acquisition system to control and record the test data. For high strength materials such as silicon, high pressure is required to deform the material and therefore loading piston, bottom plate and sleeve material should be of high strength material. In this simulation, for simplicity, the geometric configuration of triaxial compression test simulation is a 2D axisymmetric part between the top and bottom rigid platens. The bottom platen is fixed while the top platen can move in the direction of loading. The axisymmetric part with its walls is modeled deformable and meshed with CAX4R element with reduced integration for axisymmetric stress analysis. Ductile-brittle transition in silicon is primarily due to hydrostatic stresses and temperature doesn't reach to the extent to cause thermal softening. Therefore thermal part of the simulation was not performed.

The height to diameter ratio of 2 is chosen as recommended for triaxial tests in order to reduce the geometry effect on the shear strength of material [11]. Fig.1 shows the schematic of 2D axisymmetric finite element model used.

An elastic modulus of 146GPa, poisson's ratio 0.27 and density 2329 kg/m³ was used. The constitutive behavior of silicon was modeled with numerically optimized Johnson's Cook (J-C) plasticity model.

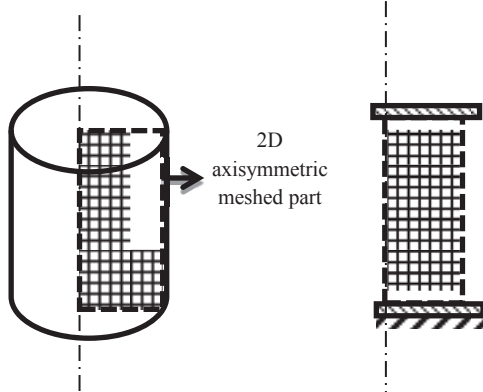


Figure 1. 2D axisymmetric model represent cylindrical specimen

In simulation, non-uniform mesh causes stress concentration in areas which doesn't undergo similar concentration in that area during experimentation. Therefore the mesh size was kept constant for the whole axisymmetric part

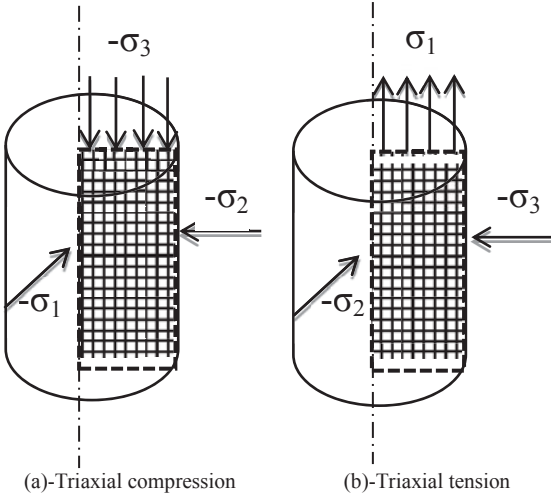


Figure 2. Triaxial compression and tension simulation

The displacement in axial and lateral directions is recorded in order to calibrate the volume change.

The axisymmetric part tested was subjected to 200,400, 1200, 2000, 3000, 4000 and 6000 (MPa) lateral confining pressures followed by an axial loading of 7500MPa. The test specimen undergoes constant fixed pressure stress throughout the axial loading. In the triaxial compression test, stress σ_3 represent the axial stress and confining pressure is represented by σ_1 and σ_2 . Fig.2 represents the pressure and loading conditions for triaxial compression and triaxial tension (initially pre-loaded) tests.

V. RESULTS AND ANALYSIS

The results obtained from the numerical simulation of triaxial compression and tension tests are analyzed and plotted to obtain the values of linear Drucker Prager

model. Stress-strain plot with different confining pressures is presented in Fig.3. It is clearly observed from the figure that increase in the confining pressure resulted in increase of elastic yield limit of the part and ultimately increased yield strength of material. At the confining pressure of 6GPa, the plastic deformation of the specimen disappeared and only the elastic limit was observed. The yield points on the onset of plastic deformation from each stress-strain curves were chosen in order to calibrate the ultimate yield surface.

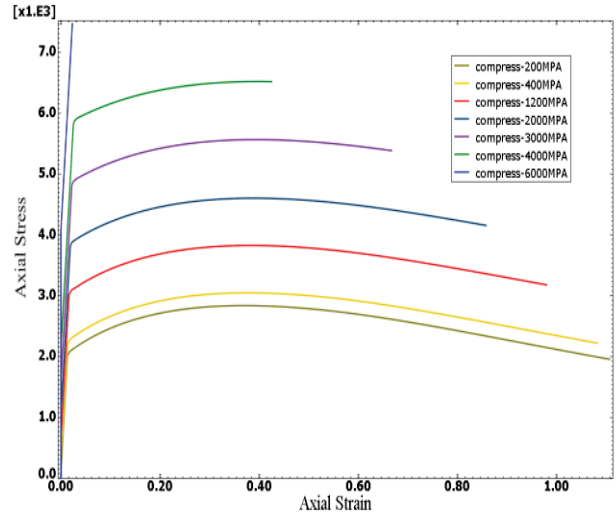


Figure 3. Axial Stress-strain plot with different confining pressures

The hydrostatic stress is expressed in the form

$$\text{Hydrostatic Stress} = \frac{1}{3}(\sigma_1 + \sigma_2 + \sigma_3) \quad (7)$$

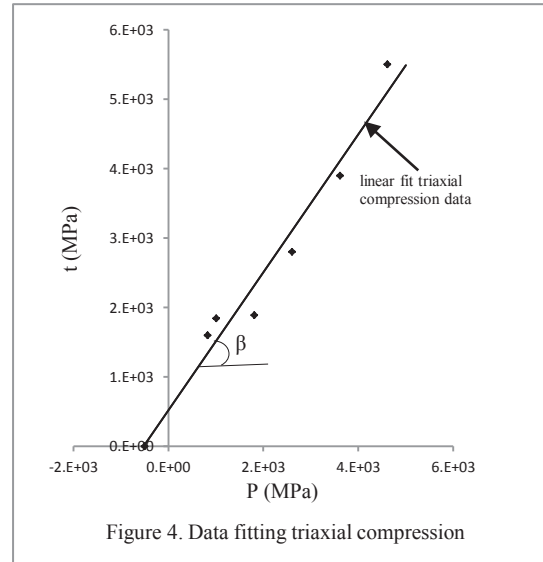


Figure 4. Data fitting triaxial compression

Since pressure is negative of hydrostatic stress, the pressure, P can be written as:

$$\text{Pressure} = -\frac{1}{3}(\sigma_1 + \sigma_2 + \sigma_3) \quad (8)$$

For triaxial compression,

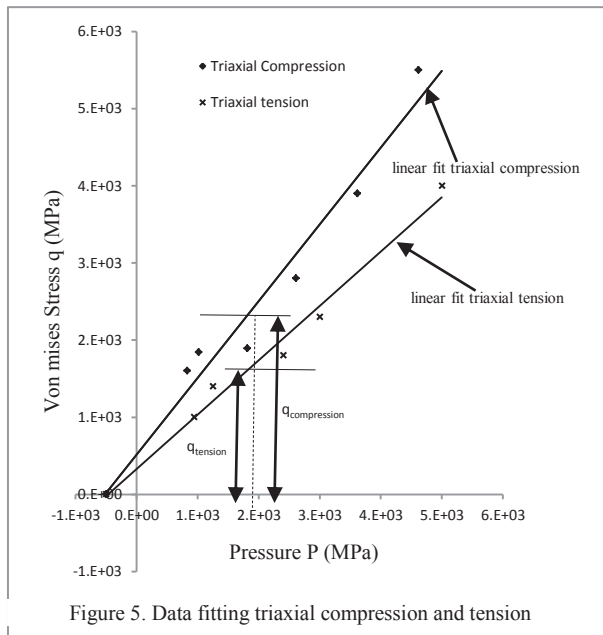
$$q = \sigma_1 - \sigma_3 \text{ and } t = q \quad (9)$$

Figure 4 is the yield surface calibration in the p-t plane from the triaxial compression results. From the compression result data, linear regression line was drawn. The values of β , and cohesion for the linear Drucker Prager model were obtained from the compression data linear trend line. The angle β is the angle made by the triaxial compression data line with the horizontal p axis and is calculated 44° with cohesion of 576MPa. The cohesion in metals and ceramics is different than usually measured for soils and therefore the cohesion obtained for silicon should be further investigated. The value of dilation angle was calculated 28.77° using flow potential equation. The dilation angle is dependent on the internal friction angle and is always less than the internal friction angle.

For triaxial tension test, the relationship of t with q is influenced by flow stress ratio

$$t = \frac{q}{K} \quad (10)$$

In order to find value of K, stress-strain data from triaxial compression and tension tests data was plotted on p-q plane. Fig 5 is the plot of linear regression of triaxial compression and tension stress-strain data plotted on p-q plane.



The flow stress ratio can be calculated from

$$K = \frac{q_{tension}}{q_{compression}} \quad (11)$$

Since in Abaqus the value of K has to follow the condition $0.78 < K < 1$, its value will be taken at pressure where K obey the condition. The value of K at pressure of 1910MPa is found to be 0.81.

VI. CONCLUSION

Numerical simulations of triaxial compression and tension tests were conducted on 2D axisymmetric model in order to obtain the linear Drucker prager model

parameters of silicon. The values of internal friction angle, flow stress ratio, cohesion and dilatancy angle were calculated by analyzing the result data. The obtained parameters can also be used to other parameters of exponent form of Drucker Prager model in Abaqus. Increase in confining pressure was found to increase the yield strength of silicon.

ACKNOWLEDGEMENT

The authors would like to thank EPSRC (EP/K018345/1) and Royal society-NSFC International exchange scheme (IE141422) to provide financial support to this research.

REFERENCES

1. Xiangjing, Huang, and Jiang Jianqing. 'Finite Element Analysis Of Triaxial Tests Of A New Composite Reinforced Soil'. *2010 International Conference on Intelligent Computation Technology and Automation* (2010): n. page
2. Shin, Hyunho et al. 'A Simulation-Based Determination Of Cap Parameters Of The Modified Drucker-Prager Cap Model By Considering Specimen Barreling During Conventional Triaxial Testing'. *Computational Materials Science* 100 (2015): 31-38
3. Okhrimenko, G. M. 'Single-Crystal Silicon, Piezoelectric Ceramics, And Ferrite Under Uniaxial Compression'. *Strength of Materials* 21.9 (1989): 1174-1180
4. Petersen, K.E. 'Silicon As A Mechanical Material'. *Proc. IEEE* 70.5 (1982): 420-457
5. Drucker, D.C., Prager, W., 1952. Soil mechanics and plastic analysis or limit design.
6. Zarudi, I. et al. 'The R8-BC8 Phases And Crystal Growth In Monocrystalline Silicon Under Microindentation With A Spherical Indenter'. *Journal of Materials Research* 19.01 (2004): 332-337
7. Wan, Haibo et al. 'A Plastic Damage Model For Finite Element Analysis Of Cracking Of Silicon Under Indentation'. *Journal of Materials Research* 25.11 (2010): 2224-2237
8. ABAQUS manual 6.13, User Documentation, software manual, Simulia
9. Kailer, A., Y. G. Gogotsi, and K. G. Nickel. 'Phase Transformations Of Silicon Caused By Contact Loading'. *J. Appl. Phys.* 81.7 (1997): 3057
10. Lee et al. "Uniaxial and triaxial compression tests of silicon carbide ceramics under quasi-static loading conditions", Sandia National Laboratories, Sandia Report (2005)
11. Jiang, Jia-Fei, and Yu-Fei Wu. 'Identification Of Material Parameters For Drucker-Prager Plasticity Model For FRP Confined Circular Concrete Columns'. *International Journal of Solids and Structures* 49.3-4 (2012): 445-456

The Design and Simulation of Beam Pumping Unit

Sun Wenlei¹, Cao Li¹, Qing Tao, Tan Yuanhua²,

¹School of Mechanical Engineering
Xinjiang University,
Xinjiang Urumqi, China

²Karamay Hong You Software Company
Xinjiang Karamay, China

Sunwenxj@163.com; 649823606@qq.com; xjutao@qq.com;

Abstract—Based on beam pumping unit, a new method of design beam hanger to dynamic loading and simulation in ADAMS software is presented, and beam pumping unit's structure and working principle are introduced in this paper first. Secondly, using the development function of macro command in ADAMS software, combining with the connect method of bushing and contact through theoretical analysis and calculation to develop beam hangers and managed to get the virtual prototype of beam pumping unit. Finally, using the function of spline to simulate the indicator diagram and loading into suspension point, making dynamic simulation in ADAMS software. The simulation results reflected the actual motion law of the horse head of beam pumping unit, proved the feasibility of the new modeling method and provided the theory base for further research.

Keywords- beam pumping unit; beam hanger; bushing; contact, dynamic simulation

I. INTRODUCTION

Beam pumping unit is the most important engineering machineries in petroleum equipment industry, and widely used in oil fields of our country [1]. In the process of traditional design, the performance of beam pumping unit can not accurately predict, so it is unable to make effective evaluation to the whole machinery. But the displacement, velocity and acceleration of suspension point are very important parameters in design and mechanical analysis of beam pumping unit, so we must be used to give a better simulation with the computer virtual prototype. And this provided strong basis to the following-up design, manufacture, research and development of beam pumping unit. And it can make more people pay more attention to the direction of reliability and security of petroleum machinery.

Due to the horse head of beam pumping unit uses adjustable arc, the effective length of the contact between beam hanger and horse head changing in the process of the movement. And the dynamic simulation software such as ADAMS and Pro/E don't have the right module to complete the design and simulation of beam hanger. At present, the researches in this area are mainly: Chen Yunxiao [2] developed heterogeneous CAD integration software of pumping units. Dai Yang [3] combined with Pro/E and ADAMS to build the soft rope to realize the simulation of this structure. Yang Youming [4] studied on the beam pumping units in CAD system. Yang Dongping [5] developed the dynamic simulation system of the double horse head pumping units in dynamic simulation

system of semi parameter flexible linkages of the double horse head pumping unit. Guo Tao [6] studied on the simulation of suspension point's motion parameters in ADAMS. But they didn't finish the dynamic loading and simulation of beam hanger. The dynamic loading and simulation of beam hanger has not been achieved which is based on the virtual prototype technology so far [7].

This paper proposes a new solution to solve the design and simulation of beam hanger based on the indicator diagram changing load, it broke through the technical bottleneck of virtual prototype of beam pumping unit, completed the dynamic loading in beam hanger's design, and carried on the virtual prototype simulation analysis to the model, the results reflected the rule of the actual movement of the beam pumping unit. The study had broad significance for the researchers to design oil facilities in the oil industry.

II. BEAM PUMPING UNIT'S STRUCTURE INTRODUCTION AND VIRTUAL PROTOTYPE MODELING PROJECT

A. The structure of Beam pumping unit

The study is on a beam pumping unit in this paper and it consists of many institutions such as crank, pitman, equalizer beam, walking beam, beam hanger, samson post, foundation and variable diameter circular arc horse head [8], the structure of beam pumping unit is shown in figure1.

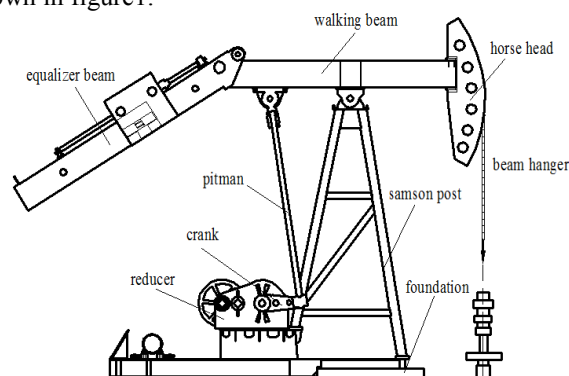


Figure 1 the model diagram of beam pumping units

The work theory of beam pumping unit is: after motor converts electrical energy into rotational motion, through two times speed reducer by the belt and retarder, then turn the rotary motion into reciprocating motion by the four bar linkage mechanism. Finally, through the

crank, pitman, horse head and the beam hanger drives the well pumping to pump oil.

B. The virtual prototype analysis of the beam pumping unit

In this paper, we build a virtual prototype model of beam pumping unit in ADAMS/View. We can build the model part of the beam pumping unit directly with the corresponding module except the beam hanger. In addition the stress and force application of beam hanger are very complex. And the length, contact point and the size of force which is contacting with horse head are changing over time. And the horse head is also affected by the radial and vertical pressure from the beam hanger, and contact point is changing in time. In this paper, used two times macro command function of ADAMS and combined with bushing and contact force to complete the design for dynamic loading to the beam hanger and get the virtual prototype model of beam pumping unit.

In order to make the model simple, we used Motion from ADAMS to simulate the motor function instead of real motor. The boundary and driving conditions of the beam pumping unit in ADAMS software are shown in sheet 1 and sheet 2.

Sheet 1 the boundary conditions of beam pumping unit

	Components	Boundary condition
1	Foundation	Fixed
2	Walking beam, horse head	Fixed
3	Walking beam, equalizer beam	Fixed
4	Samson post, foundation	Fixed
5	Walking beam, pitman	Revolute
6	Walking beam, samson post	Revolute
7	Pitman, crank	Revolute
8	Crank, reducer	Revolute
9	equalizer beam, clump weight	Translation

Sheet 2 the driving conditions of beam pumping unit

	Component	Driving condition
1	Crank	Rot joint motion
2	Clump weight	Trans joint motion

III. THE THEORY OF BEAM HANGER'S DESIGN

This paper proposed a new method to design beam hanger in ADAMS. Firstly, divided the beam hanger into many rigid cylinder, and every two rigid cylinders are connected by the bushing and contact force. And we must ensure every rigid body's kinematic parameters of (displacement at any time, velocity, acceleration etc.), the physical parameters (force, the moment of inertia etc.), and the kinetic parameters (the relative displacement, angle, action and reaction between each other etc.) are similar to the actual beam hanger as far as possible. When the length of each segment is too small than the total beam hanger, we can use this combination model approximately replace the beam hanger model. The specific modeling method of beam hanger will be introduced next.

A. Apply bushing in beam hanger

The force between each small cylindrical of the beam hanger is bushing. The bushing force is a kind of method applied to two component interactions, defined force and torque's six components $\{F_x, F_y, F_z, T_x, T_y, T_z\}$, applied a flexible force between the two components. The bushing force between every two small cylindrical is shown in figure 2.

$T_y, T_z\}$, applied a flexible force between the two components. The bushing force between every two small cylindrical is shown in figure 2.

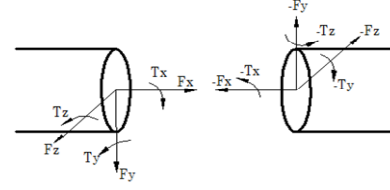


Figure 2 the bushing between small cylindrical

When used bushing force, built two coordinate marks at the force point of two interactional components. The first mark is i , and the next is j . The formula of the bushing force as following:

$$\begin{bmatrix} F_x \\ F_y \\ F_z \\ T_x \\ T_y \\ T_z \end{bmatrix} = \begin{bmatrix} K_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & K_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & K_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & K_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & K_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & K_{66} \end{bmatrix} \begin{bmatrix} R_x \\ R_y \\ R_z \\ \theta_x \\ \theta_y \\ \theta_z \end{bmatrix} + \begin{bmatrix} C_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & C_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & C_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & C_{66} \end{bmatrix} \begin{bmatrix} V_x \\ V_y \\ V_z \\ \dot{\theta}_x \\ \dot{\theta}_y \\ \dot{\theta}_z \end{bmatrix} + \begin{bmatrix} F_{x0} \\ F_{y0} \\ F_{z0} \\ T_{x0} \\ T_{y0} \\ T_{z0} \end{bmatrix} \quad (1)$$

In this formula: F is force; C is damping coefficient; K_{11} : The tensile stiffness coefficient; K_{22}, K_{33} : The shear stiffness coefficient; K_{44} : The torsion rigidity coefficient; K_{55}, K_{66} : The bend rigidity coefficient; R_x, R_y, R_z : The relative displacement from J -Marker coordinate system of the second component to the first one I -Marker; $\theta_x, \theta_y, \theta_z$: The relative angular displacement from J -Marker coordinate system of the second component to the first one I -Marker; V_x, V_y, V_z : The relative velocity from J -Marker coordinate system of the second component to the first one I -Marker; $\dot{\theta}_x, \dot{\theta}_y, \dot{\theta}_z$: The relative angular from J -Marker coordinate system of the second component to the first one I -Marker;

The reactive force of bushing can get from (2) (3):

$$F_j = -F_i \quad (2)$$

$$T_j = -T_i - \delta F_i \quad (3)$$

In this formula: δ is the torque coefficient of beam hanger; F_i and T_i are bushing force and torque of the component 1.

The flexible force of two components is closely related to the relative displacement, angle, velocity and angular velocity from the bushing force formula (1), (2), (3). The deformation, vibration and other physical properties, dynamics performance of model will be similar to the real beam hanger which we just need to control the stiffness coefficient and damping coefficient.

Major coefficient of bushing force can be calculated by the mechanics of materials, as shown in formula (4):

$$\begin{cases} K_{11} = \frac{EA}{L} \\ K_{22} = K_{33} = \frac{GA}{L} \\ K_{44} = \frac{G\pi D^4}{32L} \\ K_{55} = K_{66} = \frac{EI}{L} \end{cases} \quad (4)$$

In this formula: E is the elastic modulus of beam hanger, $E = 200GPa$; G is the shear modulus of beam hanger, $G = 80GPa$; A , D and L respectively, are cross section area of beam hanger, diameter and the length of the time of small pieces of steel wire rope, $A = 706.5mm^2$, $D = 30mm$, $L = 110mm$. I is the moment of inertia of each section of the beam hanger, $I = 3.97 \times 10^{-8} kg \cdot m^2$. So we can get:

$$\begin{cases} K_{11} = \frac{EA}{L} = 12.85 \times 10^8 N/m \\ K_{22} = K_{33} = \frac{GA}{L} = 5.138 \times 10^8 N/m \\ K_{44} = \frac{G\pi D^4}{32L} = 57804.54 N \cdot m/d \\ K_{55} = K_{66} = \frac{EI}{L} = \frac{E\pi D}{64L} = 2.68 \times 10^8 N \cdot m/d \end{cases}$$

B. Apply contact in beam hanger

In the process of beam hanger's movement, we can't ignore the collision force which is produced by the contact between horse head and beam hanger. Because of the limitation of the collision theory, the selection of contact collision force parameters generally based on experience or test. Before add contact force constraint between the beam hanger and horse head, we need to define the contact stiffness, collision force, maximum damping coefficient, the maximum penetration depth parameter and so on.

Impact function was used to describe the problem of the rigid body's collision in ADAMS software,

The direction of contact force is normal direction of two collision contact surface, the contact force as shown in formula (5):

$$IMPACT(x; \dot{x}, x_1, K, e, c_{max}, d) \quad (5)$$

In this formula: the contact stiffness K is the collision force; e is nonlinear collision force index; c_{max} is maximum damping coefficient; d is custom penetration depth.

So the contact stiffness can be calculated by formula (5):

$$K = \frac{1}{3} R^{1/2} E^* \quad (6)$$

In this formula: $1/E^* = (1 - \nu_1^2)/E_1 + 1 - \nu_2^2/E_2$, E_1 and E_2 are elastic modulus of two contact materials, $E_1 = E_2 = 200GPa$; ν_1 and ν_2 are position ratio of two contact materials, $\nu_1 = \nu_2 = 0.3$; R_1 and R_2 are equivalent radius of two contact materials, $1/R = 1/R_1 + 1/R_2$,

$R_1 = 1054.5mm$, $R_2 = 15mm$; so we can calculate the contact stiffness: $K = 1.63 \times 10^8 N/m^{1.5}$.

The contact force between beam hanger and horse head in the virtual prototype model as follows: contact stiffness is $K = 1.63 \times 10^8 N/m^{1.5}$, nonlinear index is 1.5, the maximum damping coefficient and the maximum penetrating depth in maximum damping are $10 N \cdot s/mm$ and $0.1 mm$.

According to the above theory, we can get the physical model of the beam hanger as shown in figure 3.

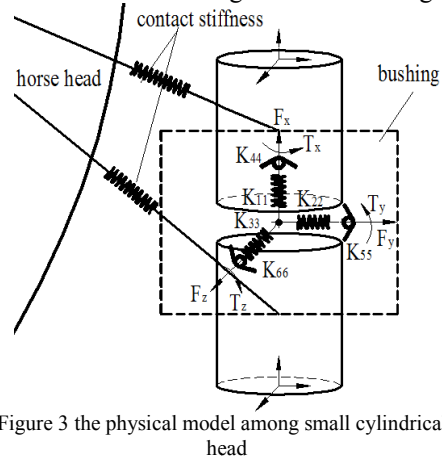


Figure 3 the physical model among small cylindrical, horse head

IV. THE SIMULATION ANALYSIS OF BEAM PUMPING UNITS BASED ON INDICATOR DIAGRAM

The load of suspension point is one of the important parameters to estimate the working ability of beam pumping unit, and it is the base of the mechanics analysis. Indicator diagram on suspension point as the picture 4 shown, the speed and acceleration of suspension point's size not only changed, but also direction.

The first of the up strokes is accelerated motion and the acceleration direction to downward; the last is decelerated motion, acceleration direction to upwards. The down stroke is the accelerated motion and the acceleration direction to downward at first; the last is the slow motion, acceleration direction upwards.

The indicator diagram is a changing curve of the force and displacement, and one displacement corresponding two loading. It is dispersed into a series of points in ADAMS and then fit a curve to load into the suspension point.

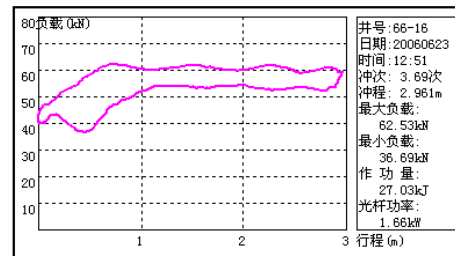


Figure 4 indicator diagram of beam pumping units

This article selected the working condition as follows: motor speed is $n = 720 r/min$, stroke is $2.1m$. Because of the collision force between the horse head and the beam

hanger, it has a high demand in the shape of the model during motion simulation analysis. The model should be simplified the rotating components before simulation. We also need to adjust the stiffness coefficient until it is appropriate. If the beam hanger cut in gears, we should increase the contact rigidity; if the beam hanger vibrate violent or be flied out, we should decrease the contact rigidity.

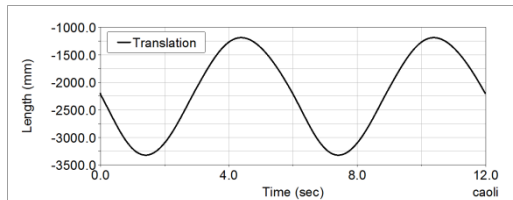


Figure 5 suspension point displacement curve of beam pumping units

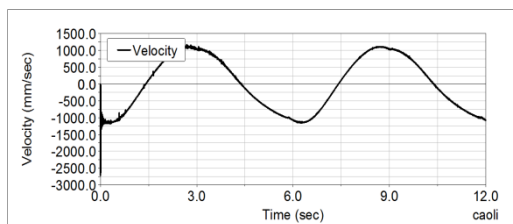


Figure 6 suspension point velocity curve of beam pumping units

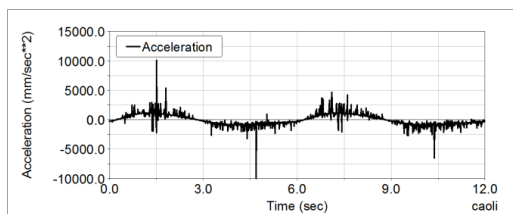


Figure 7 suspension point acceleration curve of beam pumping units

Figure 5 is suspension point displacement curve as a periodic sine curve of beam pumping unit. The down-stroke displacement is -3396.9637mm and the up-stroke displacement is -1198.3276mm. The total displacement is 2.0864m, and there was 0.0136m between design displacement and actual displacement because of some assembly and design error.

Figure 6 is the velocity curve of the beam pumping unit's suspension point. From the figure, we can see when the suspension point reaches the equilibrium position, the speed reaches the maximum, and when the suspension

point to the bottom dead center position, the speed reaches the minimum.

Figure 7 is the acceleration curve of the beam pumping unit's suspension point. It's a periodic oscillation and irregular sine curve. The figure shows the vibration tends to be more violent in the lower dead point.

Therefore, the simulation analysis of suspension point's motion characteristic curve reflects the real motion law of beam pumping unit, which proves the feasibility and validity of the digital virtual prototype of beam pumping unit.

CONCLUSION

In this paper, using macro command in ADAMS and combining with bushing and contact to complete the dynamic loading design of beam hanger and the loading of indicator diagram. The analysis proved the feasibility of this approach that beam pumping unit dynamic load design and simulation based on the indicator diagram of variable load. And the paper provided the theory basis for beam pumping unit's further research.

REFERENCES

- [1] Yang You-Ming, Li Ye-Fei, Chen Ying-Hua. CAD system for beam pumping unit[J]. Drilling Process, 2002, 25(6): 60-62.
- [2] Chen Yu-Xiao, Yu Qing-Zhong, Zhen Jun-Sheng. Interphase pumping unit integrated CAD software research [J]. Petroleum Machinery, 2002, 31(9): 57-59.
- [3] Yang Dng-Ping, Go Xe-Shi, Di Yng. Dynamic simulation system of variable parameter flexible linkage mechanism of dal horse head pumping unit[J]. Journal of Mechanical Engineering, 2010, 09(10): 59-65.
- [4] Guo Tao, Sun Wen-lei, Gao Ji-hong. Research of analog simulation on suspension movement parameter of walking beam type pumping unit[J]. machine with hydraulic, 2009, 07(8): 224-225+257.
- [5] Liu Zhen-Yu, Tan Jian-Rong. Research on process oriented assembly modeling in virtual environment[J]. Journal of Mechanical Engineering, 2004, 40(3): 93-99.
- [6] Zi Bin, Duan Bao-Yan, Du Jin-Li. Dynamic modeling and numerical simulation of cable-driven parallel manipulator[J]. Journal of Mechanical Engineering, 2007, 43(11): 82-88.
- [7] Yng Dng-ping, Go Xe-shi, Di Yng. Dynamic simulation system of variable parameter flexible linkage mechanism of dal horse head pumping unit[J]. Journal of Mechanical Engineering, 2010, 09(5): 59-65.

Quad-rotor Lifting-Transporting Cable-Suspended Payloads Control

Yaser Alothman
School of Computer Science
and Electronic Engineering
University of Essex, Colchester, UK
Email: ynalao@essex.ac.uk

Wesam Jasim
School of Computer Science
and Electronic Engineering
University of Essex, Colchester, UK
Email: wmjasi@essex.ac.uk

Dongbing Gu
School of Computer Science
and Electronic Engineering
University of Essex, Colchester, UK
Email: dgu@essex.ac.uk

Abstract—This paper presents the control of quadrotor UAV with cable-suspended stability. A linear quadratic regulator (LQR) control algorithm is proposed for lifting and transporting the load. The nonlinear dynamic model of the vehicle is represented with considering the cable-suspended load in eight degree of freedom, then the model is linearized at the hovering point. Two modes of taking-off are used, starting with taking-off without the load effect then switching to taking-off with the effect of load. The simulation presents the results to show the system stability and verify the LQR gains. The results are compared with the PD controller results.

I. INTRODUCTION

In recent years, the Unmanned Aerial Vehicles (UAVs) have significantly become common aerial robotics for researchers and it has been implemented in different applicable situations. Transportation of a suspended load via a quadrotor has been of the influential field of research and application. However, because of the load is fluctuation during the transportation proceeding, controlling the quadrotor and the suspended load is complicated. Moreover, the quadrotor system is unstable with high nonlinearities. Several quadrotor capacities such as vertical take-off, landing, single-point hover, provide it with the exemplary model for transporting autonomously the cargo delivery [1].

Many researcher's works have been published in order to solve the problem of lifting and transporting stability. For instance, Sadr, S. et al.[2] developed a dynamic model system and designed a nonlinear controller for position and attitude of a quadrotor with slung load based on an anti-swing algorithm. An adaptive controller had been designed based on a combination between the least-square estimation and the geometric control to transport the load from point to the other [3]. A combination of adaptive PD and neural network controller was proposed in [4] to cancel the effect of unmodeled dynamics. While a nonlinear PD controller was presented in [5] for stabilization and trajectory following the problem. The effect of the payload mass uncertainty was compensated using a retrospective cost adaptive controller.

Authors in [6] were proposed a baseline controller for the quadrotor with suspended load to generate a trajectory with swing free maneuver. An adaptive controller was added to cover the change in the center of gravity. A batch reinforcement learning approach was implemented in [1] to overcome

the problem of swing-free trajectory generation. The proposed controller was tested in a real vehicle to verifying its validity. A technique based on dynamic programming was presented in [7] for swing-free trajectory tracking of a quadrotor with cable-suspended load. The presented controller was tested practically to illustrate the model performance.

A nonlinear control technique was based on the dynamic model was proposed in [8] for a swing free stabilization and trajectory tracking of a quadrotor with a suspended load. The dynamic model was represented via eight degree of freedom. The simulation results of the presented controller illustrate the improvement of the quadrotor performance. A three stages geometric nonlinear control strategy was implemented in [9] for a quadrotor with suspended load path tracking. The quadrotor dynamic model was represented in eight degree of freedom and it was based a differentially-flatness. The hybrid control system stability analysis was guaranteed for stabilization and path tracking tasks.

Another distinct load transportation approach is a cooperation of multi-quadrotor, in which one load is suspended by more than one vehicle in different ways. Authors in [10] were proposed a differential flatness method for trajectory planning for the suspended load transportation problem. The load was suspended by three quadrotors via three cables. A simple PD controller was presented in [9] for tracking and formation control of a team of quadrotors with a suspended point mass load. A geometric feedback controller was implemented in [11] to track a predefined trajectory of the load attitude and position. The load was suspended by multiple quadrotor as a point mass. The proposed controller was used to control the quadrotors yaw angle as well.

In this paper, optimal LQR and PD controllers are implemented for attitude and translation stabilization in lifting and transporting tasks for a single quadrotor carrying a specific load by a flexible cable. To the best of our knowledge, this is the first use of LQR approach for single quadrotor with suspended load control.

In the following, Section II presents the quadrotor with the cable-suspended load mathematical model. Section III illustrates the LQR control approach fundamentals. Section IV describes simulation results. Our conclusion and future work are given in Section V.

II. QUADROTOR WITH SUSPENDED LOAD MODEL

A. Nonlinear Dynamic Model

The full quadrotor and payload system has been presented as a translation-rotation dynamic model. We consider that the dynamic model of two systems is subjected to specific following assumptions

- 1) symmetrical rigid body of quadrotor structure with four propellers which generate both torque and thrust for each.
- 2) The air drag is negligible.
- 3) A point mass is suspended by a massless cable attached at the center of quadrotor.

The point mass load suspended with single quadrotor is demonstrated by a derivation of a dynamic model according to the above assumptions. Figure 1 illustrate the representation of the dynamic model coordinate of the quadrotor with carrying payload suspended by a cable. There are two coordinate reference frames, inertial frame (earth fixed frame) denoted by **E** and rigid body fixed frame denoted by **B**. Their coordinate positions are denoted as x_E, y_E, z_E and x_B, y_B, z_B respectively. The payload attitude is represented in two spheres and its position is respected to the quadrotor position. Symbols and acronyms are listed as

R	Rotation matrix from inertia frame to body frame
ϕ	Roll angle along x-coordinate of the quadrotor
θ	Pitch angle along y-coordinate of the quadrotor
ψ	Yaw angle along z-coordinate of the quadrotor
ϕ_L	Roll angle along x-coordinate of the suspended load
θ_L	Pitch angle along y-coordinate of the suspended load
$\chi_Q, \chi_L \in R^3$	Positions of the quadrotor center of gravity and load
$v_Q, v_L \in R^3$	Velocity of the quadrotor center of gravity and load
F, M	Total thrust and moment produced by the quadrotor
Ω	Angular velocity of the quadrotor
I	Inertia matrix of the quadrotor
τ	Torque on airframe body
b	Thrust factor
d	Drag factor
L	Cable length
T	Cable tension
m_Q	Mass of the quadrotor
m_L	Mass of the suspended load
e_1, e_2, e_3	Three coordinate unit vectors
ρ	Unit vector from quadrotor attached point to the load

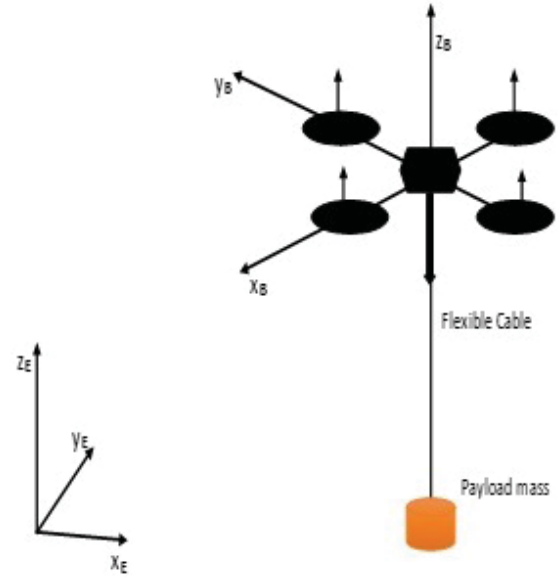


Fig. 1. Quadrotor carrying payload with cable

The cable suspended load is modeled with eight degree of freedom, which is comprised of six for the quadrotor as a rigid body and the rest for the spherical pendulum. Because of the change of the cable tension from slack to taut, two mathematical models are taken into account; a quadrotor with no load and with cable-suspended load dynamic models. A switching process is used to transfer the operation from first to second model depends on the cable tension situation. The equations of motion for no load are modeled as

$$\dot{\chi}_Q = v_Q \quad (1)$$

$$m_Q \dot{v}_Q = F R e_3 - m_Q g e_3 \quad (2)$$

$$I \dot{\Omega} = -\Omega \times I \Omega + M \quad (3)$$

where mg is a gravity force, $e_3 = [0 \ 0 \ 1]$ is a unit vector along z axis. The state condition $m = m_Q + m_L$ is the total mass. If $m = m_Q + \delta_m$ and $z_q < 1 + \delta_z$, the zero cable tension model system is used. Otherwise, the dynamic model is changing to the second non-zero cable tension model. Where δ_m is a small positive number which represent arbitrary mass change due to additional types of equipment and δ_z is a distortion cable length. The non-zero cable tension model is represented as follows

$$\dot{\chi}_Q = v_Q \quad (4)$$

$$m_Q \dot{v}_Q = F R e_3 - m_Q g e_3 - T \rho \quad (5)$$

$$\dot{\chi}_L = v_L \quad (6)$$

$$m_L \dot{v}_L = -m_L g e_3 + T \rho \quad (7)$$

$$I \dot{\Omega} = -\Omega \times I \Omega + M \quad (8)$$

where ρ is a unit vector from quadrotor center of gravity to the load which represents a cable direction vector in the body frame [3].

$$\rho = \begin{bmatrix} \sin(\theta_L)\cos(\phi_L) \\ \sin(\theta_L)\sin(\phi_L) \\ \cos(\theta_L) \end{bmatrix}$$

The cable tension T is equal to the magnitude of the cable force multiplied by the unit vector ρ [2].

$$T = |f| \begin{bmatrix} \sin(\theta_L)\cos(\phi_L) \\ \sin(\theta_L)\sin(\phi_L) \\ \cos(\theta_L) \end{bmatrix}$$

where the magnitude of the cable forces $|f|$ is

$$|f| = m_L \dot{v}_L \quad (9)$$

The quadrotor with load dynamical model (4)-(8) can be rewritten as

$$\begin{aligned} \ddot{X} = & (\sin \psi \sin \phi + \cos \psi \sin \theta \cos \phi) \frac{F}{m_Q} \\ & - (\sin \theta_L \cos \phi_L) \frac{T}{m_Q} \end{aligned} \quad (10)$$

$$\begin{aligned} \ddot{y} = & (-\cos \psi \sin \phi + \sin \psi \sin \theta \cos \phi) \frac{F}{m_Q} \\ & + (\sin \theta_L \sin \phi_L) \frac{T}{m_Q} \end{aligned} \quad (11)$$

$$\ddot{z} = -g + (\cos \theta \cos \phi) \frac{F}{m_Q} - (\cos \theta_L) \frac{T}{m_Q} \quad (12)$$

$$\ddot{\phi} = \frac{I_{yy} - I_{zz}}{I_{xx}} \dot{\theta} \dot{\psi} - \frac{I_r}{I_{xx}} \dot{\theta} \Omega + \frac{M_\phi}{I_{xx}} \quad (13)$$

$$\ddot{\theta} = \frac{I_{zz} - I_{xx}}{I_{yy}} \dot{\phi} \dot{\psi} - \frac{I_r}{I_{yy}} \dot{\phi} \Omega + \frac{M_\theta}{I_{yy}} \quad (14)$$

$$\ddot{\psi} = \frac{I_{xx} - I_{yy}}{I_{zz}} \dot{\phi} \dot{\theta} + \frac{M_\psi}{I_{zz}} \quad (15)$$

$$\ddot{\phi}_L = - (L \sin \theta_L \cos \phi_L) \frac{T}{m_L} + \frac{M_\phi}{m_L} \quad (16)$$

$$\ddot{\theta}_L = - (L \sin \theta_L \sin \phi_L) \frac{T}{m_L} + \frac{M_\theta}{m_L} \quad (17)$$

B. Linearised Model

Linearisation of the mathematical model around an operating point is the necessary requirement for LQR algorithm. For the take-off and landing vehicle, the operating point is the hovering point. The following assumptions are useful for a quadrotor at hovering point

$$\begin{aligned} F & \simeq m_Q g \\ \dot{\theta} & \simeq \dot{\phi} \simeq \dot{\psi} \simeq \dot{\theta}_L \simeq \dot{\phi}_L \simeq 0 \\ T & \simeq F - m_Q g \simeq 0 \\ \sin \psi & \simeq 0 \\ \sin \theta & \simeq \theta \\ \sin \phi & \simeq \phi \\ \sin \theta_L & \simeq \theta_L \\ \sin \phi_L & \simeq \phi_L \end{aligned}$$

According to these assumptions, equations (10)-(17) can be rewritten as

$$\ddot{x} = \theta g \quad (18)$$

$$\ddot{y} = -\phi g \quad (19)$$

$$\ddot{z} = -g + \frac{F}{m_Q} \quad (20)$$

$$\ddot{\phi} = \frac{M_\phi}{I_{xx}} \quad (21)$$

$$\ddot{\theta} = \frac{M_\theta}{I_{yy}} \quad (22)$$

$$\ddot{\psi} = \frac{M_\psi}{I_{zz}} \quad (23)$$

$$\ddot{\phi}_L = \frac{M_\phi}{m_L} \quad (24)$$

$$\ddot{\theta}_L = \frac{M_\theta}{m_L} \quad (25)$$

The model equations (18)-(25) can be represented in state space as follows

$$\dot{X} = AX + BU \quad (26)$$

$$Y = CX + DU \quad (27)$$

where $X = [x, y, z, \phi, \theta, \psi, \theta_L, \phi_L, \dot{x}, \dot{y}, \dot{z}, \dot{\phi}, \dot{\theta}, \dot{\psi}, \dot{\phi}_L, \dot{\theta}_L]^T$, $U = [F, M_\phi, M_\theta, M_\psi]^T$, $Y = [x, y, z, \psi]$,

$$A = \begin{bmatrix} \mathbf{0}_{8 \times 3} & \mathbf{0}_{8 \times 1} & \mathbf{0}_{8 \times 1} & \mathbf{0}_{8 \times 3} & \mathbf{J}_{8 \times 8} \\ \mathbf{0}_{1 \times 3} & 0 & g & \mathbf{0}_{1 \times 3} & \mathbf{0}_{1 \times 8} \\ \mathbf{0}_{1 \times 3} & -g & 0 & \mathbf{0}_{1 \times 3} & \mathbf{0}_{1 \times 8} \\ \mathbf{0}_{6 \times 3} & \mathbf{0}_{6 \times 1} & \mathbf{0}_{6 \times 1} & \mathbf{0}_{6 \times 3} & \mathbf{J}_{8 \times 8} \end{bmatrix} \quad (28)$$

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1/m_Q & 0 & 0 & 0 \\ \mathbf{0}_{8 \times 1} & \mathbf{0}_{8 \times 1} & \mathbf{0}_{8 \times 1} & \mathbf{0}_{8 \times 1} \\ 0 & 1/I_{xx} & 0 & 0 \\ 0 & 0 & 1/I_{yy} & 0 \\ 0 & 0 & 0 & 1/I_{zz} \\ 0 & 1/m_L & 0 & 0 \\ 0 & 0 & 1/m_L & 0 \end{bmatrix} \quad (29)$$

$$C = \begin{bmatrix} \mathbf{J}_{3 \times 3} & \mathbf{0}_{3 \times 2} & \mathbf{0}_{1 \times 1} & \mathbf{0}_{3 \times 10} \\ \mathbf{0}_{1 \times 3} & \mathbf{0}_{1 \times 2} & 1 & \mathbf{0}_{1 \times 10} \end{bmatrix} \quad (30)$$

and $D = [\mathbf{0}_{4 \times 16}]$.

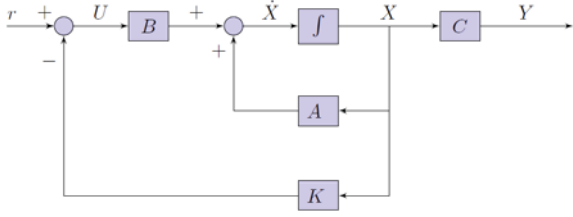


Fig. 2. LQR Control Block Diagram

III. LQR CONTROLLER

Linear Quadratic Regulator LQR is a linear state feedback optimal controller. It is presented based on dynamic model linearization of the quadrotor with cable-suspended load system in order to achieve minimum cost of the desired parameters. The error minimisation of the dynamic model is enforced by the convenient parameters of weight matrices [12] using cost objective function Γ of the form

$$\Gamma = \int_{t_o}^{t_f} \frac{1}{2} (X^T Q X + U^T R U) dt \quad (31)$$

where the initial and final time of the control horizon are t_o and t_f , matrices $Q \geq 0$ and $R > 0$ are the cost of the state X and control input U gain of the linear system represented in state space as follows

$$\dot{X} = AX + BU \quad (32)$$

$$Y = CX + DU \quad (33)$$

The goal is to minimize the cost function Γ via a calculated control input

$$U^* = -KX = -R^{-1}B^T P X \quad (34)$$

where P can be calculated from the continuous Riccati equation

$$\dot{P}(t) + P(t)A + A^T P(t) - P(t)BR^{-1}B^T P(t) + Q = 0 \quad (35)$$

Consequently, the state feedback optimal control gain K can be calculated using the following formula

$$K = lqr(A, B, Q, R) \quad (36)$$

The LQR controller is designed by choosing positive parameters for Q and R matrices to determined the desired thrust and orientations. This controller is presented to estimate state feedback tuning parameter, which is similar to individually tuning as in PD controller parameters[13]. The full system block diagram is shown in figure 2.

IV. SIMULATION RESULTS

An MATLAB simulator of a quadrotor with a cable-suspended load is implemented to test the stability of the proposed controller. Table I shows the quadrotor with load parameters used in this simulation. In order to achieve lifting and transporting of the quadrotor with load. The proposed

Symbol	Definition	Value	Units
I_x	Roll Inertia	4.4×10^{-3}	$kg.m^2$
I_y	Pitch Inertia	4.4×10^{-3}	$kg.m^2$
I_z	Yaw Inertia	8.8×10^{-3}	$kg.m^2$
m_Q	Mass	0.5	kg
m_L	Mass	0.2	kg
g	Gravity	9.81	m/s^2
l	Arm Length	0.17	m
L	Cable Length	1	m
I_r	Rotor Inertia	4.4×10^{-5}	$kg.m^2$

TABLE I
QUADROTOR PARAMETERS

LQR controller was tested and the results were compared with that of a PD controller. The first step was to find the controller of the four outputs $Y = [x, y, z, \psi]$ parameters. Then in the second step, by controlling the directions x and y , the desired remaining states $\phi_d, \theta_d, \phi_{Ld}, \theta_{Ld}$ can be found and controlling using the proposed controller as well. The matrices Q and R for $Y = [x, y, z, \psi]$ were chosen to be

$$Q = \text{diag}([0.039, 5, 0.039, 5, 10, 50, 1.44, 0.00001])$$

$$R = \text{diag}([10, 10, 1, 1])$$

and those for $\phi, \theta, \phi_L, \theta_L$ were

$$Q = \text{diag}([0.65, 0.035, 0.65, 0.035, 1, 1, 1, 1])$$

$$R = \text{diag}([1, 1, 100, 100])$$

Applying these cost matrices to the formula (36), the following state feedback controller parameters is obtained for the quadrotor translation $[x, y, z]$ as

$$K = \begin{bmatrix} 0.0624 & 0.716 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0624 & 0.716 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.162 & 7.29 \end{bmatrix}$$

and for the quadrotor rotation $[\phi, \theta, \psi]$ is

$$K = \begin{bmatrix} 0.806 & 0.205 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.806 & 0.205 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.2 & 0.145 \end{bmatrix}$$

and for the load rotation $[\phi_L, \theta_L]$ is

$$K = \begin{bmatrix} 0.1 & 0.1043 & 0 & 0 \\ 0 & 0 & 0.1 & 0.1043 \end{bmatrix}$$

The first simulation test of LQR controller is to lift the load and hover over $z_d = 2m$ desired quadrotor height, which means the desired load height is $z_{Ld} = 1m$, the desired quadrotor direction is $x_d = y_d = 0$, the desired quadrotor rotation is $\phi_d = \theta_d = \psi_d = 0$ and the desired load rotation is $\phi_{Ld} = \theta_{Ld} = 0$. The second test strategy is to lift and transport the load to a desired point $z_d = 2m, x_d = y_d = 1m$ with desired

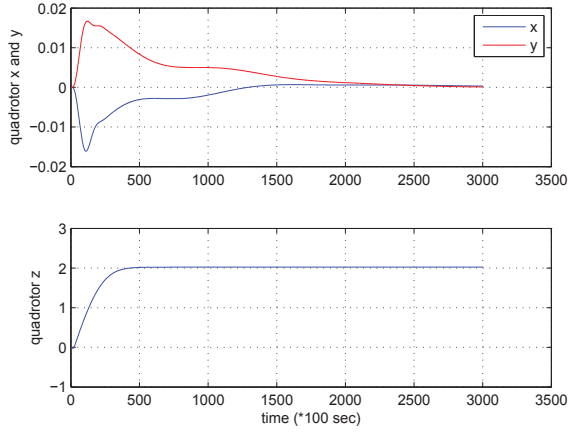


Fig. 3. Quadrotor hovering positions, LQR Controller

quadrotor rotation $\phi_d = \theta_d = \psi_d = 0$ and desired load rotation $\phi_{Ld} = \theta_{Ld} = 0$.

The lifting and hovering performance is illustrated in figures 3-5. Figure 3 shows the quadrotor positions performance, while figures 4 and 5 illustrates the quadrotor attitude angles and load angles performance respectively. These results are describe a stable performance with small steady state error. The vehicle and load angles were stabilized in less than $3sec$ and reached a zero steady state error after $15sec$.

Simulation of PD controller was tested in lifting and transporting tasks for the comparison purpose. Figure 6 shows the quadrotor positions performance using LQR controller in transporting task compared with that of PD controller. The quadrotor attitude angles and load angles performance in transporting task compared with that of PD were illustrated in Figures 7 and 8. From these figures, it is obvious that the performance of LQR controller was faster than that of PD controller with smaller steady state error as well. The results of the PD controller can't reach zero steady state error. In general, LQR controller performs better than PD controller in terms of time-consuming and steady state error.

V. CONCLUSIONS

This paper presents an optimal LQR controller to stabilize the quadrotor and a cable-suspended load in lifting and transporting tasks. The dynamic model of the quadrotor was obtained considering the effect of the suspended load to be eight degree of freedom system. The dynamic model was linearised around the hovering point in order to satisfy the proposal control approach requirements. The controller was applied to the nonlinear dynamic system and the results compared with that of PD controller. The simulation results show that LQR controller performs are efficiently minimizing the steady state error and time-consuming to reach the stability conditions. Our future work is to apply this controller practically in the real vehicle to verify the controller validity.

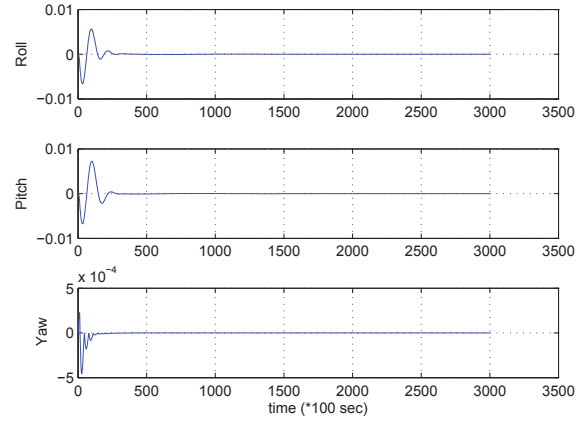


Fig. 4. Quadrotor hovering angles, LQR Controller

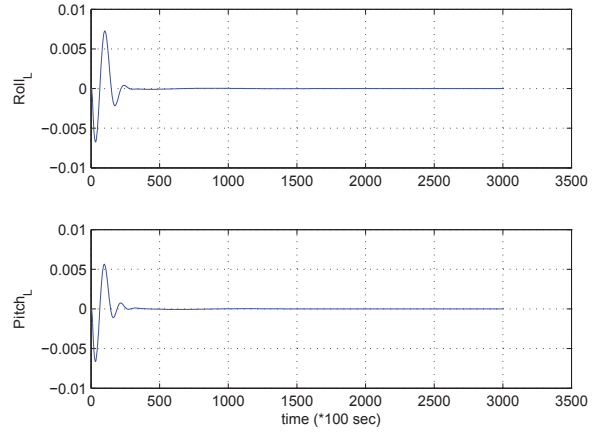


Fig. 5. Load hovering angles, LQR Controller

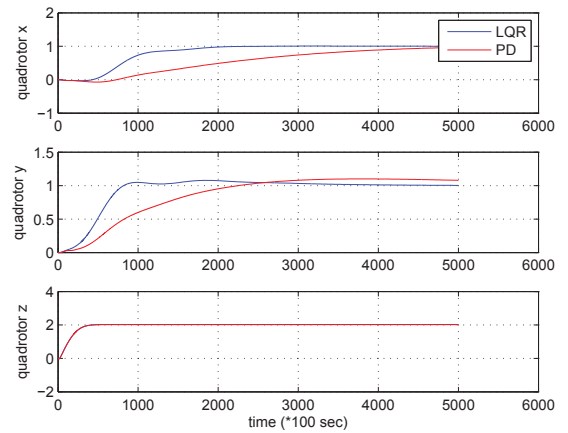


Fig. 6. Quadrotor transporting positions, LQR and PD Controllers

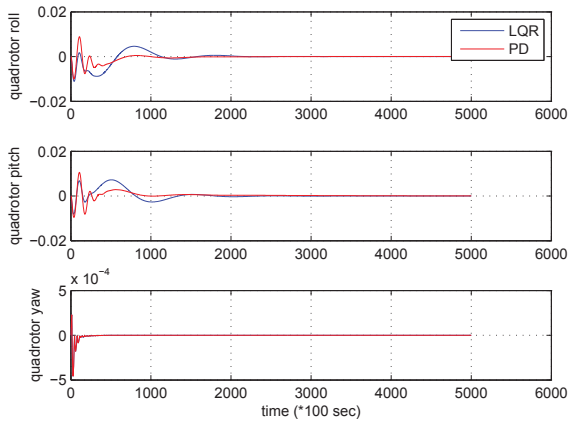


Fig. 7. Quadrotor transporting angles, LQR and PD Controllers

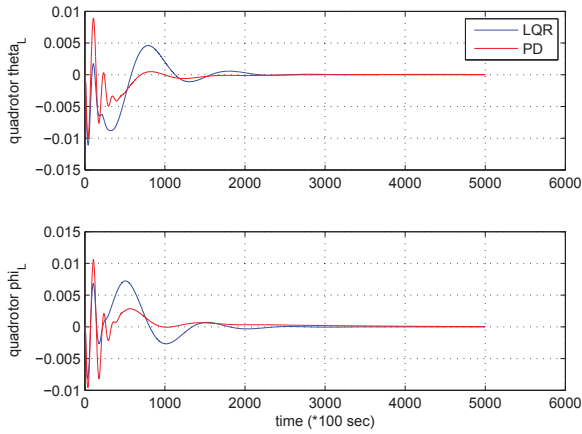


Fig. 8. Load transporting angles, LQR and PD Controllers

REFERENCES

- [1] A. Faust, I. Palunko, P. Cruz, R. Fierro, and L. Tapia, "Learning swing-free trajectories for UAVs with a suspended load," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pp. 4902–4909, IEEE, 2013.
- [2] S. Sadr, S. A. A. Moosavian, and P. Zarafshan, "Dynamics modeling and control of a quadrotor with swing load," *Journal of Robotics*, vol. 2014, 2014.
- [3] P. Cruz and R. Fierro, "Autonomous lift of a cable-suspended load by an unmanned aerial robot," in *Control Applications (CCA), 2014 IEEE Conference on*, pp. 802–807, IEEE, 2014.
- [4] C. Raimúndez and J. L. Camaño, "Transporting hanging loads using a scale quad-rotor," in *CONTROLO2014—Proceedings of the 11th Portuguese Conference on Automatic Control*, pp. 471–482, Springer, 2015.
- [5] S. Dai, T. Lee, and D. S. Bernstein, "Adaptive control of a quadrotor UAV transporting a cable-suspended load with unknown mass," *ratio*, vol. 1, p. 3, 1991.
- [6] I. Palunko, P. Cruz, and R. Fierro, "Agile load transportation: Safe and efficient load manipulation with aerial robots," *Robotics & Automation Magazine, IEEE*, vol. 19, no. 3, pp. 69–79, 2012.
- [7] I. Palunko, R. Fierro, and P. Cruz, "Trajectory generation for swing-free maneuvers of a quadrotor with suspended payload: A dynamic programming approach," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pp. 2691–2697, IEEE, 2012.
- [8] K. Sreenath, T. Lee, and V. Kumar, "Geometric control and differential flatness of a quadrotor UAV with a cable-suspended load," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pp. 2269–2274, IEEE, 2013.
- [9] T. Lee, K. Sreenath, and V. Kumar, "Geometric control of cooperating multiple quadrotor UAVs with a suspended payload," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pp. 5510–5515, IEEE, 2013.
- [10] K. Sreenath and V. Kumar, "Dynamics, control and planning for cooperative manipulation of payloads suspended by cables from multiple quadrotor robots," vol. 1, June 2013.
- [11] G. Wu and K. Sreenath, "Geometric control of multiple quadrotors transporting a rigid-body load," 2014.
- [12] A. Sorensen, "Autonomous control of a miniature quadrotor following fast trajectories," *Control Engineering Masters Thesis, Aalborg University, Denmark*, 2010.
- [13] S. Bouabdallah, A. Noth, and R. Siegwart, "PID vs LQ control techniques applied to an indoor micro quadrotor," in *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3, pp. 2451–2456, IEEE, 2004.

Output Feedback Sliding Mode Controller with \mathcal{H}_2 Performance for Robot Manipulator

P. Righettini and R. Strada, S. Valilou, E. KhademOlama

Department of Engineering and Applied science

Università degli Studi di Bergamo, Bergamo, Italy

Email: paolo.righettini@unibg.it, roberto.strada@unibg.it,

s.valilou@studenti.unibg.it, e.khademolama@studenti.unibg.it

Abstract—In this paper, an output feedback sliding mode controller with \mathcal{H}_2 performance for position control of uncertain robot manipulators without joint velocity measurement based on a second-order sliding mode observer is presented. First, a second order sliding mode observer with finite time convergence is developed for velocity estimation. Then the robot manipulator is formulated as an error dynamics and a sliding mode control with \mathcal{H}_2 performance is developed to tackle the unmodeled dynamics and disturbances. The stability properties of the controllers are proved by the Lyapunov method. Asymptotic stability of the closed loop system is proved under a set of nonrestrictive assumptions on the plant and the disturbances. This control law providing smaller errors and better performance to deal with uncertainty and disturbances. Finally, we will use a 2-dof planar robot manipulator to verify the effectiveness of the proposed control scheme.

Keywords—Robot manipulator, Dynamic control, Output feedback tracking control, Sliding mode control, \mathcal{H}_2 performance.

I. INTRODUCTION

In recent years, robot manipulators showed a vast usage in industries. Due to their wide range of practical applications they have attracted the attention of many researchers. However design a high performance controller for rigid robot manipulators is a difficult problem due to its non-linearities, parameter uncertainties, external disturbances and the coupling effects that are typical of robotic systems. Various control techniques have been developed during recent decade for robot control, for instances, traditional feedback control (PID, PD) ([1]-[3]), sliding mode control ([4]-[11]), adaptive control ([12]-[15]) and robust control ([16]-[18]). To enhance the abilities of trajectory tracking of manipulators, robust control system confronted the uncertainties and disturbances is necessary.

Sliding mode control is a powerful robust technique to control uncertain nonlinear systems ([19] and [20]). However In the face of large-scale parametric uncertainties and disturbances the sliding mode control approach demands high gains for the controller to achieve satisfactory tracking performance. The main practical problem of having high-gain-based design is that it amplifies the input and output disturbance as well as excites hidden unmodeled dynamics, causing poor tracking performance. For solving this problem, designing observer for disturbance estimation is suggested in some papers due to the unmodelled friction effects and sensor noise ([6], [9] and [10]). However, parameter identification of friction model is quite difficult, and precise friction model is impossible to obtain in practical applications. So the difference between the identified

model of the robotic manipulator and the real plant can make the performances of the controlled system quite poor.

In this paper, we design an output feedback sliding mode controller with \mathcal{H}_2 performance based on the dynamic of manipulators. First, an observer is introduced for velocity estimation of manipulators based on second order sliding mode (SOSM) observer (see [25] and [26]). The using of the SOSM observer is suggested in this paper because linear observers do not achieve adequate performance for nonlinear systems such as manipulators and model based observers needed the knowing the exact model of system, while for the SOSM observer it is not needed the exact knowledge of system model and also the observer can be designed separately from the controller. Secondly, a sliding mode controller with \mathcal{H}_2 performance is proposed to track the desired trajectory of the manipulator, although the dynamic model of manipulator is even with uncertainties and disturbances. Reasonably, the \mathcal{H}_2 norm in controller design can improve the performance of closed-loop system. The proposed control law consists of a linear state feedback part and a nonlinear part with an additional variable structure control component. By defining performance measure z , we can limit the energy of states and control inputs. Finally, we propose an linear matrix inequality (LMI)-based solution to the sliding mode controller design with the \mathcal{H}_2 performance.

This paper is organized as following: In Section 2, some notations and preliminaries are introduced. In section 3, the mathematical representation of the manipulator dynamic and problem is introduced. In Section 4 a second order sliding mode observer for velocity estimation is presented. In Section 5, the robust output feedback dynamic controller is discussed. Computer simulation results of the proposed dynamic control are given in Section 6. Finally, conclusion is presented in Section 7.

II. NOTATIONS AND PRELIMINARIES

The notations used in this paper are fairly standard. For a given matrix A , A^T denotes its transpose. I denotes unity matrix with appropriate dimension. We define

$$\|z\|_\infty = \sup_{t \geq 0} \sqrt{z^T(t)z(t)}, \quad \|w\|_2 = \sqrt{\int_0^\infty w^T(t)w(t) dt},$$

$$\forall w \in L_2[0, \infty].$$

Definition 2.1: [21] The $\mathcal{L}_2 - \mathcal{L}_\infty$ induced norm (or energy to peak norm) of the system S is defined as

$$\|S\|_{\mathcal{H}_2} = \sup_{0 < \|w\|_2 < \infty} \frac{\|z\|_\infty}{\|w\|_2}. \quad (1)$$

where z is the output and w is the input of the system S . The equation (1) is called as a \mathcal{H}_2 norm.

Therefore, the \mathcal{H}_2 norm measures the peak amplitude of the output signal for the worst case input.

We use an asterisk (*) to represent a term that is induced by symmetry. The following results are used in the paper.

Lemma 2.1: [22] For any $x, y \in \mathbb{R}^n$ and any positive definite matrix $P \in \mathbb{R}^{n \times n}$, we have

$$2x^T y \leq x^T P x + y^T P^{-1} y$$

Lemma 2.2: [23] The following LMI

$$\begin{pmatrix} Q(x) & S(x) \\ S(x)^T & R(x) \end{pmatrix} > 0$$

where $S(x)$ depends affinely on x , is equivalent to

$$R(x) > 0, \quad Q(x) - S(x)R(x)^{-1}S(x) > 0$$

III. SYSTEM DESCRIPTION AND PROBLEM FORMATION

In this section, we describe the dynamic model of manipulator. The dynamic equation of general n -link robotic manipulators in the joint space, by using the Lagrangian approach can be written as [24]:

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + g(q) + T_d(t, q, \dot{q}) = \tau \quad (2)$$

where $q \in \mathbb{R}^n$ is the joint position vector, $\dot{q} \in \mathbb{R}^n$ is the joint velocity vector, $\tau \in \mathbb{R}^n$ is the vector of motor control torques, $M(q) \in \mathbb{R}^{n \times n}$ is the symmetric and uniformly positive definite inertia matrix, $C(q, \dot{q})\dot{q} \in \mathbb{R}^n$ is the coriolis and centrifugal torques vector, $G(q) \in \mathbb{R}^n$ is the gravitational torques vector and $T_d(t, q, \dot{q}) \in \mathbb{R}^n$ is the vector of generalized input due to disturbances or unmodeled dynamics.

In the dynamic model of manipulator we can write

$$\begin{aligned} M(q) &= M_n(q) + \Delta M(q) \\ C(q, \dot{q}) &= C_n(q, \dot{q}) + \Delta C(q, \dot{q}) \\ g(q) &= g_n(q) + \Delta g(q) \end{aligned} \quad (3)$$

where $M_n(q)$, $C_n(q, \dot{q})$ and $g_n(q)$ are the known nominal function and $\Delta M(q)$, $\Delta C(q, \dot{q})$ and $\Delta g(q)$ are the uncertain part of the matrix.

In the next section we design an observer for velocity estimation.

IV. SECOND ORDER SLIDING MODE OBSERVER

In this section, we use the SOSM observer in order to reconstruct the velocity from the position measurements. For this purpose and simplify the subsequent design and analysis, we introduce the variables $x_1 = q$ and $x_2 = \dot{q}$, the equation (2) can be rewritten in the space form

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= f(t, x_1, x_2, \tau) + \Delta(t, x_1, x_2) \\ y &= x_1 \end{aligned} \quad (4)$$

where in this equation $f(t, x_1, x_2, \tau)$ is the nominal part and $\Delta(t, x_1, x_2)$ represents the uncertainties and frictions in the dynamic model. So

$$f(t, x_1, x_2, \tau) = -M_n^{-1}(x_1)(C_n(x_1, x_2)x_2 + g_n(x_1) - \tau)$$

Now based on equation (4) the SOSM observer is design as [25]

$$\begin{aligned} \dot{\hat{x}}_1 &= \hat{x}_2 + k_1 \|x_1 - \hat{x}_1\|^{\frac{1}{2}} \text{sign}(x_1 - \hat{x}_1) \\ \dot{\hat{x}}_2 &= f(t, x_1, \hat{x}_2, \tau) + k_2 \text{sign}(x_1 - \hat{x}_1) \end{aligned} \quad (5)$$

where \hat{x}_1 and \hat{x}_2 are the state estimations, and k_i s are the sliding mode gains. Taking $\tilde{x}_1 = x_1 - \hat{x}_1$ and $\tilde{x}_2 = x_2 - \hat{x}_2$, we obtain the estimation errors

$$\begin{aligned} \dot{\tilde{x}}_1 &= \tilde{x}_2 - k_1 \|\tilde{x}_1\|^{\frac{1}{2}} \text{sign}(\tilde{x}_1) \\ \dot{\tilde{x}}_2 &= F(t, x_1, x_2, \hat{x}_2) - k_2 \text{sign}(\tilde{x}_1) \end{aligned} \quad (6)$$

where

$$F(t, x_1, x_2, \hat{x}_2) = f(t, x_1, x_2, \tau) - f(t, x_1, \hat{x}_2, \tau) + \Delta(t, x_1, x_2)$$

Now suppose that the system states can be assumed bounded, then the existence is ensured of a constant η^+ such that the inequality

$$|F(t, x_1, x_2, \hat{x}_2)| < \eta^+$$

holds for any possible t , x_1 , x_2 and $|\hat{x}_2| \leq 2\text{sup}|x_2|$. Therefore, the sliding mode gains of the observer scheme in equation (5) are chosen as

$$\begin{aligned} k_1 &> \eta^+ \\ k_2 &> \sqrt{\frac{2}{k_1 - \eta^+}} \frac{k_1 + \eta^+ (1+p)}{(1-p)} \end{aligned}$$

where $0 < \rho < 1$ is an arbitrary constant. Then the observer scheme is stable, and the states of the observer in equation (5), (\hat{x}_1, \hat{x}_2) converge to the true states (x_1, x_2) of the system in the equation (4) in finite time. The stability of the observer has been proved in [25].

In the next section, the controller of the system will be discussed.

V. CONTROLLER DESIGN

In this section a dynamic controller based on sliding mode control with \mathcal{H}_2 performance for the dynamic model of manipulator in the equation (2) is introduced. In the equation (3), the nominal part of the inertia matrix $M_n(q)$ is composed of two kinds of elements: first Inertia terms that do not depend on the robot's configuration and is constant. And the second terms that depend on the robot's configuration. Therefore

$$M_n(q) = \bar{M}_c + \bar{M}_{nc}(q)$$

where \bar{M}_c is a matrix with constant elements and $\bar{M}_{nc}(q)$ is the nonconstant part of $M_n(q)$. From the dynamic model we can write

$$\bar{M}_c \ddot{q} + \phi(q, \dot{q}, \ddot{q}) + T_d(t, q, \dot{q}) = \tau \quad (7)$$

where

$$\phi(q, \dot{q}, \ddot{q}) = (\bar{M}_{nc}(q) + \Delta M(q))\ddot{q} + C(q, \dot{q})\dot{q} + g(q)$$

so the dynamic equation (7) can be describe as follows:

$$\ddot{q} = \bar{M}_c^{-1}(\tau - \phi(q, \dot{q}, \ddot{q}) - T_d(t, q, \dot{q})) \quad (8)$$

Let us define the error vector such that $e_1 = q_d - q$ and $e_2 = \dot{q}_d - \dot{q}$ where q_d and \dot{q}_d is the desired trajectory and its first derivative respectively. So, the system model (8) can be defined in the error space form as

$$\dot{e}_2 = \ddot{q}_d - \bar{M}_c^{-1}(\tau - \phi(q, \dot{q}, \ddot{q}) - T_d(t, q, \dot{q}))$$

Therefore the dynamic model of manipulator can be defined in the error-state space form as follows

$$\begin{bmatrix} \dot{e}_1 \\ \dot{e}_2 \end{bmatrix} = \begin{bmatrix} e_2 \\ -\bar{M}_c^{-1}e_2 - \bar{M}_c^{-1}\tau + \bar{M}_c^{-1}(\phi(q, \dot{q}, \ddot{q}) + e_2) + \ddot{q}_d + \bar{M}_c^{-1}T_d(t, q, \dot{q}) \end{bmatrix} \quad (9)$$

After some simple manipulation,, the equation (9) can be integrated into the following state space

$$\dot{e} = Ae + B(\tau + h(t, e, q_d, \dot{q}_d, \tau)) + Ew \quad (10)$$

where

$$\begin{aligned} e &= \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}, \quad \dot{e} = \begin{bmatrix} \dot{e}_1 \\ \dot{e}_2 \end{bmatrix} \\ A &= \begin{bmatrix} 0_{n \times n} & I_{n \times n} \\ 0_{n \times n} & -\bar{M}_c^{-1} \end{bmatrix}, \quad B = \begin{bmatrix} 0_{n \times n} \\ -\bar{M}_c^{-1} \end{bmatrix} \\ h(t, e, q_d, \dot{q}_d, \tau) &= \begin{bmatrix} 0_{n \times 1} \\ -\phi(q, \dot{q}, \ddot{q}) - e_2 - \bar{M}_c \ddot{q}_d \end{bmatrix} \\ E &= \begin{bmatrix} 0_{n \times n} \\ \bar{M}_c^{-1} \end{bmatrix}, \quad w = T_d(t, q, \dot{q}) \end{aligned}$$

The matrixes A , B and E are known and $h(t, e, q_d, \dot{q}_d, \tau)$ is the nonlinearities and parametric uncertainties that caused by inaccurate measurement of the robot parameters such as mass and moment of inertia. And w is nonparametric uncertainties that may be caused by unmoulded dynamics and frictions. We also assume that the $h(t, e, q_d, \dot{q}_d, \tau)$ is bounded as

$$\|h(t, e, q_d, \dot{q}_d, \tau)\| \leq \beta\|\tau\| + \phi(t, e),$$

where $\phi(t, e) \geq 0$ is a known function, $\beta < 1$ is a known constant and $\|\bullet\|$ denotes the 2-norm. For the system (10), we consider the following controller

$$\tau = \tau_l + \tau_n \quad (11)$$

where the linear component τ_l is defined by

$$\tau_l = Ge,$$

and the nonlinear discontinues component τ_n is given by

$$\tau_n = \begin{cases} -\rho(e) \frac{\sigma}{\|\sigma\|}, & \text{if } \sigma \neq 0 \\ 0, & \text{if } \sigma = 0 \end{cases}$$

where

$$\rho(e) = \frac{1}{1 - \beta}[\alpha + \beta\|Ge\| + \phi(t, e)],$$

and $\alpha > 0$ is a positive scalar. The sliding surface σ is given by $\sigma = Fe$ where F is a constant matrix that will be designed latter. This nonlinear component is used to reject the matching uncertainty $h(t, e, q_d, \dot{q}_d, \tau)$ [27]. We consider a performance measure z and then we can define a generalized system as follows

$$\Sigma : \begin{cases} \dot{e} = Ae + B(\tau_l + \tau_n + h(t, e, q_d, \dot{q}_d, \tau)) + Ew \\ z = L_1 e + L_2 \tau_l \end{cases} \quad (12)$$

where L_1 and L_2 are design matrices with appropriate dimensions.

Theorem 5.1: Consider the generalized plant Σ in (12) and the control law in (11). Given $\gamma_2 > 0$, there exists the feedback gain G and the surface matrix F which stabilizes Σ and guarantees a \mathcal{H}_2 performance, that is

$$\|z\|_\infty < \gamma_2 \|w\|_2$$

if there exists a matrix $P = P^T > 0$ and a matrix G such that

$$PA + A^T P + PEE^T P + PBG + G^T B^T P < 0, \quad (13)$$

$$(L_1 + L_2 G)^T (L_1 + L_2 G) < \gamma_2^2 P,$$

Furthermore, if a feasible solution (P, G) exists in the above matrix inequalities, then

$$F = B^T P.$$

If $w = 0$ for all t , the reachability condition is satisfied and for the case $w \neq 0$ a stable sliding motion may not occur in finite time but the closed-loop system is uniform stability as long as $w \in \mathcal{L}_\infty$.

Proof: By Theorem 5.1, we have only to show $\|z\|_\infty < \gamma_2 \|w\|_2$. Assume that (13) holds. Then

$$PA + A^T P + PEE^T P + PBG + G^T B^T P < 0 \quad (14)$$

The inequality (13) implies

$$e^T (L_1 + L_2 G)^T (L_1 + L_2 G) e < e^T \gamma_2^2 P e$$

This inequality and (12) imply

$$e^T P e > \frac{1}{\gamma_2^2} e^T (L_1 + L_2 G) (L_1 + L_2 G)^T e = \frac{1}{\gamma_2^2} z^T z \quad (15)$$

To show that a stable sliding mode exists, we define the Lyapunov function as $V = e^T P e$, where P satisfies (13). By substituting the control input (11) in to the time derivative of V , we can obtain

$$\begin{aligned} \dot{V} &= 2e^T P(Ae + B[\tau + h(t, e, q_d, \dot{q}_d, \tau)] + Ew) \\ &= e^T (PA + A^T P)e + 2e^T PBGe \\ &\quad - 2e^T PB\rho \frac{\sigma}{\|\sigma\|} + 2e^T PBh(t, e, q_d, \dot{q}_d, \tau) \\ &\quad + 2e^T PEw \end{aligned}$$

where $\sigma = Fe$ and $F = B^T P$. Using Lemma 2.1, we can show

$$2e^T PEw \leq e^T PEE^T P e + w^T w$$

we define new variables ν_1 and ν_2 as

$$\begin{aligned} \nu_1 &= e^T (PA + A^T P + PEE^T P + PBG + G^T B^T P)e \\ \nu_2 &= 2\|\sigma\|(\beta\|Ge\| - \rho + \beta\rho + \phi(t, e)) + w^T w \end{aligned}$$

Thus $\dot{V} < \nu_1 + \nu_2$, by using inequality (14) we can write

$$\begin{aligned} \nu_1 &< 0 \\ \nu_2 &< -2\alpha\|\sigma\| + w^T w < w^T w \end{aligned}$$

Thus, we have

$$\dot{V} < w^T w \quad (16)$$

Using the inequalities (16) and (15), we can show that

$$\int_0^t w^T w dt - \frac{1}{\gamma_2^2} z^T z \geq e^T P e - \frac{1}{\gamma_2^2} z^T z > 0$$

which implies $\|z\|_\infty < \gamma_2 \|w\|_2$.

If $w = 0$ for all t in the inequality (16), we conclude that the closed-loop system (12) is globally quadratically stable and it satisfies for some $\alpha_1 > 0$ and $\alpha_2 > 0$ [27]

$$\|e\| \leq \alpha_1 \exp(-\alpha_2 t)$$

Motivated by developments in [27], we define $\sigma_0 = (B^T P B)^{-\frac{1}{2}} \sigma$. After some algebra, we can obtain

$$\sigma_0^T \dot{\sigma}_0 < \|\sigma\|(\delta \|e\| - \alpha) \quad (17)$$

where $\delta = \|(B^T P B)^{-1} B^T P (A + B G)\|$. We can obtain

$$\sigma_0^T \dot{\sigma}_0 < -\|\sigma\|(\alpha - \delta e) \leq -\alpha_3 \|\sigma_0\|(\alpha - \delta \|e\|) \quad (18)$$

where $\alpha_3 > 0$. Using the inequalities (17) and (18), it is easy to show that for any $0 < \varepsilon < \alpha$

$$\begin{aligned} \alpha_1 \exp(-\alpha_2 t) < q &\rightarrow t \geq \frac{1}{\alpha_2} \ln\left(\frac{\alpha_1}{q}\right) = t_1 \\ \|e\| < q = \frac{\alpha - \varepsilon}{\delta}, \quad \forall t \geq t_1 \end{aligned} \quad (19)$$

The inequality (18) and (19) imply that for all $t \geq t_1$, $\sigma_0^T \dot{\sigma}_0 \leq -\varepsilon \alpha_3 \|\sigma_0\|$ and reachability condition for $w = 0$ is satisfied. For the case $w \neq 0$ a stable sliding motion may not occur in finite time but the closed-loop system is uniform stability as long as $w \in \mathcal{L}_\infty$, see [27]. The following result proposes an LMI-based solution to Theorem 5.1. ■

Corollary 5.1: Given $\gamma_2 > 0$, the matrix inequalities in (13) are feasible if there exists a matrix $X = X^T > 0$ and a matrix W such that

$$\begin{aligned} \begin{pmatrix} AX + XA^T + BW + W^T B & E^T \\ * & -I \end{pmatrix} &< 0, \\ \begin{pmatrix} \gamma_2^2 X & XL_1^T + W^T L_2^T \\ * & I \end{pmatrix} &> 0 \end{aligned} \quad (20)$$

Furthermore, if a feasible solution (W, X) exists in the above LMIs, then

$$F = B^T X^{-1}, \quad G = W X^{-1}.$$

Proof: Define $X = P^{-1}$, $W = G X$. Using Lemma 2.2 and the inequality (13), we can easily obtain the LMIs in (20). ■

VI. NUMERICAL EXAMPLE

To demonstrate the effectiveness of this approach we consider the two-link manipulators adopted from the given by [28]. The configuration of the two link robot manipulator is shown in Fig. 1. The dynamic equation for this robot system can be defined as

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \ddot{q}_1 \\ \ddot{q}_2 \end{bmatrix} + \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \end{bmatrix} + \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} + \begin{bmatrix} T_{d1} \\ T_{d2} \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix} \quad (21)$$

where

$$\begin{aligned} M_{11} &= d_1 + 2d_3 \cos q_2 + 2d_4 \sin q_2, \\ M_{12} &= M_{21} = d_2 + d_3 \cos q_2, \quad M_{22} = d_2, \\ C_{11} &= -d_7 \dot{q}_2, \quad C_{12} = -d_7(\dot{q}_1 + \dot{q}_2), \\ C_{21} &= h \dot{q}_1, \quad C_{22} = 0, \\ G_1 &= d_5 \cos q_1 + d_6 \cos(q_1 + q_2), \\ G_2 &= d_6 \cos(q_1 + q_2), \end{aligned} \quad (22)$$

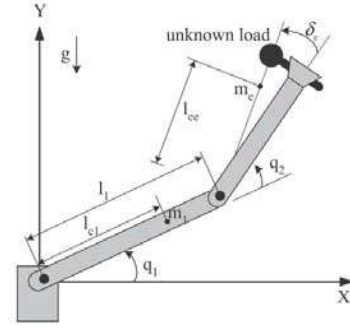


Fig. 1. Two link planar manipulator.

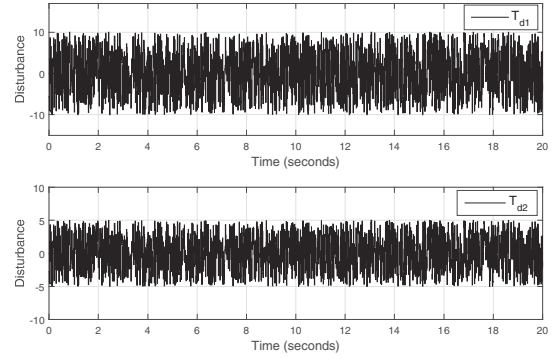


Fig. 2. The disturbance profiles which are random signals with mean 0.

and

$$\begin{aligned} d_1 &= I_1 + m_1 l_{c1}^2 + I_e + m_e l_{ce}^2 + m_e l_1^2 \\ d_2 &= I_e + m_e l_{ce}^2, \quad d_3 = m_e l_1 l_{ce} \cos(\delta_e) \\ d_4 &= m_e l_1 l_{ce} \sin(\delta_e), \quad d_5 = m_1 g l_{c1} + m_e g l_1 \\ d_6 &= m_e g l_{ce}, \quad d_7 = d_3 \sin q_2 - d_4 \cos q_2 \end{aligned} \quad (23)$$

where $q = [q_1 \ q_2]^T$ and $\dot{q} = [\dot{q}_1 \ \dot{q}_2]$ are the joints displacement and velocities, respectively. m_1 and m_e are links masses, l_1 is the length of the first link, l_{c1} and l_{ce} represent the lengths of the center of masses and I_1 and I_e denotes the moments of inertia of links. The nominal parameters of the two-link manipulators are chosen as follows:

$$\begin{aligned} m_1 &= 5 \text{ kg}, \quad m_e = 2.5 \text{ kg}, \quad \delta_e = 0^\circ \\ l_1 &= 1.0 \text{ m}, \quad l_{c1} = 0.5 \text{ m}, \quad l_{ce} = 0.5 \text{ m} \\ I_1 &= 0.36 \text{ Kg m}^2, \quad I_e = 0.24 \text{ Kg m}^2. \end{aligned} \quad (24)$$

The disturbances T_{d1} and T_{d2} are random signals with mean 0. These disturbances have been depicted in Fig. 2. The uncertain mass of joints 1 and 2 is illustrated in Fig. 3, it can be seen that the load of the manipulators is changed at $t = 5, 8, 12, 16$ s, respectively. The simulation sets are divided into two parts. In the first term, we verify the capability of SOSM observer in terms of velocity estimation. In the second term, the tracking performance of the proposed sliding mode controller with \mathcal{H}_2 Performance is shown.

For the first term of the simulations, we take the SOSM observer as equation (5) with assuming $k_1 = 5$ and $k_2 = 30$.

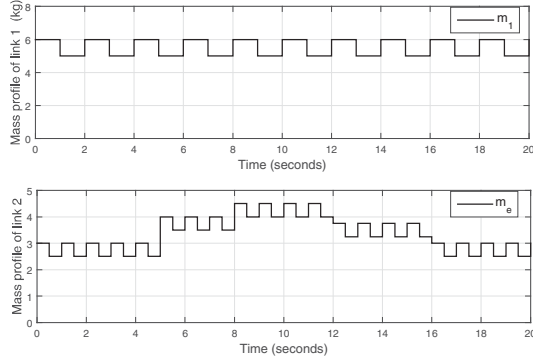


Fig. 3. The uncertain mass of joints.

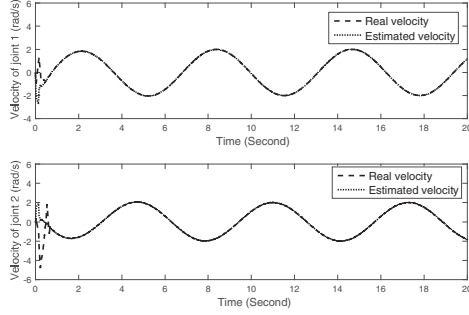


Fig. 4. Results of SOSM observer (25) for velocity estimation of joints. dashed line: Real velocity and dotted line: estimated joint velocity.

Thus, the proposed velocity observer has the form:

$$\begin{aligned}\dot{\hat{x}}_1 &= \hat{x}_2 + 5\|x_1 - \hat{x}_1\|^{\frac{1}{2}} \text{sign}(x_1 - \hat{x}_1) \\ \dot{\hat{x}}_2 &= M_n^{-1}(x_1)(\tau - C_n(x_1, x_2)x_2 - G_n(x_1)) + 30\text{sign}(x_1 - \hat{x}_1)\end{aligned}\quad (25)$$

where $x_1 = [q_1 \ q_2]$ and $x_2 = [\dot{q}_1 \ \dot{q}_2]$. Fig. 4 shows the performance of the SOSM observer to estimate the velocities. It can be seen that the observer provides a good estimate of the joint velocities of the robot manipulator.

In the second term of the simulation, we design a controller based on equation (11). For this purpose we assume the performance measure z in equation (12) as

$$z = \begin{bmatrix} 0.1 & 0 & 0.1 & 0.1 \\ 0.1 & 0 & 0.1 & 0 \\ 0.2 & 0 & 0.1 & 0.1 \\ 0 & 0.1 & 0 & 0.1 \end{bmatrix} e + \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.1 \\ 0 & 0.1 \\ 0.1 & 0.1 \end{bmatrix} \tau \quad (26)$$

Here, we use the results of Theorem 5.1 and Corollary 5.1 to design a sliding mode control with a \mathcal{H}_2 performance for manipulator dynamic errors in (12). By solving the LMI in equation (20), for $\gamma_2 = 0.06$, we obtain the sliding surface matrix F and the state feedback G as follows

$$\begin{aligned}F &= \begin{bmatrix} -0.4492 & -0.0555 & -1.1734 & -0.2205 \\ -0.2154 & -0.1778 & -0.5082 & -0.4135 \end{bmatrix} \\ G &= \begin{bmatrix} 0.9062 & -0.2611 & 2.1970 & 0.1644 \\ -0.4287 & 0.5092 & 0.4509 & 0.0420 \end{bmatrix}\end{aligned}\quad (27)$$

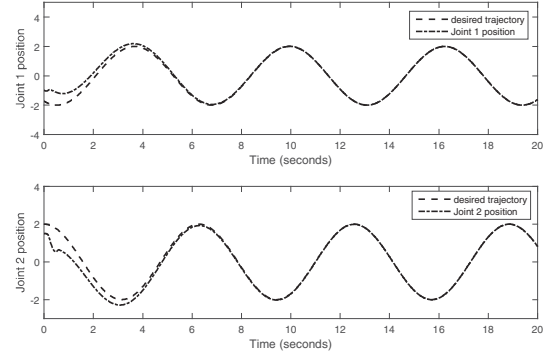


Fig. 5. simulation results of trajectory tracking of joints with control law (28). dashed line: desired trajectory and dash-dot line: actual trajectory

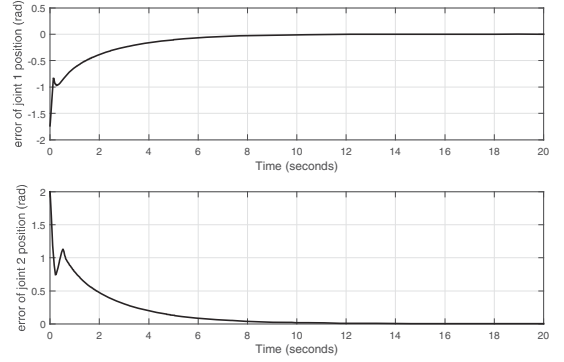


Fig. 6. Tracking errors of joints.

For simulation, the manipulator starts from $q = [-1 \ 1.5]$ and the desired joint trajectories for tracking are

$$\begin{bmatrix} q_{d1} \\ q_{d2} \end{bmatrix} = \begin{bmatrix} -2\sin(\frac{1}{3}\pi t) \\ 2\sin(\frac{1}{2}\pi t) \end{bmatrix} \text{rad}.$$

Since $\|h(t, e, q_d, \dot{q}_d)\| < 30 + \|C_n(\hat{x}_1, \hat{x}_2)x_2 + G_n(\hat{x}_1)\|$ we can set $\phi(e, t) = 30 + \|C_n(\hat{x}_1, \hat{x}_2)x_2 + G_n(\hat{x}_1)\|$. Therefore the controller law is given by

$$u = Gx - (2 + \phi(e, t)) \frac{\sigma}{\|\sigma\|}$$

where $\sigma = Fe$, $\alpha = 2$. In order to avoid the chattering problem, we can use the following control law

$$u = Gx - (2 + \phi(e, t)) \frac{\sigma}{(\|\sigma\| + \epsilon)} \quad (28)$$

where ϵ is a small positive scalar. The simulation results are shown in Figs. 4-7. Fig. 4. shows that the proposed SOSM observer can effectively estimate the velocities with parametric uncertainties and disturbances by using the position measurements. Trajectory tracking of joints 1 and 2 are shown in Fig. 5. This figure implies that the position of the joints are affected by the uncertainties and disturbances, but the controller law (28) guarantees the stability and it performs approximate rejection of the disturbances. The tracking errors are shown in Fig. 6. Fig. 7 shows the control signals and the sliding surfaces. As shown in this figure replacement of control

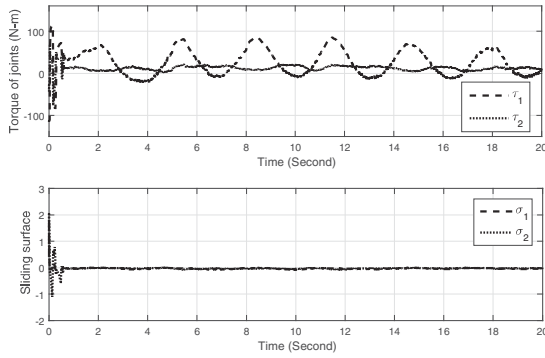


Fig. 7. Top: Control input of joints, dashed line: control input of joint 1 and dotted line: control input of joint 2. Bottom: Sliding surface value of $\sigma = Fe$, dashed line: sliding surface of joint 1 and dotted line: sliding surface of joint 2

law (28), can eliminate the chattering. These simulation results show that the output feedback sliding mode control with a \mathcal{H}_2 performance approach gives an acceptable performance and robustness in trajectory tracking in the presence of parametric and nonparametric uncertainties.

VII. CONCLUSION

In this paper, we have proposed an output feedback sliding mode controller with \mathcal{H}_2 performance for dynamic model of manipulator with parametric and nonparametric uncertainties. First a SOSM observer is used for velocity estimation of uncertain manipulator with position measurements. Then an sliding mode with \mathcal{H}_2 performance is designed. The controller parameter can be easily obtained using LMI and available scientific softwares. Using this approach, one can improve the system robustness and performance in the presence of uncertainties and disturbances. A numerical example has shown the effectiveness of the proposed method.

REFERENCES

- [1] ZH. Qu, "Global stability of trajectory tracking of robot under PD control," *Dynamics and Control*, vol. 4, no. 1, pp. 59-71, 1994.
- [2] R. Kelly, "PD control with desired gravity compensation of robotic manipulators: a review," *The International Journal of Robotics Research*, vol. 16, no. 1, pp. 660-672, 1997.
- [3] Q. Chen, H. Chen, YJ. Wang, PY. Woo, "Global stability analysis for some trajectory tracking control schemes of robotic manipulators," *Journal of Robotic Systems*, vol. 18, no. 1, pp. 69-75, 2001.
- [4] E.M. Jafarov, M.N.A. Parlakci, and Y. I Stefanopoulos, "A New Variable Structure PID-Controller Design for Robot Manipulators," *IEEE Transactions on Control Systems and Technology*, vol. 13, no. 1, pp. 122130, 2000.
- [5] G. Bartolini, A. Pisano, E. Punta, and E. Usai, "A Survey of Applications of Second-order Sliding Mode Control to Mechanical Systems," *International Journal of Control*, vol. 76, no. 9, pp. 875892, 2003.
- [6] W.H. Chen, "Disturbance observer based control for nonlinear systems," *Mechatronics, IEEE/ASME Transactions on*, vol. 9, no. 4, pp. 706-710, 2004.
- [7] A. Ferrara, and L. Magnani, "Motion Control of Rigid Robot Manipulators via First and Second Order Sliding Modes," *Journal of Intelligent and Robotic Systems*, vol. 48, no. 1, pp. 2336, 2007.
- [8] S. Islam and X.P. Liu, "Robust Sliding Mode Control for Robot Manipulators," *IEEE Transaction on Industrial Electronics*, vol. 58, no. 6, 2011.
- [9] M.S. Kang, "Disturbance Observer Based Sliding Mode Control for Link of Manipulator Driven by Elastic Cable," *Transactions of the Korean Society for Noise and Vibration Engineering*, vol. 22, no. 10, pp. 949-958, 2012.
- [10] V. Venkatesan, S. Mohan and J. Kim, "Disturbance observer based terminal sliding mode control of an underwater manipulator," *In Control Automation Robotics and Vision (ICARCV)*, 2014 13th International IEEE Conference on, pp. 1566-1572, 2014.
- [11] T.V. Tran, Y. Wang, H. Ao and T.K. Truong "Sliding Mode Control Based on Chemical Reaction Optimization and Radial Basis Functional Link Net for De-Icing Robot Manipulator," *Journal of Dynamic Systems Measurement and Control*, vol. 137, no. 2, pp. 051009, 2015.
- [12] JY. Choi, JS. Lee, "Adaptive iterative learning control of uncertain robotic systems," *IEEE Proc. Control Theory Appl*, vol. 147, no. 2, pp. 217-223, 2000.
- [13] C.C. Cheah, C. Liu, and J.J.E. Slotine, "Adaptive Tracking Control for Robots with Unknown Kinematic and Dynamic Properties," *The International Journal of Robotics Research*, vol. 25, no. 3, pp. 283296, 2006.
- [14] L. Tang, Y.J. Liu, S. Tong, "Adaptive neural control using reinforcement learning for a class of robot manipulator," *Neural Comput and Applie*, vol. 25, 125-141, 2014.
- [15] V. Pilania, K. Gupta "Hierarchical and adaptive mobile manipulator planner with base pose uncertainty," *Autonomous Robots*, pp. 1-21, 2015.
- [16] L. Cu villon, E. Laroche, J. Gangloff, M. de Mathelin, "Dynamic model and robust control of flexible link robot manipulator," *Robotics and Automation, Proceedings of the 2005 IEEE International Conference on*, pp. 4044-4049, 2005.
- [17] R.J. Wai, P.C. Chen, "Robust neural-fuzzy-network control for robot manipulator including actuator dynamics," *IEEE Transactions on Industrial Electronics*, vol. 53, no. 4, 2006
- [18] W. Shang and S. Cong, "Robust nonlinear control of a planar 2-DOF parallel manipulator with redundant actuation," *Robotics and Computer-Integrated Manufacturing*, vol. 30, no. 6, pp. 597-604, 2014.
- [19] Utkin, V. I., "Variable structure system with sliding Modes," *IEEE Transaction on Automatic Control*, AC-22, pp. 212-222 (1977).
- [20] Edwards, C., S. Spurgeon, and T. Francis, *Sliding Mode Control: Theory and Applications*, Taylor and Francis, London (1998).
- [21] Scherer, C. and S. Weiland, *Linear Matrix Inequalities in Control*, Lecture Note, Delft University of Technology (2005).
- [22] Wang, Y., L. Xie, and C. E. deSousa, "Robust control of a class of uncertain nonlinear systems," *Systems and Control Letters*, Vol. 19, pp. 139-149 (1992).
- [23] Boyd, S., L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*, SIAM Studies in Applied Mathematics (1994).
- [24] M. Sun, S.S. Ge, I.M.Y. Mareels, "Adaptive repetitive learning control of robotic manipulators without the requirement for initial repositioning," *IEEE Transactions on Robotics*, vol. 22, no. 3, pp. 563568, 2006.
- [25] J. Davila, L. Fridman, and A. Levant, "Second-order sliding-mode observer for mechanical systems," *IEEE transactions on automatic control*, vol. 50, no. 11, pp. 1785-1789, 2005.
- [26] J. A. Moreno and M. Osorio, "A Lyapunov approach to second-order sliding mode controllers and observers," *In Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, pp. 2856-2861, 2008.
- [27] H. H. Choi, "Output feedback variable structure control design with an \mathcal{H}_2 performance bound constraint," *Automatica*, vol. 44, pp. 2403-2408, 2008.
- [28] T.H.S. Li and Y.C. Huang, "MIMO adaptive fuzzy terminal sliding-mode controller for robotic manipulators," *Information Sciences*, vol. 180, no. 23, pp. 4641-4660, 2010.

On Periodically Pendulum-Driven Systems for Underactuated Locomotion: a Viscoelastic Jointed Model

Pengcheng Liu¹ (PhD student), *Student Member, IEEE*, Hongnian Yu¹ and Shuang Cang²

¹ *Faculty of Science and Technology, Bournemouth University, Poole BH125BB, UK*

² *Faculty of Management, Bournemouth University, Poole BH125BB, UK*

Abstract—This paper investigates the locomotion principles and nonlinear dynamics of the periodically pendulum-driven (PD) systems using the case of a 2-DOF viscoelastic jointed model. As a mechanical system with underactuation degree one, the proposed system has strongly coupled nonlinearities and can be utilized as a potential benchmark for studying complicated PD systems. By mathematical modeling and non-dimensionalization of the physical system, an insight is obtained to the global system dynamics. The proposed 2-DOF viscoelastic jointed model establishes a commendable interconnection between the system dynamics and the periodically actuated force. Subsequently, the periodic locomotion principles of the actuated subsystem are elaborately studied and synthesized with the characteristic of viscoelastic element. Then the analysis of qualitative changes is conducted respectively under the varying excitation amplitude and frequency. Simulation results validate the efficiency and performance of the proposed system comparing with the conventional system.

Keywords—pendulum-driven systems; periodic motion; underactuated; viscoelasticity

I. INTRODUCTION

Applications of underactuated mechanical systems (UMSs) have been penetrated into extensive branches of technology in the domain of robotics and control communities. These systems excel in performing complicated tasks with a reduced number of actuators, which imply an increased manoeuvrability, optimized energy consumptions as well as reduced cost.

Starting with these viewpoints, the motions with a repeated pattern at periodically intervals raise interests for various applications, for instance, the walking or running of the creatures, which under a regular pattern in their implementation. This attracts significant devotions to the trajectory planning and nonlinear control of UMSs by the robotics and control communities during the past few decades. The researchers are addressing both the theoretical difficulties [1]–[3] and the practical challenges [4]–[6]. Among these researches, the UMSs employing a pendulum or a system of the pendulums, which is referred to as PD UMSs, permits the investigations on selecting different important nonlinear effects. Attentions have been paid to the classical pendulum UMSs, as benchmarks, including the Acrobot [7], the Pendubot [8], the cart-pole system [9], the crane systems [10], Furuta pendulum systems [11]. Besides, numerous applications of such systems are known in engineering, for instance, in vibro-

absorption problems [12], [13], in trajectory tracking control of PD systems [14]–[16]. However, making a stabilized periodic motion trajectory (limit cycle) through feedback laws has been proved to be essential for nonlinear control.

The employment of viscoelastic property in the applications of UMSs has many advantages. For instance, higher bandwidth mechanical compliance, larger working space, better manoeuvrability, higher convergence rate and lower energy consumption are regarded as important indexes to evaluate the performance of the robot systems. Viscoelasticity has been studied extensively in the past two decades, including impact force reduction [17], trajectory planning [18], nonlinearities analysis such as hysteresis and friction [19], dynamic and static stability [20], etc. However, challenges are still remained in trajectory planning and controller design for the UMSs in the presence of strong coupled nonlinear dynamics, i.e., how to govern the nonlinear dynamics for the underactuated locomotion.

This paper investigates the periodic locomotion principles in the case of a 2-DOF PD system, which has potential applications such as pipeline inspection, medical assistance and information acquisition in disaster rescues. The aim of this paper is to shed light on the aforementioned nontrivial challenges by calling attentions to the issue of periodic motion trajectory synthesis and nonlinear dynamic analysis through numerical investigations of the characteristics of the proposed system.

The rest of the paper is organized as follows. Section II describes the formulation of the problem. Periodic locomotion principles synthesis is provided in Section III. Section IV investigates the system nonlinearities, and analyses the periodic and chaotic behaviours under varying excitation amplitude and frequency. Simulation results are presented in Section V. Finally, conclusions are given in Section VI.

II. PROBLEM FORMULATION

This study considers the nonlinear viscoelastic model shown in Fig.1, which consists of an inverted pendulum coupled with a 2-DOF spring-mass-damper system and is subjected to a periodical actuation applied at the pivot. The masses, spring and the dashpot in the 2-DOF system

are all identical and the locomotion of the proposed system is in horizontal plane.

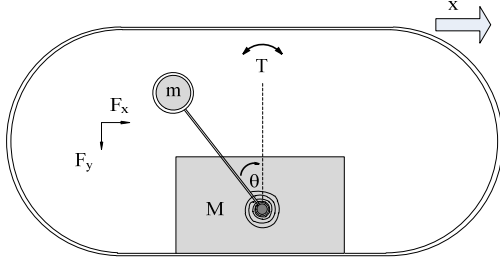


Figure 1. The 2-DOF pendulum-driven system with a viscoelastic joint

It is assumed that the centre of mass of the rotating mechanism is coinciding with the centre of the ball which is fixed rigidly at the end of the pendulum. Furthermore, the air frictional resistance is supposed to be zero when the pendulum is rotating. The torsional spring is unstretched when the inverted pendulum is upright. M and m are the masses of the base and the ball, respectively. l is the length of the inverted pendulum, θ and x depict the configuration variables of the rotational and the horizontal movements, i.e. $q_i = [q_1 \ q_2]^T = [\theta \ x]^T$, k and c represent the stiffness and damping coefficients, respectively. It is also assumed that the configuration variables θ and x are measured from the equilibrium position of the inverted pendulum and the original point of the base. Employing the Coulomb friction model to describe the resistance force between the proposed PD locomotive system and the environmental surface, gives

$$f = \begin{cases} 0, & \dot{x} = 0 \\ \mu F_y \text{sgn}(\dot{x}), & \dot{x} \neq 0 \end{cases} \quad (1)$$

where μ is the Coulomb friction coefficient, F_y represents the force applied on the platform in the vertical direction.

Based on the aforementioned assumptions and definitions, the equations of motion governing the dynamic behaviour the proposed model can be derived using the Euler-Lagrangian approach

$$\frac{d}{dt} \frac{\partial L(q_i, \dot{q}_i)}{\partial \dot{q}_i} - \frac{\partial L(q_i, \dot{q}_i)}{\partial q_i} + \frac{\partial D(\dot{q}_i)}{\partial \dot{q}_i} = Bu + Q_\zeta \quad (2)$$

where $L(q_i, \dot{q}_i)$ reflects the difference between the kinetic energy and the potential energy, $D(\dot{q}_i)$ describes the dissipative energy. $B \in \mathbb{R}^{n \times n}$ is a constant matrix, u is the control input, Q_ζ represents the effects of uncertainties and disturbances.

The dynamic equations of motion are given by

$$ml^2\ddot{\theta} - ml\cos\theta\ddot{x} - mgl\sin\theta + k\theta + c\dot{\theta} = T \quad (3)$$

$$\begin{aligned} -ml[\cos\theta + \mu\sin\theta\text{sgn}(\dot{x})]\ddot{\theta} + (M+m)\ddot{x} + ml[\sin\theta \\ - \mu\cos\theta\text{sgn}(\dot{x})]\dot{\theta}^2 + \mu[(M+m)g \\ - (k\theta + c\dot{\theta})\sin\theta/l]\text{sgn}(\dot{x}) = 0 \end{aligned} \quad (4)$$

where T is the rotational controlled torque applied to the inverted pendulum.

Our goal is to create periodic progression of the proposed system via an elaborate design of the rotational trajectory of the inverted pendulum and an appropriate

feedback action. Consequently, the following desired periodic function is adopted

$$T = A\cos(\Omega t) \quad (5)$$

where A and Ω are the amplitude and frequency of the periodic excitation, respectively.

We further define the following non-dimensional parameters

$$\begin{aligned} \tau = \omega_n t, \quad X = x/l, \quad \omega_n = \sqrt{g/l}, \quad \omega = \Omega/\omega_n, \quad \lambda = M/m, \\ \rho = k/ml^2\omega_n^2, \quad v = c/ml^2\omega_n, \quad h = A/ml^2\omega_n^2 \end{aligned} \quad (6)$$

Adopting the desired periodic function and the parameters above, Eq. (3) and Eq. (4) reduce to the following non-dimensional form

$$[\mathcal{M}]\{\mathfrak{X}\}'' + [\mathcal{C}]\{\mathfrak{X}\}' + [\mathcal{N}]\{\mathfrak{X}\} + [\mathcal{Q}] = \{\mathcal{U}\} \quad (7)$$

where

$$\begin{aligned} [\mathcal{M}] &= \begin{bmatrix} 1 & -\cos\theta \\ -[\cos\theta + \mu\sin\theta\text{sgn}(\dot{X})] & \lambda + 1 \end{bmatrix}, \\ [\mathcal{C}] &= \begin{bmatrix} v & \\ \sin\theta - \hat{a}\cos\theta & \end{bmatrix}, \quad [\mathcal{N}] = \rho \begin{bmatrix} 1 \\ -\hat{a}\sin\theta \end{bmatrix}, \\ [\mathcal{Q}] &= \begin{bmatrix} \sin\theta \\ \hat{a}(\lambda + 1) \end{bmatrix}, \quad \hat{a} = \mu\text{sgn}(\dot{X}), \quad \{\mathcal{U}\} = \begin{bmatrix} h\cos(\omega\tau) \\ 0 \end{bmatrix}. \end{aligned}$$

It is noted that the derivations above are conducted with respect to the dimensionless time τ and the configuration variables in the dimensionless time coordinate become $\{\mathfrak{X}\} = [\xi_1 \ \xi_2]^T = [\theta \ X]^T$.

Remark 1: In essence, the proposed PD locomotive system is a 2-DOF mechanical system with underactuation degree one. This nature results in the unavailability in the direct control of the platform's locomotion. Moreover, notwithstanding the fact that the proposed system is simple in structure, strong coupling and high nonlinearity exist in the dynamics which are originated from the trigonometric functions and the signal function. It is important to note that the sliding friction in the horizontal direction plays a vital role in the locomotion of the platform. These motivates the authors to scrutinize the characteristics of the periodic rotational torque and precisely design the locomotion principles for the actuated θ -subsystem.

III. PERIODIC LOCOMOTION PRINCIPLES SYNTHESIS

In this section, the periodic locomotion principles are generated for synthesizing the rotational motion of the inverted pendulum and the harmonic property of the viscoelastic element. It is considered that the nontrivial characteristic of viscoelastic element is equivalent to the existence of the periodic trajectory manifold with homologous arguments.

To effectively utilize the rotational motion of the pendulum and optimally drive the proposed 2-DOF system moving forward, the viscoelastic property is considered to synthesize the different periodic motions between the dissipated pendulum and the torsional spring, thus synthesized periodic locomotion principle is developed. In particular, three stages below are defined to generate the desired periodic locomotion.

Initialization ($\tau = 0$) and *re-initialization* stages ($\tau = \tau_7$): one cycle of progressive motion begins and ends respectively with the initialization and re-initialization stages. In initialization stage, the pendulum and the torsional spring are constrained and kept stationary at a predesigned negative angle to the opposite direction of the retraction of spring, which stores potential energy in such a manner that more mechanical power is injected into the entire system; at the end of the motion, the pendulum gradually returns to the initial position by following the motion profile, the system then is reinitialized with stored elastic energy for the new cycle.

Progressive stage ($\tau \in (0, \tau_3)$): the torque motor drives the pendulum fast in the forward direction, together with the energy-releasing of the torsional spring, leads the system to overcome the maximal dry friction and a continuous progression of the base is obtained;

Restoring stage ($\tau \in (\tau_4, \tau_7)$): the pendulum gradually returns to the initial position since the resultant force in the horizontal direction is less than the maximal dry friction, that is, the entire system is kept stationary in this stage of duration.

Therefore, the synthesized periodic locomotion profile is generated by Eq. (8) and shown in Fig. 2, wherein the zoom up window demonstrates the detailed profile in the progressive stage.

$$\dot{\theta}_{Td} = \begin{cases} P_1 \omega \sin(\omega \tau), & \tau \in [0, \tau_1) \\ P_1 \omega, & \tau \in [\tau_1, \tau_2) \\ P_1 \omega \sin(\omega \tau - \tau_2), & \tau \in [\tau_2, \tau_3) \\ \frac{\tau_3 - \tau}{\tau_3 - \tau_2} P_2, & \tau \in [\tau_3, \tau_4) \\ \frac{\tau_4 - \tau}{\tau_4 - \tau_3} P_3, & \tau \in [\tau_4, \tau_5) \\ -P_3, & \tau \in [\tau_5, \tau_6) \\ \frac{\tau_6 - \tau}{\tau_6 - \tau_5} P_3, & \tau \in [\tau_6, \tau_7) \end{cases} \quad (8)$$

where P_1 and P_2 respectively describe the upper and lower boundaries of the motion trajectory, P_3 is the critical boundary when the system begins to keep stationary, ω is frequency of periodic excitation.

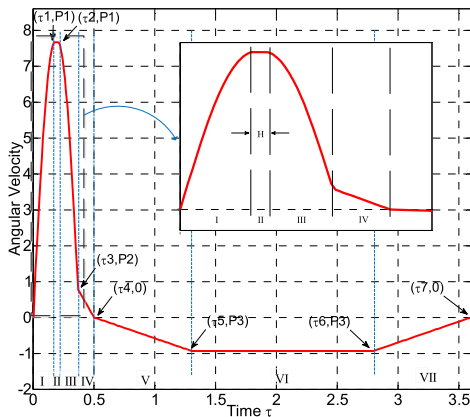


Figure 2. Synthesized periodic locomotion profile for one full cycle

The explicit description of the periodic locomotion principles are as follows:

Initialization $\tau = 0$: $\theta(\tau) = \theta_{min} = -\theta_0$, $X(\tau) = 0$, $\dot{\theta}(\tau) = 0$, $\dot{X}(\tau) = 0$, $\ddot{\theta}(\tau) = 0$, $\ddot{X}(\tau) = 0$. The pendulum

together with the torsional spring is kept stationary at a predesigned negative angle $-\theta_0$ to the opposite direction of the retraction of spring, which stores potential energy such that more mechanical power will be provided.

Phase I $\tau \in (0, \tau_1)$: $\theta(\tau) = \theta > 0$, $X(\tau) = x$, $\dot{\theta}(\tau) > 0$, $\dot{X}(\tau) > 0$, $\ddot{\theta}(\tau) \gg 0$, $\ddot{X}(\tau) > 0$. The torque motor begins to move under the synthesized angular velocity and simultaneously the stored potential energy is released from the stretched torsional spring. This results in a motion with maximal angular acceleration of the pendulum pushing the base moving forward with relatively high acceleration.

Phase II $\tau \in [\tau_1, \tau_2)$: $\theta(\tau) = \theta > 0$, $X(\tau) = x$, $\dot{\theta}(\tau) > 0$, $\dot{X}(\tau) > 0$, $\ddot{\theta}(\tau) = 0$, $\ddot{X}(\tau) > 0$. It is noted that once the potential energy is released, a short period of time is required to let the potential energy fully transfer into kinetic energy of the proposed system. This leads to more efficient energy consumption. Thus a short period of uniform motion of the pendulum is designed. During this period, the pendulum swings forward with the maximal angular velocity while driving the base accelerating continuously.

Phase I $\tau \in [\tau_2, \tau_3)$: $\theta(\tau) = \theta > 0$, $X(\tau) = x$, $\dot{\theta}(\tau) > 0$, $\dot{X}(\tau) > 0$, $\ddot{\theta}(\tau) < 0$, $\ddot{X}(\tau) < 0$. The torque actuation exerts an opposing force on the pendulum under the synthesized angular velocity together with the contractility of the torsional spring. This leads to a forward deceleration motion of the pendulum as well as the base.

Phase $\tau \in [\tau_3, \tau_4)$: $\theta(\tau) = \theta_{max} > 0$, $X(\tau) = x \rightarrow 0$, $\dot{\theta}(\tau) \rightarrow 0$, $\dot{X}(\tau) = 0$, $\ddot{\theta}(\tau) < 0$, $\ddot{X}(\tau) = 0$. In phase IV, a slow deceleration motion of the pendulum results in the stationary of the base, which is subjected to the constraints under the dissipative force lie in the sliding surface as well as the pivot. Moreover, the angular displacement of the pendulum is constrained at θ_{max} to avoid over-actuation and system failure.

Phase $\tau \in [\tau_4, \tau_5)$: $\theta(\tau) = \theta < 0$, $X(\tau) = x$, $\dot{\theta}(\tau) < 0$, $\dot{X}(\tau) = 0$, $\ddot{\theta}(\tau) < 0$, $\ddot{X}(\tau) = 0$. Phase V is designed to be a short duration and to generate a relatively low angular acceleration of the pendulum which keeps the base stands still.

Phase $\tau \in [\tau_5, \tau_6)$: $\theta(\tau) = \theta < 0$, $X(\tau) = a\Delta x$, $\dot{\theta}(\tau) = -P_3 < 0$, $\dot{X}(\tau) = 0$, $\ddot{\theta}(\tau) = 0$, $\ddot{X}(\tau) = 0$. A uniform angular velocity of is designed for the purpose of gradually stretching the torsional spring such that enough potential energy is restored for the next cycle. The base remains stationary in this phase. $a\Delta x$ represents the net displacement of the base after the a^{th} cycle.

Phase I $\tau \in [\tau_6, \tau_7)$: $\theta(\tau) = \theta < 0$, $X(\tau) = a\Delta x$, $-P_3 < \dot{\theta}(\tau) < 0$, $\dot{X}(\tau) = 0$, $\ddot{\theta}(\tau) > 0$, $\ddot{X}(\tau) = 0$. In phase VII, a low angular deceleration motion is generated in a short duration to decelerate the pendulum while the base keeps stationary;

Re-Initialization $\tau = 0$: $\theta(\tau) = \theta_{min} = -\theta_0$, $X(\tau) = a\Delta x$, $\dot{\theta}(\tau) = 0$, $\dot{X}(\tau) = 0$, $\ddot{\theta}(\tau) = 0$, $\ddot{X}(\tau) = 0$. When the pendulum reaches to the initial angle, the torsional spring

is constrained to θ_{min} such that enough elastic energy is stored for the next cycle.

Remark 2: It is worth mentioning that the net progression during one full motion cycle occurs in the progressive stage, in which the synthesis procedure is mainly carried out. Moreover, as one of the key elements regarding to the progression of the whole system, the friction between the platform and the sliding surface is taken into account for designing the restoring stage through the consideration of the system constraints. The proposed periodic locomotion principles can be utilized for generating a class of appropriate trajectory profiles for PD underactuated mechanical systems with viscoelastic elements. The trajectory synthesis occurred at the progressive stage is the main enhancement comparing with the work in [21]. To obtain an optimal progression of the proposed system for one full motion cycle and to avoid unpredictable chaotic motions, it is necessary to find the optimal amplitude and frequency of the periodic force.

IV. NONLINEAR DYNAMIC ANALYSIS

Due to the fact that the proposed system is analytically unsolvable, a sequence of solutions is numerically calculated using the first order Euler algorithm in Matlab. In this section, we employ a visual interpretation on the dynamic behaviour affected respectively by the amplitude and frequency of the periodic force, and the stability variance of solutions accompanied by the varying values.

A. Qualitative Analysis of Amplitude h

The bifurcation diagram in Fig. 3 presents a projection of the Poincaré map on the dimensionless configuration axis. It clearly illustrates the richness of the system dynamics along with various transitions in the system response. It is noted that a large region of period-one response can be observed for $h \in [0.15, 1.6419]$. Accompanied by increasing the excitation amplitude, a large window of chaotic motion is depicted for h in $(1.6419, 5.1]$.

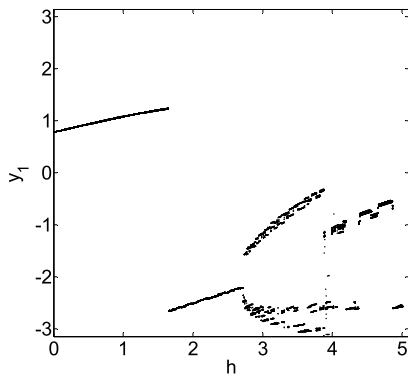


Figure 3. Bifurcation diagram of progression of the PD system under varying excitation amplitude

Figure 4 describes the time trajectories of the angular displacement of the inverted pendulum computed for various amplitudes of excitations. The characteristic of the irregular transitions under varying excitation amplitude, behaves atypical responses. These originate from the complicated interactions between different coexisting

periodic obits and bifurcations. The time histories of the angular displacement are important to appreciate the behaviours illustrated. At relatively low amplitude of excitation as shown in Fig. 4 (a), the pendulum employs simple but steady oscillation after the initial transients have decayed, which repeat continuously. On the other hand, the motion contained in Fig. 4 (b) becomes chaotic at relatively high amplitude of excitation, which are extremely complex nonrepeating functions of time.

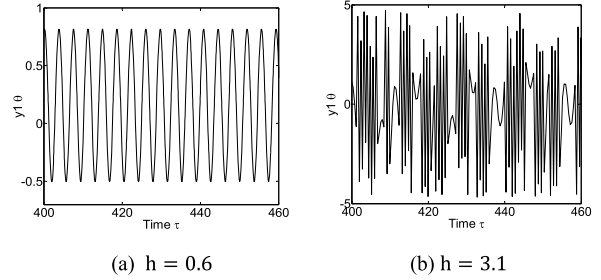


Figure 4. The time histories of the angular displacement of the inverted pendulum on dimensionless time coordinate

B. Qualitative Analysis of Frequency ω

The parameter dependence on varying frequency ω is studied as the second branching parameter and clearly shown as a bifurcation diagram in Fig. 5. It can be seen that the motion of the proposed system behaves atypical chaotic response for $\omega \in [0.01, 1.2215]$. On the other hand, for $\omega \in (1.2215, 5.1]$, a response of period-one is recorded for the rest of the values of excitation frequency.

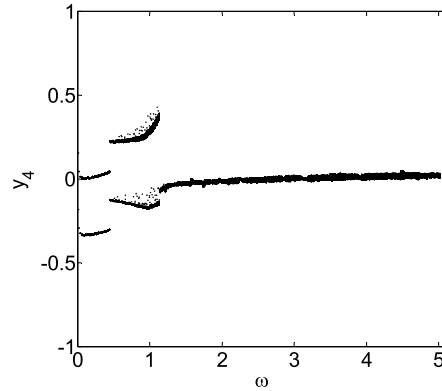


Figure 5. Bifurcation diagram of average velocity of the PD system under varying excitation frequency

The numerical investigation also reveals that the characteristic of the irregular transitions, accompanied with the increases of the values of the branching parameter, behaves atypical responses from chaotic to periodic. These are resulted from the complicated interaction between different coexisting periodic obits and bifurcations.

The time histories of the angular displacement computed for various frequencies of excitations are presented as well, in which the dynamic behaviors are illustrated. The motion contained in Fig. 6 (a) behaves chaotic response when at relatively lower frequency, which is an extremely complex nonrepeating function of time. On the other hand, at relatively higher frequency of

excitation as shown in Fig. 6 (b), the pendulum employs simple but steady-state rotations after the initial transients have decayed, which would repeat continuously.

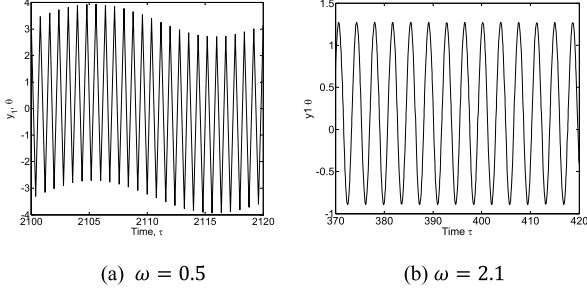


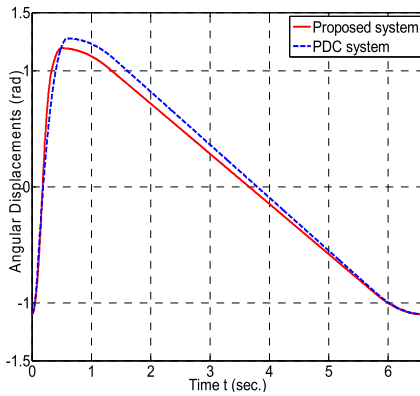
Figure 6. The time histories of the angular displacement of the inverted pendulum on dimensionless time coordinate

V. SIMULATIONS

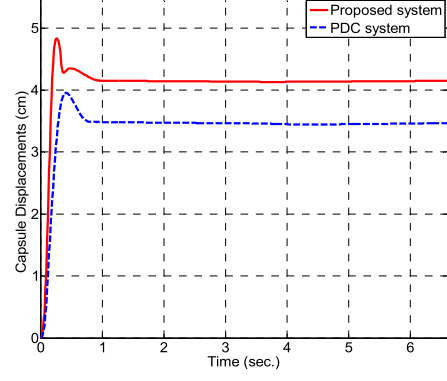
In this section, numerical simulations are presented to verify the effectiveness of the proposed system by employing the closed-loop controller designed in [16]. The performance of the synthesized periodic locomotion principle is also evaluated. The control objective is to make the pendulum track the synthesized locomotion trajectory and simultaneously drive the whole system moving rectilinearly overcoming the environmental resistances.

The numerical simulation results are obtained through MATLAB. The system parameters are selected as $M = 0.5 \text{ kg}$, $m = 0.05 \text{ kg}$, $l = 0.3 \text{ m}$, $g = 9.81 \text{ m/s}^2$, $\mu = 0.01 \text{ N} \cdot \text{m}^{-1} \text{s}^{-1}$, $\rho = 0.9$, $v = 0.6$, $h = 1$, $\omega = 1.7$. The initial conditions are adopted as $\theta(0) = -\theta_0 = -\pi/3$, $\dot{\theta}(0) = 0$, $x(0) = 0$, $\dot{x}(0) = 0$.

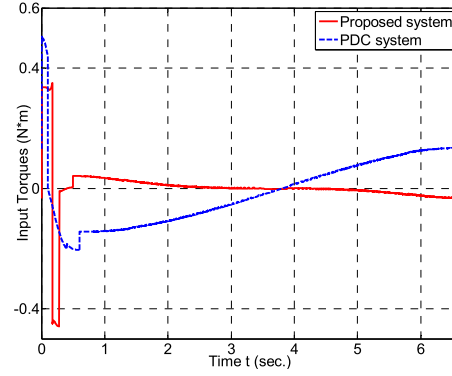
A series of simulations are conducted in comparison to the pendulum-driven cart-pole system proposed in [21], which is referred to as PDC system. Heuristically, the parameter selection approach employed here is trying to find the optimal values such that the best system response is achieved. Fig. 7 shows the comparison results for one cycle in time histories of actuated and passive subsystems, respectively. It is noted that the proposed system periodically actuated under the synthesized locomotion principles behaves steady and intermittent progressive motions. More interestingly, the proposed system and the PDC system travel 4.15cm and 3.5cm, respectively.



(a) Angular displacement



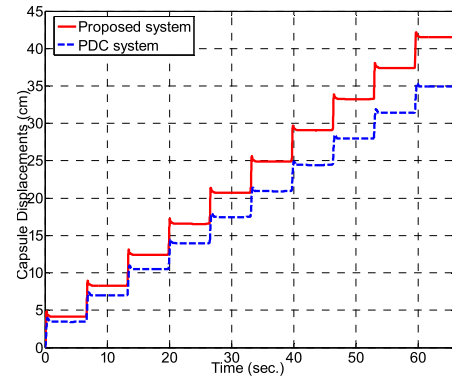
(b) Displacements of the base



(d) Input torques

Figure 7. Time histories under closed-loop control for one full cycle

To further evaluate the sequential performance of the proposed system, simulation for ten cycles are conducted as shown in Fig. 8. The proposed system and the PDC system advance 41.5278cm and 34.9335cm, which demonstrate that the PD system has higher efficiencies of 15.88% in progression calculated from Fig. 8 (a). On the other hand, the maximum input torque respectively for the proposed system and PDC system is 0.4582N*m and 0.5037N*m. The maximum angular displacement respectively for the proposed system and PDC system are 1.2059rad and 1.2775rad. Therefore the energy consumptions for the proposed system and the PDC system, respectively, are 0.5525J and 0.6435J, which means the proposed system has a 16.46% higher energy efficiency calculated from Fig. 8 (b).



(a) Displacements of the base

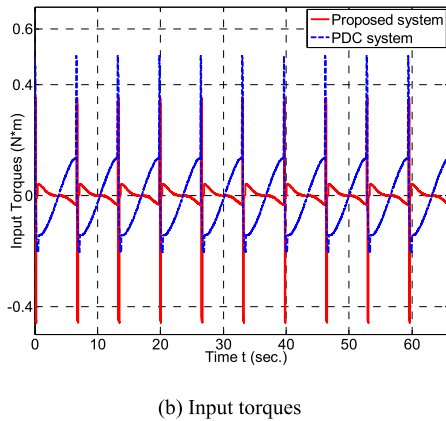


Figure 8. Time histories under closed-loop control for ten full cycles

VI. CONCLUSIONS

The issues of periodic locomotion principles synthesis and nonlinear dynamic analysis are studied in this paper. Mathematical models have been established and utilized as a benchmark of numerical analysis to optimize the excitation parameters. The periodic locomotion principles of the actuated subsystem are elaborately studied and synthesized with the characteristic of viscoelastic element. Then the qualitative changes are conducted respectively under the varying excitation amplitude and frequency. Time histories of the pendulum demonstrate a wide variety of system responses, which vary from periodic to chaotic. It is noted that based on the qualitative analysis of the system dynamics, a series of optimal parameters can be obtained, which sheds light on the linkage between nonlinear analysis and trajectory planning for efficient underactuated locomotion. The simulation results demonstrate the promising performance in progression and energy consumption.

ACKNOWLEDGMENT

This work is supported by the match funded studentship provided by Bournemouth University and Zhongyuan University of Technology and also supported by the EU Real-time Adaptive networked control of rescue roBOTS (RABOT) project.

REFERENCES

- [1] R. Olfati-Saber, "Nonlinear control of underactuated mechanical systems with application to robotics and aerospace vehicles," Massachusetts Institute of Technology, 2000.
- [2] M. W. Spong, "Underactuated mechanical systems," in *Control problems in robotics and automation*, Springer, 1998, pp. 135–150.
- [3] N. Sun and Y. Fang, "New energy analytical results for the regulation of underactuated overhead cranes: an end-effector motion-based approach," *Ind. Electron. IEEE Trans. On*, vol. 59, no. 12, pp. 4723–4734, 2012.
- [4] Y. Fang, B. Ma, P. Wang, and X. Zhang, "A motion planning-based adaptive control method for an underactuated crane system," *Control Syst. Technol. IEEE Trans. On*, vol. 20, no. 1, pp. 241–248, 2012.
- [5] J. W. Grizzle, G. Abba, and F. Plestan, "Asymptotically stable walking for biped robots: Analysis via systems with impulse effects," *Autom. Control IEEE Trans. On*, vol. 46, no. 1, pp. 51–64, 2001.
- [6] C.-L. Hwang and H.-M. Wu, "Trajectory tracking of a mobile robot with frictions and uncertainties using hierarchical sliding-mode under-actuated control," *IET Control Theory Appl.*, vol. 7, no. 7, pp. 952–965, 2013.
- [7] X. Xin and T. Yamasaki, "Energy-based swing-up control for a remotely driven Acrobot: Theoretical and experimental results," *Control Syst. Technol. IEEE Trans. On*, vol. 20, no. 4, pp. 1048–1056, 2012.
- [8] F. B. Mathis, R. Jafari, and R. Mukherjee, "Impulsive Actuation in Robot Manipulators: Experimental Verification of Pendubot Swing-Up," *Mechatron. IEEEASME Trans. On*, vol. 19, no. 4, pp. 1469–1474, 2014.
- [9] L.-C. Hung and H.-Y. Chung, "Decoupled control using neural network-based sliding-mode controller for nonlinear systems," *Expert Syst. Appl.*, vol. 32, no. 4, pp. 1168–1182, 2007.
- [10] N. Sun, Y. Fang, Y. Zhang, and B. Ma, "A novel kinematic coupling-based trajectory planning method for overhead cranes," *Mechatron. IEEEASME Trans. On*, vol. 17, no. 1, pp. 166–173, 2012.
- [11] L. Freidovich, A. Shiriaev, F. Gordillo, F. Gomez-Estern, and J. Aracil, "Partial-energy-shaping control for orbital stabilization of high frequency oscillations of the Furuta pendulum," in *Decision and Control, 2007 46th IEEE Conference on*, 2007, pp. 4637–4642.
- [12] C. Shi and R. G. Parker, "Modal properties and stability of centrifugal pendulum vibration absorber systems with equally spaced, identical absorbers," *J. Sound Vib.*, vol. 331, no. 21, pp. 4807–4824, 2012.
- [13] R. Lima, C. Soize, and R. Sampaio, "Robust design optimization with an uncertain model of a nonlinear vibro-impact electro-mechanical system," *Commun. Nonlinear Sci. Numer. Simul.*, 2014.
- [14] H. Yu, Y. Liu, and T. Yang, "Closed-loop tracking control of a pendulum-driven cart-pole underactuated system," *Proc. Inst. Mech. Eng. Part J. Syst. Control Eng.*, vol. 222, no. 2, pp. 109–125, 2008.
- [15] H. Yu, T. Yang, Y. Liu, and S. Wane, "A further study of control for a pendulum-driven cart," *Int. J. Adv. Mechatron. Syst.*, vol. 1, no. 1, p. 44, 2008.
- [16] P. Liu, H. Yu, and S. Cang, "Modelling and control of an elastically joint-actuated cart-pole underactuated system," in *2014 20th International Conference on Automation and Computing (ICAC)*, 2014, pp. 26–31.
- [17] M. H. Korayem, H. N. Rahimi, and A. Nikoobin, "Mathematical modeling and trajectory planning of mobile manipulators with flexible links and joints," *Appl. Math. Model.*, vol. 36, no. 7, pp. 3229–3244, 2012.
- [18] E. G. Papadopoulos and D. A. Rey, "A new measure of tipover stability margin for mobile manipulators," in *Robotics and Automation, 1996. Proceedings., 1996 IEEE International Conference on*, 1996, vol. 4, pp. 3111–3116.
- [19] M. Ruderman, "Modeling of Elastic Robot Joints with Nonlinear Damping and Hysteresis," in *Robotic Systems - Applications, Control and Programming*, A. Dutta, Ed. InTech, 2012.
- [20] U. Bhattiprolu, P. Davies, and A. K. Bajaj, "Static and Dynamic Response of Beams on Nonlinear Viscoelastic Unilateral Foundations: A Multimode Approach," *J. Vib. Acoust.*, vol. 136, no. 3, p. 031002, 2014.
- [21] H. Yu, Y. Liu, and T. Yang, "Closed-loop tracking control of a pendulum-driven cart-pole underactuated system," *Proc. Inst. Mech. Eng. Part J. Syst. Control Eng.*, vol. 222, no. 2, pp. 109–125, 2008.

Fault Tolerant BLDC Motor Control for Hall Sensors Failure

Sova V., Chalupa J., Grepl R.

Mechatronics laboratory (www.mechlab.cz)
Brno University of Technology, Faculty of Mechanical Engineering
Brno, Czech Republic
sova@fme.vutbr.cz

Abstract—In most brushless direct current (BLDC) motor drives, there are three hall sensors as a position reference. This paper presents a method that allows the operation of a BLDC motor with one faulty hall sensor. The situations considered are when the output from a hall sensor stays continuously at low or high levels, or a short-time pulse appears on a hall sensor signal. For fault detection, identification of a faulty signal and generating a substitute signal, this method only needs the information from the hall sensors. So this method can be added as a standalone subsystem to existing drives. The paper provides many simulations on how the presented method reacts for various types of faults. An experimental evaluation is provided too. Due to the demand to operate at a high speed, the method is implemented into the FPGA.

Keywords- fault detection; brushless motor; BLDC

I. INTRODUCTION

Brushless direct current (BLDC) motors are becoming more popular. Their advantages are: a simple control mechanism, power density and long service life. They are increasingly used in different industrial and commercial applications. They are also used in applications where the failure of a BLDC motor drive can lead to big economic losses or even safety threats, e.g. in the automotive [1][2] or aerospace [3][4] industries. An example could be a turboprop engine fuel pump powered by a BLDC motor. There is a demand for a high degree of reliability. In case of fault, the running fuel pump must remain functional, but the start-up of a pump is not necessary [5].

Many parts in a BLDC motor drive can malfunction. For example: the open-circuit or short-circuit fault of a switch, voltage or current sensor fault in a power inverter. Motor faults: the inter-turn short circuit in winding, overheating, bearing failure, phase open-circuit fault, etc. And position sensors are subjected to faults too. In addition to these faults, every part and signal of a BLDC motor drive is subjected to interferences.

On almost all BLDC motors, three hall sensors shifted by 120° el. (electrical degrees) are used as a position reference for control algorithm [6]. Our proposed method tries to preserve the functionality of a BLDC motor drive during a hall sensor fault. It should preserve the functionality of a BLDC motor drive during a long-term fault of a hall sensor or even during short-term pulses on hall sensor signals caused mostly by interferences.

Fault detection and controlling various systems during faults is the current subject of scientific research. In areas

around BLDC motors new contributions still arise. Jeong et al. [7] deals with various faults in interior permanent magnet motor drives. In the case of a position sensor fault, they very briefly say, it can be detected when the estimated rotor position differs from the expected and in that case, the algorithm switches to a sensorless control scheme. The faults in a power inverter and remedial strategies for that case are in [8][9]. Especially for hall sensor faults [10][11] are dedicated.

Tashakori and Ektesabi [10] only address the situation when a hall sensor stays permanently at a low or high level. The proposed algorithm is divided into three parts: fault detection, identification of a faulty signal and generating a substitute signal. The time between a hall sensor going wrong and a substitute signal being used is quite long.

Firmansyah et al. [11] check the sequence of hall signals states, which are defined by hall sensor signals. The transition to a following state is allowed only if the change of hall sensor signal defines a following state and if the change happens within an expected time.

II. METHOD PROPOSAL

There was a demand to use only hall sensors signals, so a fault tolerant BLDC motor control algorithm based only on checking hall sensor signals was created. The algorithm can be implemented in a device which can form a standalone unit in between a motor and an ECU (Fig. 1). Obviously the algorithm can be implemented into the existing ECU.

In Fig. 2 there is a block diagram illustrating the algorithm functionality. There are three main building blocks in the algorithm. The blocks shown in red are responsible for generating a substitute signal from two fine signals.

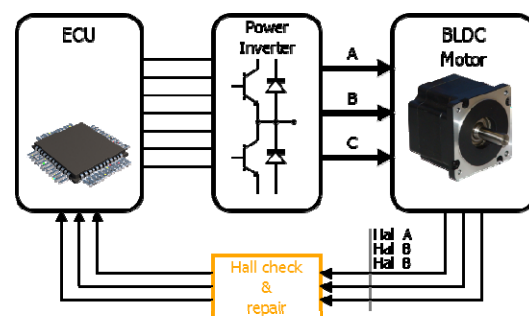


Figure 1. Location of proposed algorithm in a BLDC motor drive

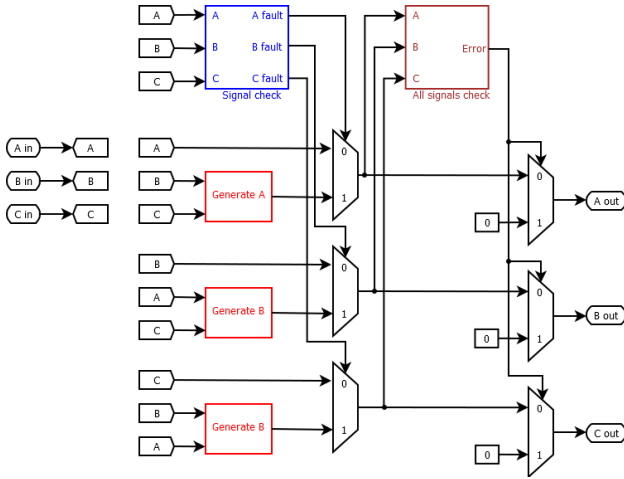


Figure 2. Overall block diagram of proposed algorithm

There are three independent generating blocks in this algorithm and they are generating the substitute signals continuously. Signal A is generated using signals B and C, signal B is generated using A and C, and C is generated using A and B.

The next important block in the algorithm is the block called the Signal check. The Signal check block is responsible for recognizing a faulty signal. When a faulty signal is detected, then this block sets the corresponding error output to the high level and the faulty signal is replaced by its substitute.

The last important block in the algorithm is the block increasing the reliability, but in some cases it can be omitted. This block sets all the signals to the low level in case of a fault, before the faulty signal is switched to the substitute signal. Because the recognition of a faulty signal by the Signal check block is not immediate, there might occur a sequence on signals, which may lead to switching the wrong transistors in a power inverter, thus causing overcurrent. We assume that setting all the signals to a low level is a forbidden combination and the ECU switches all the transistors in a power inverter off. The time of leaving out is very short (less than rotating by 180° el.) and the rotor overcomes it by its inertia.

Because a lot of the parts of the algorithm are based on measuring time intervals in signals and generating signals, the proposed algorithm requires good resolution on the time axis and multiple simultaneous measurements, so the algorithm was targeted into the FPGA.

In accordance with the Rapid Control Prototyping technique, dSPACE [12] hardware as a target device and Xilinx System Generator for DSP™[13] as a high-level tool for generating the FPGA code were used.

The three main building blocks are described in more detail:

A. Generating a substitute signal from two fine signals

Signal waveforms from hall sensors are periodic, rectangular, with a phase shift of 120° (Fig. 3). Edges in hall sensor signals indicate the time of commutation for a motor.

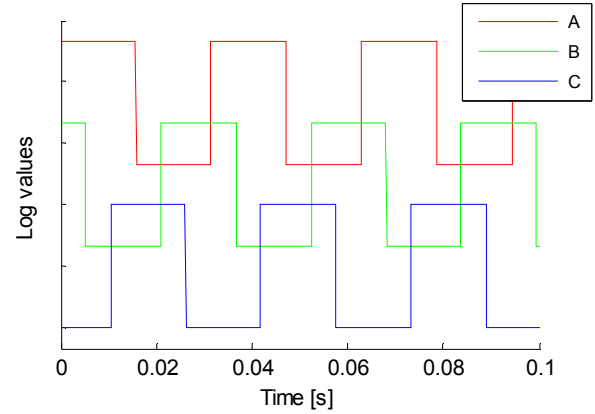


Figure 3. Waveforms of hall sensor signals

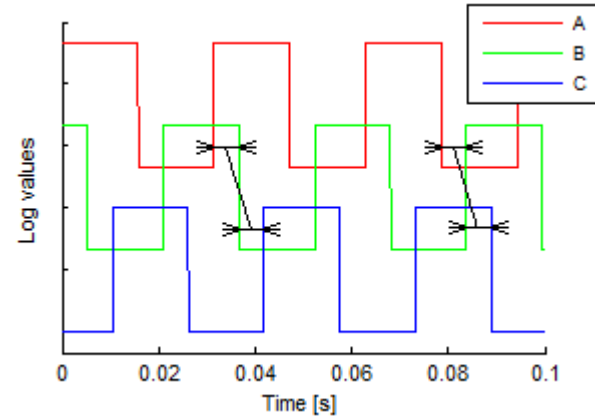


Figure 4. Generating a substitute signal according to two fine signals

Theoretically it should be able to generate two other signals from one functioning signal. If the signal period can be measured, divided by 3 (a phase shift of 120°) and this time measured out from the signal edges, it should be possible to generate two other signals from one functioning signal. In reality, it turns out that this approach is unacceptable. Thanks to manufacturing inaccuracies in hall sensor mount positions and inaccuracies in the sensing ring for hall sensors, the resulting signal edges come too far from the actual moment of commutation. This results in increased noise, bigger vibrations and worse efficiency. The substitute signals can be generated from an average time from a few last periods of the signal. This improves the behaviour in a steady state, but leads to worse transient states.

So the focus was kept on generating one signal from two others, and using the most recent time information, not averaging. The time interval between the falling edge of one signal and rising edge of a second signal was measured. This time we measured out from the rising edge of the second signal and then the falling edge of the generated signal was made. And similarly for the second edge (Fig. 4).

FPGA implementation of this building block is in Fig. 5. The upper part of the diagram is responsible for setting the generated signal to the low level and the bottom part for setting it to the high level. The upper part is described in more detail: The falling edge of signal A

resets a counter which increments with every FPGA clock tick. The rising edge of signal B stores the value of the first counter and resets the second counter. When the value of the second counter equals the stored value from the first counter, the generated signal is set to the low level.

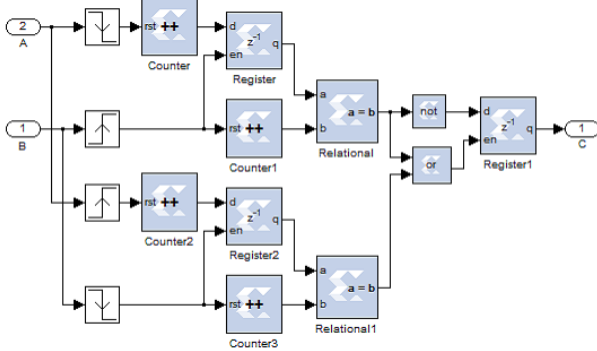


Figure 5. FPGA implementation of Generating block algorithm

B. Signal check - recognizing the faulty signal

There exist many options for how to recognize a faulty signal. The faulty signal can be recognized by the fact that the change in hall signals is expected at a certain time [11]. This method fails during fast transients because the expected time of a signal change differs from the actual one. In [10] they recognize the faulty signal by analyzing the waveforms of the terminal voltages. Even sensorless BLDC motor control methods can be used for recognizing the faulty signal [7].

In our proposed method, the sequence of signal changes is checked to recognize the faulty signal. This is primarily due to the fact that it only needs the signals from the hall sensors. Also good behaviour during transient states can be expected, because the time information of the signals is not needed.

The sequences of signal changes are checked for three pairs of hall signals (AB, BC, AC). In the case of an individual signal fault, two pairs report the fault and we know that the faulty signal is the one which is common for both pairs.

For each signal pair, during the positive direction of rotation, the repeating sequence 00, 10, 10, 11, 01, 01 (Table 1) can be seen. It is checked to see whether this sequence is kept and, if not, the algorithm reports a fault for the signal pair.

TABLE I. HALL SENSORS STATES DURING ONE ELECTRICAL REVOLUTION

State number	\square [°el.]	A	B	C
1.	0-60	0	1	0
2.	60-120	0	1	1
3.	120-180	0	0	1
4.	180-240	1	0	1
5.	240-300	1	0	0
6.	300-360	1	1	0

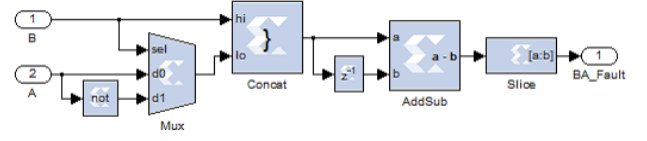


Figure 6. FPGA implementation of signal pair check

The FPGA implementation of this algorithm is in Fig. 6. The sequence 00, 10, 11, 01 (repeating states are omitted) is transformed to sequence 00, 01, 10, 11, which represents binary counting by inverting the second signal when the first one is at the high level. Then these two signals are concatenated into one two-bit signal and the delayed sample of this signal is subtracted from the current sample. If the result of subtracting is 0 or 1, the change of the signal state was correct, otherwise it was incorrect. Taking the first bit (counting from zero) from the results, the fault indicator signal is obtained.

The fault indicator signal indicates a problem in one or both signals. To identify the specific signal, every pair of the fault indicator signals is carried to AND gates. At the end of this part which recognizes the faulty signal, there are registers in which the values of individual signal fault indicators are stored until a new edge on the signals occurs.

C. All signals check

This block is based on checking the sequence of signal changes on all three signals. If the change of the state number is directly following the previous state, we know that it's all right, otherwise there is a fault. A look-up table for recognizing the state number was implemented into the FPGA. Individual hall signals were concatenated into one 3-bit signal and this signal is used to choose the state number from the block which behaves like a ROM memory. In the ROM memory block is information from table 1. The state number is then subtracted from the delayed one and the result is checked to see whether it is within the expected range. The error signal is stored in a register and can change only if there is a change in some signal from the hall sensors.

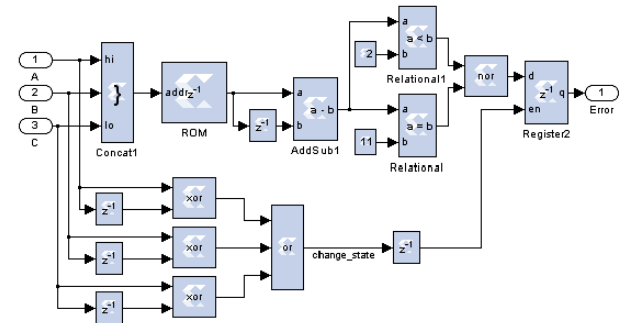


Figure 7. FPGA implementation of all signals check block

III. SIMULATION RESULTS

First of all, the behaviour of the proposed method under various conditions was simulated. The focus was kept on situations where the signal from the hall sensor stays permanently at the low or high level and when there

is a short time pulse on the signal. These errors were simulated for the motor in steady state conditions and also during the changes of rotational speed.

A. Permanent low or high fault

When one of the signals stays permanently at the low or high level, a very short moment when all signals are at the low level occurs. This suspends transistors switching in a power inverter, thus preventing them from wrong switching, which may lead to overcurrent. After this short leave out moment, the substitute signal is used instead of the faulty one.

In Fig. 8 the leave out time (all signals at the low level) is very short. It is represented only by a separate peak at time 0.018s. Generally the leave out time can last between 0-180°el.

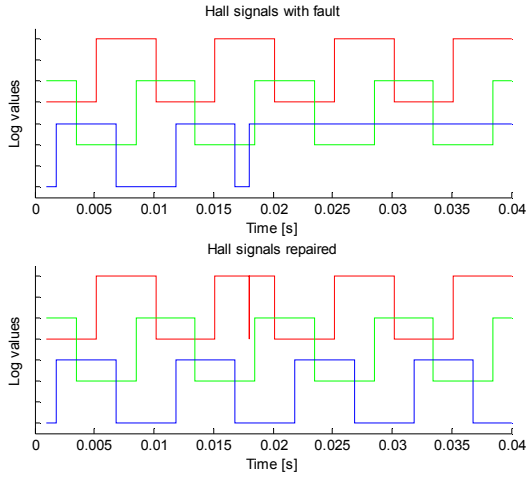


Figure 8. Hall signals with a permanent high fault

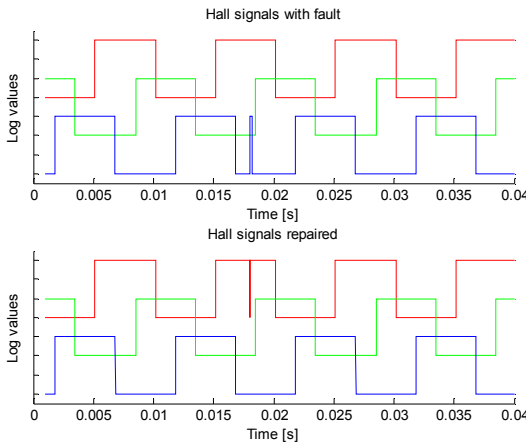


Figure 9. Hall signals with a short time pulse fault

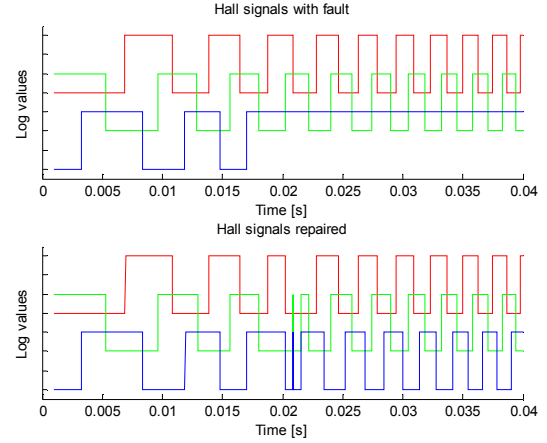


Figure 10. Hall signals with a permanent high fault during motor acceleration

B. Short time pulse fault

In the simulation for a short time pulse on otherwise fine signals (typically caused by interferences), a situation where all signals are at the low level for safety reasons appears for a brief moment, until the previous parts of the algorithm react correctly. For the situation in Fig. 9, the time where all the signals are at the low level is very short and almost immediately it is switched to the simulated signal. After the pulse is out and the original signals are fine, the simulated signal is replaced by the original one.

C. Permanent low fault during acceleration

There is no problem to repair a faulty signal even during acceleration of the motor. As usual a brief moment appears when all the signals are at the low level and after that the substitute signal is used (Fig. 10). The substitute signal is generated using the most recent edges in signals, without filtering, so the missing signal is substituted quite well.

IV. EXPERIMENTAL EVALUATION

Various simulations showed good results, so the experimental evaluation followed. The algorithm was tested on our laboratory BLDC motor test bench, which consists of a BLDC motor [15], torque meter and a DC motor which acts as a load. The power inverter is a Microchip MC1L [16]. The proposed algorithm is implemented on the modular dSPACE Hardware with a DS5203 FPGA board [17].

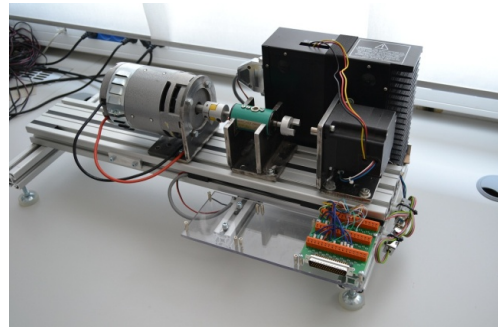


Figure 11. BLDC motor test bench available in the Mechatronics laboratory [14]

For the experimental evaluation, besides the proposed algorithm, a simple controller for the BLDC motor and an algorithm for simulating faults were implemented on the dSPACE FPGA board (Fig. 12). The controller of the BLDC motor was represented only by a simple commutation table, which assigns transistor switching commands for every combination of hall sensor signals.

To manage the experiment, the control GUI in ControlDesk® [18] that allowed us to switch between normal mode and safe mode and simulate long-term and short-term errors on hall sensor signals was created.

A. Permanent low or high fault

In a time of 55ms a high level fault on signal A was simulated. From the plots in Fig. 13, we can see that at the moment of the fault, the current starts to grow until a leave-out time occurs. A moment later a substitute signal was generated.

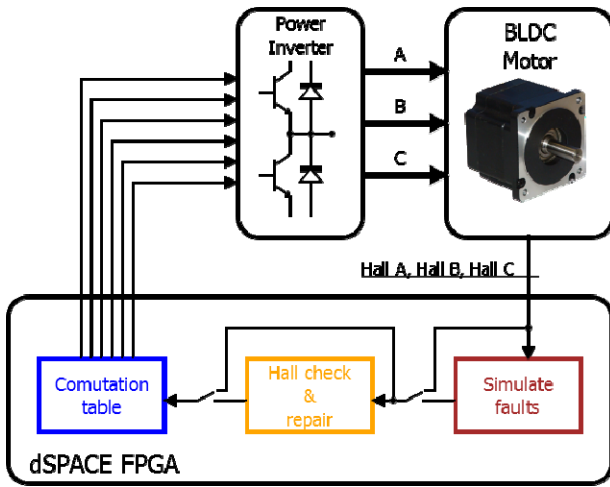


Figure 12. Setup for experimental evaluation

TABLE II. IMPORTANT PARAMETERS OF BLDC MOTOR B8672-48 [15]

Nominal voltage	48 V
Nominal power	117 W
No load speed	3700 rpm
Nominal speed	3102 rpm
Nominal torque	0.359 Nm
Nominal current	19.4 A
Number of pole pairs	4

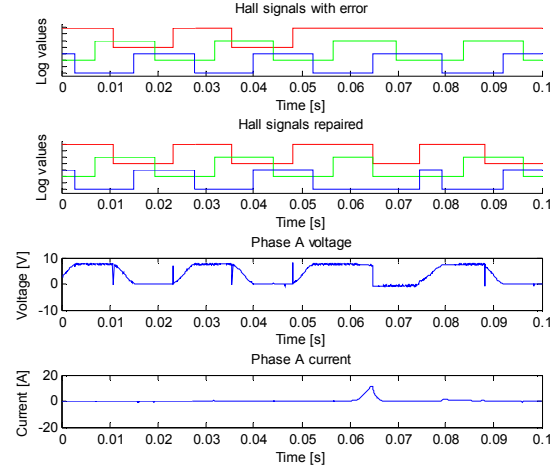


Figure 13. Hall signals, phase A terminal voltage and current under a permanent high fault

B. Short time pulse fault

Short-time pulses lasting 3 ms were generated with a period of 40 ms. In Fig. 14 we can see that the first pulse was immediately repaired, even without a leave-out time. The second pulse caused a short moment where all the signals are at the low level.

We can see that the presented waveforms are in accordance with the simulation results.

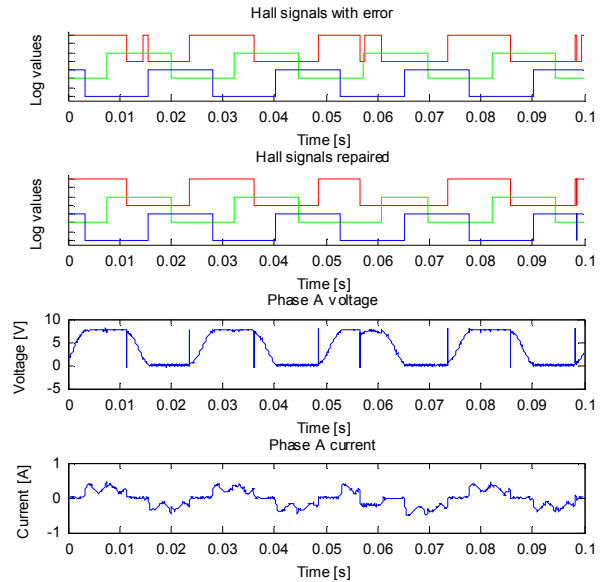


Figure 14. Hall signals, phase A terminal voltage and current waveforms under a pulse fault

C. Permanent low fault during acceleration

The behavior when a fault happens during acceleration was investigated too. The presented waveforms are in accordance with the simulation results.

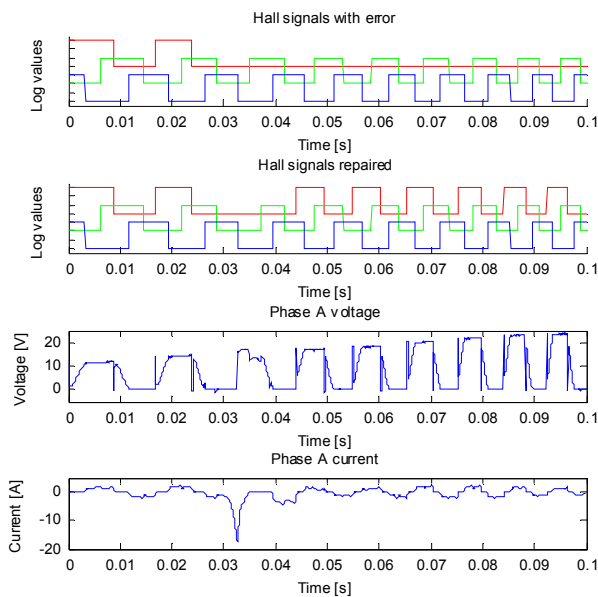


Figure 15. Hall signals, phase A terminal voltage and current under a permanent low fault during acceleration

V. CONCLUSION

The algorithm that allows the operation of a BLDC motor drive with one faulty hall position sensor was presented. Faults when the hall sensor stays permanently at the low or high level, or a short-time pulse appears on a hall sensor signal, are repaired. The algorithm was verified on numerous simulations and in a real experiment too. FPGA was used for implementation, so the reaction time is very short and the motor speed is almost unlimited. The presented algorithm does not allow start-up or near zero speed operation of a motor, because it is impossible to measure the pulse length on hall sensors signals when the motor is not rotating. But once the motor is rotating, the functionality during a fault is maintained. Plenty of methods exist on how to maintain an open-loop start-up of a BLDC motor. The proposed algorithm only needs the information from the hall sensors so it can be implemented into an existing solution as a standalone subsystem. One of the possible ways to improve the proposed method is to merge our algorithm for recognizing a faulty signal with an algorithm that checks the time when a signal change occurs [11].

ACKNOWLEDGMENT

This work was supported by the European Commission within the FP7 project Efficient Systems and Propulsion for Small Aircraft "ESPOSA", contract No. ACP1-GA-2011-284859-ESPOSA, and by NETME CENTRE PLUS (LO1202) created with financial support from the Ministry of Education, Youth and Sports under the „National Sustainability Programme I“.

REFERENCES

- [1] Jianwen Shao; Nolan, D.; Teissier, M.; Swanson, D., "A novel sensorless brushless DC (BLDC) motor drive for automotive fuel pumps," *Power Electronics in Transportation*, 2002, vol., no., pp.53,59, 24-25 Oct. 2002
- [2] Jun-Hyuk Choi; Se-Hyun You; Jin Hur; Ha-Gyeong Sung, "The Design and Fabrication of BLDC Motor and Drive for 42V Automotive Applications," *Industrial Electronics*, 2007. ISIE 2007. IEEE International Symposium on, vol., no., pp.1086,1091, 4-7 June 2007
- [3] Laxminarayana, Y.; Tarakalyani; Ravindranath, "“Design of a compact BLDC motor for high power, high bandwidth Rotary ElectroMechanical Actuator for aerospace application”," *India Conference (INDICON)*, 2011 Annual IEEE, vol., no., pp.1,4, 16-18 Dec. 2011
- [4] Banerjee, T.P.; Roychoudhury, J.; Das, S.; Abraham, A., "Hybrid Intelligent Predictive Control System for High Speed BLDC Motor in Aerospace Application," *Emerging Trends in Engineering and Technology (ICETET)*, 2010 3rd International Conference on, vol., no., pp.258,262, 19-21 Nov. 2010
- [5] Andrs, O.; Hadas, Z.; Kovar, J., "Introduction to design of speed controller for fuel pump," *Mechatronics - Mechatronika (ME)*, 2014 16th International Conference on, vol., no., pp.672,676, 3-5 Dec. 2014
- [6] Elvich L.N., 3-Phase BLDC Motor Control with Hall Sensors Using 56800/E Digital Signal Controllers, URL: <http://cache.freescale.com/files/product/doc/AN1916.pdf>, June 2015
- [7] Yu-Seok Jeong; Seung-Ki Sul; Schulz, S.E.; Patel, N.R., "Fault detection and fault-tolerant control of interior permanent-magnet motor drive system for electric vehicle," *Industry Applications*, IEEE Transactions on, vol.41, no.1, pp.46,51, Jan.-Feb. 2005
- [8] Byoung-Gun Park; Tae-Sung Kim; Ji-Su Ryu; Dong-seok Hyun, "Fault Tolerant Strategies for BLDC Motor Drives under Switch Faults," *Industry Applications Conference*, 2006. 41st IAS Annual Meeting. Conference Record of the 2006 IEEE, vol.4, no., pp.1637,1641, 8-12 Oct. 2006
- [9] Tashakori, A.; Ektesabi, M., "Fault diagnosis of in-wheel BLDC motor drive for electric vehicle application," *Intelligent Vehicles Symposium (IV)*, 2013 IEEE, vol., no., pp.925,930, 23-26 June 2013
- [10] Tashakori, A.; Ektesabi, M., "A simple fault tolerant control system for Hall Effect sensors failure of BLDC motor," *Industrial Electronics and Applications (ICIEA)*, 2013 8th IEEE Conference on, vol., no., pp.1011,1016, 19-21 June 2013
- [11] Firmansyah, E.; Wijaya, F.D.; Aditya, W.P.R.; Wicaksono, R., "Six-step commutation with round robin state machine to alleviate error in hall-effect-sensor reading for BLDC motor control," *Electrical Engineering and Computer Science (ICEECS)*, 2014 International Conference on, vol., no., pp.251,253, 24-25 Nov. 2014
- [12] dSPACE GmbH, URL: <https://www.dspace.com>, April 2015
- [13] System Generator for DSP, Xilinx Inc., URL: <http://www.xilinx.com/products/design-tools/vivado/integration/sysgen.html>, April 2015
- [14] Mechatronics Laboratory, Faculty of Mechanical Engineering, Brno University of Technology, URL: <http://mechlab.fme.vutbr.cz/>, April 2015
- [15] B86 series brushless dc motors, Transmotec Inc., URL: <http://www.transmotec.com/brushless-dc-motors/no-gear/b86-series.aspx>, April 2015
- [16] MC1L 3-Phase Low Voltage Power Module Microchip Technology Inc., URL: <http://ww1.microchip.com/downloads/en/DeviceDoc/70097A.pdf>, April 2015
- [17] DS5203 FPGA Board Overview, dSPACE GmbH, URL: https://www.dspace.com/en/pub/home/products/hw/modular_hardware_introduction/i_o_boards/ds5203_fpga_board.cfm, April 2015
- [18] dSPACE ControlDesk Overview, dSPACE GmbH, URL: <https://www.dspace.com/en/pub/home/products/sw/expsoft/controldesk.cfm>, April 2015

Design of a Fault Tolerant Redundant Control for Electro Mechanical Drive System

Grepl, R., Matejasko, M., Bastl, M., Zouhar, F.

Mechatronics laboratory (www.mechlab.cz)

Brno University of Technology, Faculty of Mechanical Engineering

Brno, Czech Republic

grepl@fme.vutbr.cz

Abstract—This paper deals with a design of a electro-mechanical actuator (EMA) control system with advanced safety requirements. Based on the Failure Mode and Effects Analysis (FMEA), a redundant structure of sensors and electronic control units (ECUs) is proposed. Following the compromise between safety and simplicity, only two channels (sensors and ECUs) are used. A Voter (hardware which routes the output signals from ECUs to power electronics) is implemented on a complex programmable logic device (CPLD) using hardware description language (HDL). As a case study, the proposed system structure is implemented on a self-balancing vehicle control system. The Model Based Design (MBD) approach is applied during the system development and testing. ECUs are based on dsPIC Microchip devices and programmed using automatically generated C code.

Fault tolerant control, redundant control system, EMA, CPLD, self-balancing vehicle

I. INTRODUCTION

One of the most common mechatronic applications is electro-mechanical actuator (EMA) control. These actuators, usually at aerospace, automotive and military fields, are replacing original hydraulic, pneumatic or mechanical solutions [1],[2],[3]. In some applications it is essential to ensure system reliability that exceeds the reliability of an ordinary solution and therefore it is necessary to apply redundancy on some components.

A. Fault tolerant control of EMA

Solution complexity depends on which components of the system are redundant or multiplied. Considering the information and energy flows, one can use redundant components in following order:

- Sensors
- Signal processing and Control electronics
- Power electronics
- Electrical part of the EMA
- Mechanical part of the EMA.

Using redundant sensors is the most common cost effective way of increasing the reliability. An example is using two potentiometers to sense a valve rotation angle in electronic throttle [4],[2]. In case of three-sensor usage it is simple to decide on correctness of the sensor output. When mounting multiple sensors is difficult or overall system cost is restricted, physical sensor might be

substituted by a software one (indirect control value calculation from other measured data).

From implementation point of view, the complete actuation system duplication is considered to be the most difficult. This might be solved by connecting both redundant EMA with the last stage of a mechanical gearbox [5]. The Engines can be connected both in parallel and in series [4]. Interesting compromise arrangement might be doubling the electrical part of EMA (redundant motor winding) while using common mechanical part [6].

In this article we are presenting a solution that increases system reliability by doubling sensors and electronic control units (ECUs). Power electronics and EMA are not redundant. This system arrangement was developed based on thorough Failure Mode and Effects Analysis (FMEA) considering the desired low cost and inability to attach redundant EMA. The contribution of this article is the description of hardware and software structure for a fault tolerant system created with emphasis on maximum simplicity of the design.

B. Control of a self-balancing transporter as a case study for fault tolerant system

In this article we demonstrate a general fault tolerant control system design on a specific example of the design of a self-balancing vehicle. The vehicle is developed as a long-term student project with high educational potential.

The control of a self-balancing transporter is an interesting scientific and also educational topic [7],[8],[9], that is still an issue [10]. Several approaches are used and tested for the control design including simple PID, state-space or fuzzy regulators [7]. An important issue is to determine tilt angle which cannot be measured directly. Remarkably practical is so called *complementary filter* [11], although there is for example an approach using a camera [9].

The rest of this paper is organized as follows: in section II. a general fault tolerant system structure will be described. The scheme is designed with respect to the required function as well as the maximum simplicity of the system. In section III. we will show a specific application of this general scheme on a case of the self-balancing vehicle. Further, in section IV., an experimental testing of the implemented fault tolerant control system is briefly mentioned.

II. STRUCTURE OF A FAULT TOLERANT SYSTEM WITH REDUNDANT SENSORS AND CONTROL UNITS

A. System requirements

Proposed fault tolerant system must be able to appropriately react on following fault states/events:

- Sensor malfunction
- ECU computation error / Total failure of ECU

B. Schematic of the fault tolerant system

The system (Fig. 1) is composed of two control units. Each unit has its own (redundant) sensors. Both subsystems are galvanically isolated.

Control outputs of both units (usually PWM, DIR, EN) are connected to a switch (Voter). In order to increase the reliability of the whole system and to provide fast reaction time, the Voter is implemented using complex programmable logic device (CPLD). As well as the control outputs, the control units are generating a *check signal*. It is a software-generated pulse signal on a digital output pin (not using a PWM peripheral). This ensures that in a case of ECU malfunction, the signal stops being generated.

Control units are connected by bus through which they are interchanging measured and calculated data.

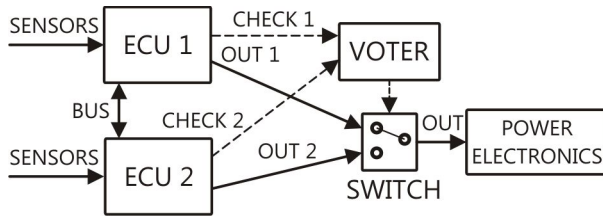


Figure 1. Schematic of a fault tolerant system with redundant ECUs

C. Functionality of the fault tolerant system

1) Voter functionality

Control logic in the Voter is shown in Fig. 2. It is a state machine that routes either one of the two control units or disconnects the whole system. After an initialization of the system, the Voter is routing first ECU. Transitions among particular states are carried out based on an evaluation of the *check signal* from both control units.

2) ECU functionality

Several parallel processes, generally with various priority and sample time, are implemented inside each control unit. These are:

- EMA control algorithm – typically PID regulator supplemented with linear or nonlinear feed-forward and/or feed-back friction compensation [2].
- Sensor signals check – it is a fundamental fault detection (FD) functionality. Basically the algorithm is comparing the measured values with

min. and max. limits. Eventually gradients of filtered signals might be checked [12].

- A communication with second ECU through bus and measured data comparison.
- *Check signal* generation (software pulses).

Standard procedure is to design and implement each unit separately in a different manner, for example using different microcontroller platform and developing by another development team.

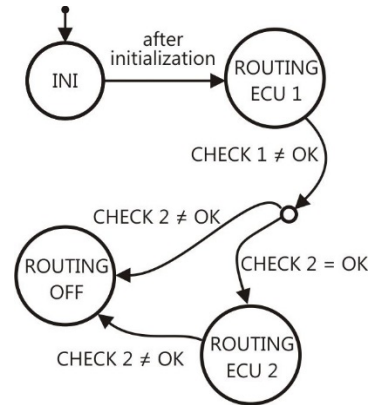


Figure 2. Schematic of the control logic in the Voter

3) System reactions to fault states/events

System reactions to fault states/events are the following:

a) *Sensor malfunction*: can be caused by a measurement out of limits (typically for analog output sensors), communication failure (sensors with SPI, I²C, SENT, etc. buses) or data gradient crossing permissible range. For sensor functionality monitoring, a simple method based on data noise might be used.

Action: Control unit checks the condition of the redundant unit (via bus). If the second unit reports no error, the first unit turns off its *check signal* generation. The Voter reacts to this by routing the second control unit outputs. The whole system enters an emergency mode. A warning for the passenger is generated and system requires a safe switch off.

b) *ECU malfunction*: can be a total or a partial microcontroller failure. In such a case it is likely that a short-time or complete check signal interruption will happen.

Action: see case a) above.

c) *ECU computation error*: this type of error can be detected by output gradient monitoring and/or I/O data comparison with FD algorithm. In this case an action performed would be the same as in previous cases.

d) *Calculated values difference between ECUs or bus malfunction*: this type of error can be detected through information interchange via bus. Based only on the outputs differences between ECUs, it cannot be determined which ECU is calculating correctly and which is not. As long as there is no significant change in the outputs (see case c) above), it usually does not have to signify serious degradation of the system behaviour. However, for full system operation it is always necessary

that both units are operating simultaneously (neither is detecting an error in the system).

Action: control unit nr. 1 stays active, but enters an emergency mode.

III. CASE STUDY: FAULT TOLERANT CONTROL OF SELF-BALANCING TRANSPORTER

In this part we will describe an application of the general structure of the fault tolerant control system on a specific example of the control of a self-balancing vehicle. The vehicle was developed in our laboratory according to the Model Based Design (MBD) approach, i.e. with maximum model usage in the design process. During the development, different control strategies were tested using offline simulations. Afterward a real prototype model was used to estimate model parameters and algorithm testing (Fig. 3). The use of the Rapid Control Prototyping (RCP), when the real system was connected through MF624 I/O card and driven directly from Simulink, significantly accelerated the design process [13],[14].

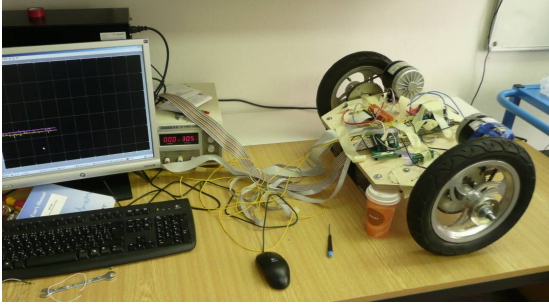


Figure 3. Testing platform at very early stage connected to Simulink via MF624

A. Estimation of tilt angle using Complementary Filter (CF)

The control of a self-balancing vehicle is similar to a control of an inverted pendulum from the control design point of view. The main difference is that tilt angle cannot be measured directly. We can estimate it indirectly by using sensors:

- **Accelerometer** – the inclination is calculated as a projection of gravity into the horizontal axis of the sensor. Unfortunately, also the forward acceleration is added to the measured signal and therefore the estimation is very imperfect.
- **Gyro** – the angle is obtained as the integration of the measured angular velocity. Unfortunately, every real gyro sensor suffers from the drift.
- **Combination of accelerometer and gyro** – this combination might eliminate disadvantages of both separate sensors above. Except Kalman Filter it is possible to use simple Complementary filter (CF) [10]. Difference equation of CF is as follows:

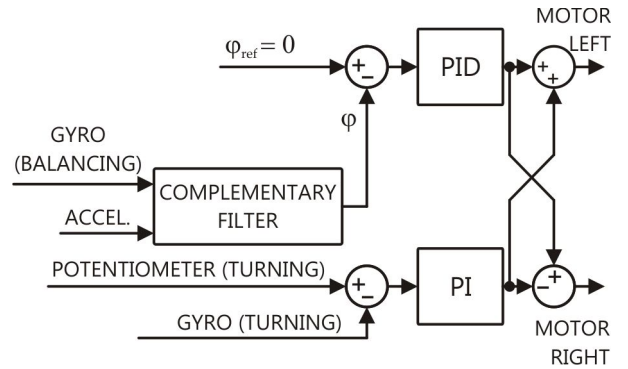
$$\varphi_k = (1-c)(\varphi_{k-1} + \dot{\varphi}_{GYRO}T_s) + c\varphi_{ACC} \quad (1)$$

where $\dot{\varphi}_{GYRO}$ is the gyro signal (angular velocity) and φ_{ACC} is the angle calculated from the accelerometer signal. Constant c represents the scale of integration

influence of angular velocity vs. acceleration projection, and needs to be specified. The actual control algorithm runs with $T_s = 10ms$ and uses $c = 0.02$. Thus the gyro is significantly the major input to the filter and only small influence of the accelerometer is used to the drift compensation.

B. Balancing and turning controllers

For driving a self-balancing vehicle it is essential to control the tilt and turning. Fig. 4 shows a diagram of both regulators. Balancing is driven by PID regulator and turning by PI regulator. The input value for turning controller is a potentiometer signal which measures handlebar tilt angle. It can be seen, that output of the turning regulator modifies the output duty cycle by adding or subtracting the difference for left and right wheel



respectively.

Figure 4. Scheme of the tilt and turning controllers (CF – Complementary Filter)

C. Mechanical design of the vehicle

The self-balancing transporter developed at our laboratory is interesting also from the mechanical construction point of view. Unlike of typical torque transmission from motor to wheels using a chain or a toothed belt, we use structurally more complicated direct attachment of the motor shaft to the driven wheel. Outer design appearance is shown in Fig.5, cross-section of the wheel mounting in Fig.6.



Figure 5. Detailed view of DC motors and wheel mounting

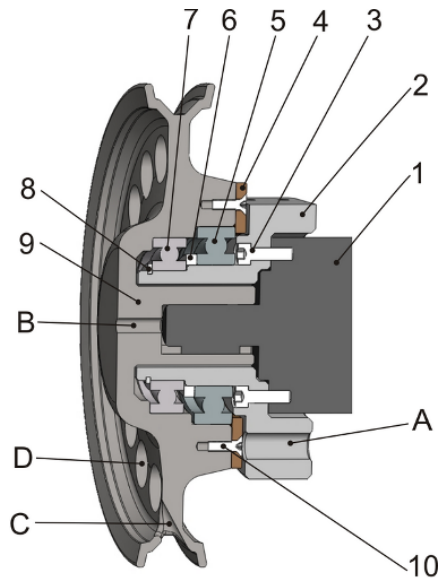


Figure 6. Cross-section of the wheel mounting (1 – motor, 2 – grip, 3 – screw, 4 – axial fixator, 5 – ball bearing, 6 – distance ring, 7 – small ball bearing, 8 – safety ring, 9 – wheel, 10 – screw fixing the wheel, A – mounting slot, B – demounting thread, C – lightening)

D. Implementation of the redundant control system

1) Hardware details

Following the scheme shown in Fig. 1, the redundant control unit was implemented using two dsPIC microcontrollers (ECUs) and CPLD (Voter). The modular board is shown in Fig. 7.

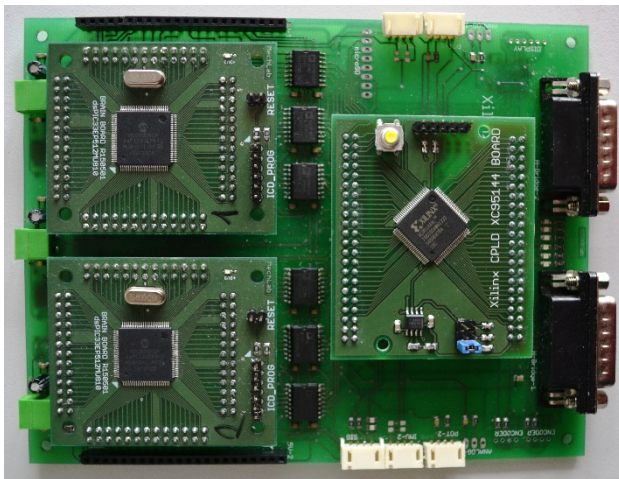


Figure 7. Redundant control system composed of two ECUs (Microchip dsPIC) and one Voter (CPLD Xilinx).

Utilizing CPLD as the Voter provides several advantages: the CPLD is highly reliable, the execution is very fast (in order of nanoseconds) and the internal logic is easily reprogrammable. In comparison with FPGA, the logic inside CPLD is active immediately after switching on.

During the development process we were using CoolRunner-II CPLD Starter Board which is equipped with XC2C256 circuit. Inter logic was designed in hardware description language (HDL) programming language using webPACK environment.

The Voter routes control outputs of the control units to the power electronics. It is monitoring the *check signal* and in a case of the signal interruption it switches the control outputs to the second ECU or disconnects both ECUs (Fig.2.).

2) Fault-tolerant software details

Fig.8 shows a scheme of data flows and algorithm sections, running on the ECUs. The block “CONTROL” implements the control algorithm shown in Fig.5. The block “LIMIT AND GRADIENT CHECKING” processes sensor signals for the “FAULT DETECTION” block.

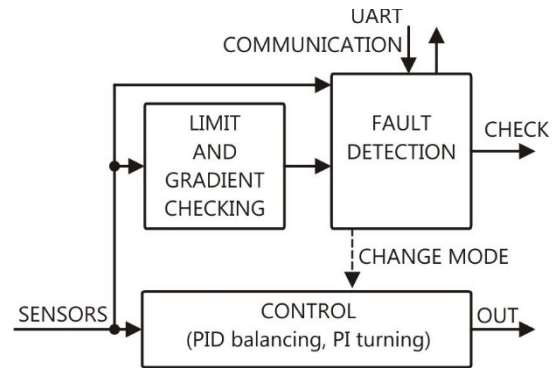


Figure 8. Data flows and algorithm sections of the ECU program

In the “FAULT DETECTION” block, there are all the control algorithms implemented. These algorithms are monitoring absolute sensor values, sensors functionality (using noise), second ECU status, number of data bytes received, output duty cycle and bus functionality. FD reactions to errors are described in II.C.3. The block also includes the state machine shown in Fig. 9, which controls the operational mode of the system.

Signal “CHANGE MODE” triggers an algorithm inside the CONTROL block that slows down and eventually stops the vehicle. This is done by decreasingly limiting the maximum duty cycle, also accompanied by a passenger warning.

The dsPICs were programmed using automatically generated C code from Simulink with Embedded Coder and MPLAB Device Blocks.

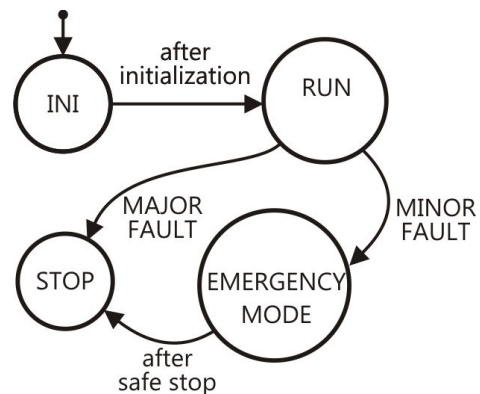


Figure 9. Operating states of the program inside ECU

IV. EXPERIMENTAL TESTING USING HIL SIMULATION

A. Description of experimental platform

The designed and implemented redundant control system (Fig. 7) was tested using the Hardware-In-the-Loop simulation (HIL). Fig. 10 shows the setup of the experiment performed on the dSPACE DS1103 real-time hardware platform.

The “VEHICLE MODEL” includes the simple nonlinear model of the system with parameters estimated using measured I/O data.

The “SENSORS MODEL” simulates the output of gyros, accelerometer and potentiometers of the real system. These signals are represented using 14-bits DACs on dSPACE hardware.

Several tests of FD algorithm were simulated using this setup including various sensor faults, ECU faults or communication problems.

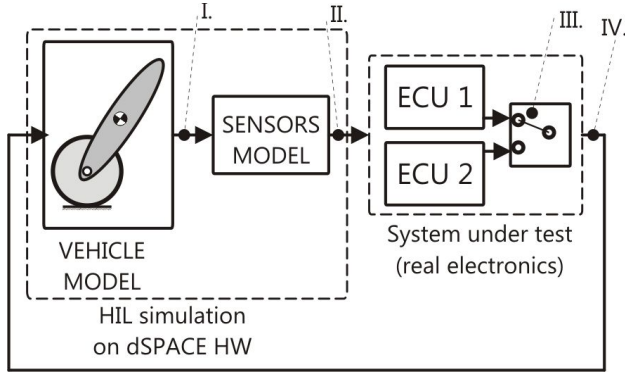


Figure 10. Scheme of the HIL simulation setup (signals I.-IV. refers to the results shown in Fig. 11)

B. Example of HIL test

One of frequent sensor malfunctions leads to saturation of its output signal. Fig. 11 shows the response of the control system to the simulation of the fault. The mechanism of the fault insertion in signal II. provides the saturation of acceleration sensor.

Approx. at 9.5 s of the simulation, the sensor output value is suddenly saturated to the lower bound. The fault is detected by master ECU 1, the *check signal* is stopped and consequently, the Voter starts routing the ECU 2 (can be seen in Fig. 11 III.). Further, the balancing is correctly maintained.

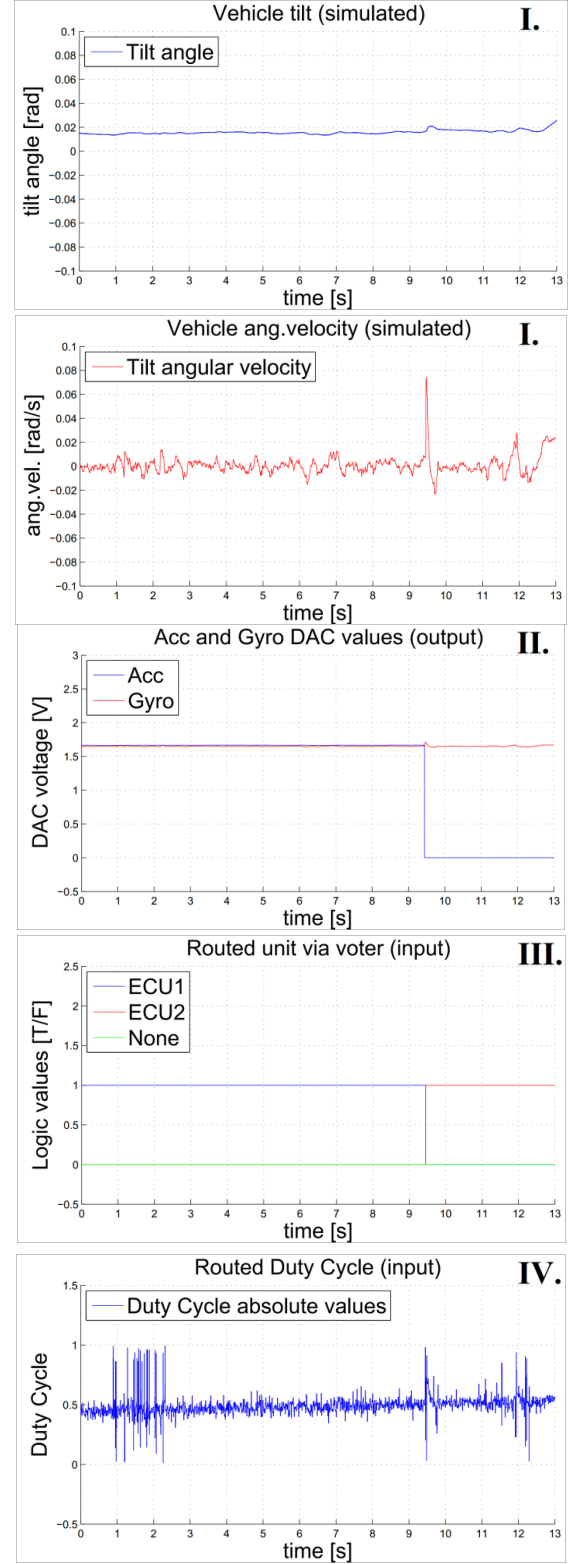


Figure 11. Experimental results: example of FD test (Figures I. – IV. corresponds with signals shown in Fig. 10).

V. CONCLUSION

This paper presents the efficient and simple fault tolerant scheme for an EMA.

A general scheme is studied and implemented for a particular task of a safe control of a self-balancing transporter. All requirements were based on a long time experiences with a student's project of the self-balancing vehicle developed in our laboratory. During the operation, several problems with sensors and power electronics were observed which creates the motivation for our research and development.

Based on the thorough FMEA, resulting requirements on the control system and also the *maximum simplicity of the design* requirement, we chose to develop a system using two channels (redundant sensors and ECUs) and connect the control outputs with the Voter implemented on CPLD.

During the process of the design, development and testing of the control system we used Model Based Design approach including tools such as Simulink, StateFlow, SimMechanics. The designed and implement redundant control system was extensively tested using HIL simulation. Fig. 12 shows the experimental self-balancing vehicle in operation.

Presented structure of the fault tolerant control system might be, with appropriate modifications, used in wide range of applications in fields such as reliable measuring instruments development, automotive, aerospace and other.



Figure 12. Overall view of the self-balancing transporter

ACKNOWLEDGMENT

This work was supported by the European Commission within the FP7 project Efficient Systems and Propulsion for Small Aircraft "ESPOSA", contract No. ACPI-GA-2011-284859-ESPOSA, and by NETME CENTRE PLUS (LO1202) created with financial support from the Ministry of Education, Youth and Sports under the „National Sustainability Programme I“.

REFERENCES

- [1] D. Pavković, J. Deur, M. Jansz and N. Perić, "Adaptive Control of Automotive Electronic Throttle," Control Engineering Practice, 2006, 14, pp. 121 – 136.
- [2] R. Grepl and B. Lee, "Modeling, parameter estimation and nonlinear control of automotive electronic throttle using a Rapid-Control Prototyping technique," Int. J. of Automotive Technology, 2010, Volume 11, Number 4, pp. 601-61.
- [3] R. Grepl, "Adaptive Composite Control of ET using Local Learning Method," IEEE International Symposium on Industrial Electronics, Bari, Italy, 2010.
- [4] M. Muenchhof, M. Beck and R. Isermann, "Fault diagnosis and fault tolerance of drive systems - Status and research," Control Conference (ECC), 2009 European , pp.3464,3479, 23-26 Aug. 2009.
- [5] N. Wang and Y. Zhou, "Research on reliability of a hybrid Three-Redundant Electro-Mechanical Actuator," ICMA 2009, pp.1066,1070, 9-12 Aug. 2009.
- [6] Dong Huifen, Zhou Yuanjun and Shen Songhua, "The performance of a brushless DC motor control system with double channel," Electrical Machines and Systems, ICEMS 2005, vol.1, pp.376-379, 27-29 Sept. 2005.
- [7] J.-H. Jean and Chih-Kai Wang, "Design and implementation of a balancing controller for two-wheeled vehicles using a cost-effective MCU," Machine Learning and Cybernetics, 2009, vol.6, pp.3329-3334, 12-15 July 2009.
- [8] R. Grepl, "Balancing Wheeled Robot: Effective Modelling, Sensory Processing and Simplified Control," J. Engineering Mechanics, 2009, Vol. 16, No. 2, 141–154.
- [9] H. G. Nguyen, J. Morrell, K. Mullens, A. Burmeister, S. Miles, N. Farrington, K. Thomas, and D.W. Gagee, "Segway Robotic Mobility Platform," SPIE Proc. 5609: Mobile Robots XVII, Philadelphia, PA, October 27-28, 2004.
- [10] B. Allouche, L. Vermeiren, A. Dequidt and M. Dambrine, "Robust control of two-wheeled self-balanced transporter on sloping ground: A Takagi-Sugeno descriptor approach," Systems and Control (ICSC), 2013, pp.372-377, 29-31 Oct. 2013.
- [11] S. Colton, "The balance filter – A simple solution for Integrating Accelerometer and Gyroscope Measurements for a Balancing problem," report, web.mit.edu/scolton/www/filter.pdf
- [12] R. Isermann, "Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance," Springer, 2006.
- [13] R. Grepl, "Real-Time Control Prototyping in MATLAB/Simulink: review of tools for research and education in mechatronics," IEEE International Conference on Mechatronics (ICM 2011-13-15 April, 2011, Istanbul), 2011.
- [14] R. Grepl, F. Zouhar, J. Stepanek and P. Horak, "The Development of Self-balancing Vehicle: a Platform for Education in Mechatronics," Technical Computing Prague 2011.
- [15] Youngsong Lee and Woon-Sung Lee, "Hardware-in-the-loop Simulation for Electro-mechanical Brake," SICE-ICASE, pp.1513-1516, 18-21 Oct. 2006.

Actuator Fault Monitoring and Fault Tolerant Control in Distillation Columns

Sulaiman A. Lawal and Jie Zhang
School of Chemical Engineering and Advanced Materials
Newcastle University
Newcastle upon Tyne NE1 7RU, UK
s.a.lawal@newcastle.ac.uk

Abstract– This work presents, from a practical viewpoint, an investigation of real-time actuator fault detection, propagation and accommodation in distillation columns. Addressing faults in industrial processes, coupled with the growing demand for higher performance, improved safety and reliability necessitates implementation of alternative control strategies in the events of malfunctions in actuators, sensors and or other system components. This work uses multivariate statistical process monitoring based fault detection and diagnosis techniques employing principal components analysis (PCA) and dynamic PCA for prompt and effective faults detection and isolation. The work also investigates effects of disturbances on fault propagation and detection. Alternative control strategy is then implemented to accommodate the actuator faults upon their identification. Specifically, the reflux and vapour boil-up control strategy used for the distillation column during normal operation is switched to one point control of the more valued composition by utilising the remaining healthy actuator. The proposed approach was implemented on a simulated methanol-water separation column to assess its effectiveness.

Keywords: *Dynamic principal component analysis; fault detection and diagnosis; distillation column; fault tolerant controller; inferential control; principal component analysis.*

I. INTRODUCTION

Application of fault tolerant control systems (FTCS) in industrial processes offers high performance, improved safety, reliability and availability in the presence of faults in sensors, actuators and some other system components. FTCS is a closed-loop control system with automatic components containment capabilities and provides desirable performance on complex automated facilities whether faults are present or not. A requirement for the design of an FTCS is an effective fault detection and diagnosis (FDD) system to detect faults and provide information on the time of fault occurrence, locations and magnitudes of faults. Billions of dollars are lost in the industry every year due to low productivity, loss of operational hours, occupational injuries and illnesses resulting from major and common minor accidents occurring on a daily basis [1-3]. It is inevitable that some processing equipment including actuators, sensors and control systems will breakdown or malfunction at some point during their operational life span. So, it will be desirable to have a fault tolerant controller (FTC) that is able to accommodate those potential failures during

operation while still maintaining acceptable level of performance, albeit with some graceful degradation.

Distillation column is among the most common and energy intensive plant units. Though, it is a complex separation unit but in principle its control is relatively straightforward [4]. The dynamics and control of distillation column has been extensively studied because of its fundamental importance to the chemical and process industries [4,5]. However, from practical viewpoint its operation and control under faulty components such as actuators and sensors have not been widely reported. Owing to its importance in the industries, this paper focuses on the implementation of an actuator fault tolerant control strategy for a comprehensive nonlinear methanol-water separation column.

This paper presents a two-stage fault tolerant control system for a binary distillation column. First, the presence of fault is detected using principal component analysis (PCA) or dynamic PCA (DPCA). DPCA incorporates time-lagged process variables to properly capture the system dynamics and its effect on actuator fault propagation. Upon detection of a fault, contribution plots are used to isolate the fault. The second stage involves restructuring the control configuration to accommodate the detected faults. If the fault is an actuator fault, then two-point control strategy cannot be functional and has to be switched to one-point control. The most valued composition is then controlled directly using the remaining healthy actuator to limit the impact of actuator fault. If the detected fault is a composition sensor fault, then composition sensor feedback control involving the faulty composition sensor will not be functional and inferential control by-passing the faulty sensor can be implemented.

This paper is structured as follows. Section 2 discusses fault detection and diagnosis while Section 3 discusses the column under consideration. Fault detection and accommodation is presented in Section 4. Section 5 presents results and discussions while Section 6 presents some concluding remarks.

II. FAULT DETECTION AND DIAGNOSIS

FDD is a crucial component of an FTCS. Its effectiveness will to a large extent determine the applicability, effectiveness and overall functionality of the resulting FTCS. As it is the case with the conventional control systems, an understanding of the

faulty actuators and their effects would be required in either a mathematical or statistical form to enable the design of suitable FTCS for the controlled process. FDD has been approached from a wide range of perspectives, which includes but not limited to model-based and data-based approaches. The focus of this paper is on data-based fault detection and diagnosis using PCA and DPCA. Readers interested in other approaches should consult [11-14].

PCA is a static multivariate statistical projection technique that is based on orthogonal decomposition of the covariance matrix of the process variables along direction that explains the maximum variation of the data. Its main function is finding factors that have a much lower dimension than the original data set which can properly describe the major trend in the original data set. Dynamic PCA works exactly the same way, but it incorporates time-lagged measurements to model the dynamic correlation behaviour of the system for effective fault propagation analysis. The PCA method can be briefly summarised as follows: let X be an $n \times p$ matrix of the scaled measurements of n samples and p variables with covariance matrix Σ . From matrix algebra, Σ may be reduced to a diagonal matrix L by a particular orthonormal $p \times p$ matrix U , i.e.,

$$\Sigma = ULU^T \quad (1)$$

where columns of U are the principal component loading vectors and the diagonal elements of L are the ordered eigenvalues of Σ which defines the amount of variance explained by the corresponding eigenvector. Then, the principal component transformation is given as:

$$T = XU \quad \text{or} \quad t_i = Xu_i \quad (2)$$

Equivalently, X is decomposed by PCA as:

$$X = TU^T = \sum_{i=1}^p t_i u'_i \quad (3)$$

The $n \times p$ matrix $T = (t_1, t_2, \dots, t_p)$ contains the so-called *principal component scores* which are linear combinations of all p observations. Generally, the first few PCs will capture the most variation in the original data if the variables are correlated. Typically the first “ a ” principal components ($a < p$) can be used to represent the majority of data variation:

$$X = t_1 u'_1 + \dots + t_a u'_a + E = \sum_{i=1}^a t_i u'_i + E \quad (4)$$

where E is the resulting residual term due to ignoring the rest of the principal components.

The statistical metrics against which the new measurements will be checked for any possible occurrence of faults include the Hotelling’s T^2 and the squared prediction error (SPE). The Hotelling’s T^2 monitoring statistics is obtained as follows.

$$T_i^2 = \sum_{j=1}^a \frac{\theta_{ij}^2}{\lambda_j} \quad (5)$$

where T_i^2 is the Hotelling’s T^2 value for sample i , θ_{ij} is the i^{th} element of principal component j , λ_j is the

eigenvalue corresponding to principal component j and a is the number of principal components retained.

SPE is simply square of the difference between the original scaled data and the estimates from the PCA model. When the process is in normal operation, both SPE and T^2 should be small and within their control limits. When a fault presents in the monitored process, the fault will cause some variables having larger than normal magnitudes (large T^2 values) and/or change the variable correlations leading to large SPE values. The control limits for SPE and T^2 are given by (6) and (7) respectively.

$$\begin{cases} SPE_{lim} = \Theta_1 \left[\frac{c_\alpha h_0 \sqrt{2\Theta_2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1^2} \right]^{\frac{1}{h_0}} \\ \Theta_i = \sum_{j=a+1}^p \lambda_j^i \\ h_0 = 1 - \frac{2\Theta_1 \Theta_3}{3\Theta_2^2} \end{cases} \quad (6)$$

$$T_{lim}^2 = \frac{a(n-1)}{(n-a)} F_{a,n-a,\alpha} \quad (7)$$

In (6) and (7), c_α is the value for normal distribution at 99% confidence level and $F_{a,n-a,\alpha}$ is the F distribution with appropriate degrees of freedom and confidence level.

III. A METHANOL-WATER SEPARATION COLUMN

The distillation column studied in this paper is a comprehensive nonlinear simulation of a methanol-water separation column. A nonlinear stage-by-stage dynamic model has been developed using mass and energy balances. The simulation has been validated against pilot plant test and is well known for its use in control system performance studies [8-10]. The following assumptions are imposed on the column: negligible vapour hold-up, perfect mixing in each stage, and constant liquid hold-up. Table 1 presents the steady-state conditions for the column.

TABLE 1. NOMINAL COLUMN OPERATING DATA

Column Parameters	Values
No of theoretical stages	10
Feed tray	5
Feed composition (Z)	50% methanol
Feed flow rate (F)	18.23 g/s
Top composition (Y_D)	0.95 (weight fraction)
Bottom composition (X_B)	0.05 (weight fraction)
Top product flow rate (D)	9.13 g/s
Bottom product flow rate (B)	9.1 g/s
Reflux flow rate (L)	10.11 g/s
Steam flow rate (V)	13.81 g/s

The column was simulated in MATLAB with 30 second sampling time using the LV control strategy. The top composition (Y_D) is controlled by the reflux flow rate

(L) and the bottom composition (X_B) by the steam flow rate (V) to the reboiler. Levels in the condenser and the reboiler are controlled by the top and bottom flow rates respectively. The disturbances in the system are feed flow rate and feed compositions. The top and bottom product compositions are measured by composition analyzers with 10 sampling time delay (5 minutes).

Low and high magnitude faults were introduced into the system at different times by restricting the flow of reflux and steam rates to represent stuck valves, thereby acting as actuator faults as shown in Table 2. Four low magnitude actuator faults (F1, F3, F6, and F7) were investigated, with values of the manipulated variables held close to their respective steady state values. The first 2 low magnitude faults (F1 and F2) were investigated for low magnitude fault detectability while the last 2 low magnitude faults (F6 and F7) were introduced to investigate effects of disturbances on low magnitude faults propagation and detectability. Also, two high magnitude actuator faults (F2 and F4) and a combination of the two high magnitude actuator faults (F5) were considered. Similarly, two sensor faults, top composition (F8) and bottom composition (F9) sensor faults, were investigated to assess the ability of the DPCA FDD technique to isolate different faults using contribution plots. The fault cases were each simulated for 750 minutes to collect 1500 samples each for the 9 fault cases.

TABLE 2. FAULT LIST

Fault	Fault description
F1	Reflux valve stuck btw 7-10 g/s after sample 750
F2	Reflux valve stuck btw 5- 8 g/s after sample 750
F3	Steam valve stuck btw 10-14 g/s after sample 750
F4	Steam valve stuck btw 10-13 g/s after sample 750
F5	Reflux and steam valves stuck @ 8 g/s and 13 g/s after samples 750 and 1150 respectively
F6	F1 repeated with Feed flow rate disturbance introduced after sample 900
F7	F3 repeated with Feed flow rate disturbance introduced after sample 900
F8	Top composition sensor fault with sensor value set at 0.75 after sample 750 (static)
F9	Bottom composition sensor fault with sensor value set at 0.03 after sample 750 (static)

There were a total of 14 variables monitored: the top and bottom product compositions, the controller outputs for the reflux flow and the steam flow, and the ten tray temperatures. Random noise with zero means and 0.15 and 0.001 standard deviations were added to the ten tray temperatures and the top and bottom compositions respectively, to represent true measurements of the data collected. Fig. 1 presents the top and bottom compositions, their respective control outputs and the ten tray temperatures for normal operating conditions.

IV. FAULT DETECTION AND ACCOMMODATION

A PCA model was developed using the first 1600 samples out of the 2,600 samples collected under normal process conditions while the last 1000 samples were used as testing data. The data was scaled to zero mean and unit variance. Also, a DPCA model was developed using the same number of samples for training and testing data with one sampling time lag. By examining the accumulated contributions of principal components (PCs), 4 PCs were kept for PCA and DPCA models and they account for 84.27% and 82.17% variations respectively. The obtained PCA and DPCA models for the fault free system are then used for process monitoring. Fig. 2 presents the T^2 and SPE plots for the normal process for both the PCA and the DPCA models. It can be observed from Fig. 2 that the normal process operation data are both classified as being normal by the PCA and DPCA models. DPCA model performs slightly better in terms of control limit violations, hence only DPCA will be used for fault detection and diagnosis.

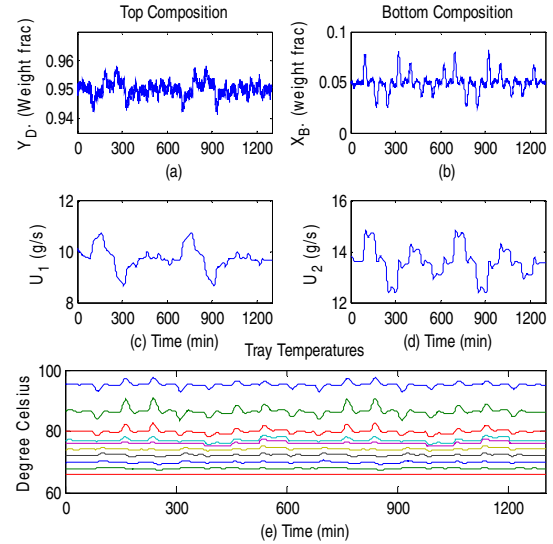


Figure 1. (a) Top composition; (b) Bottom composition; (c) Controller output for top comp.; (d) Controller output for bottom comp.; (e) Tray temperatures

DPCA model is applied to the nine faulty data sets to detect the faults. When the monitoring statistics, SPE or T^2 , violate their control limits, we allow four sampling time to elapse before declaring detection of a fault in the system. This is to reduce the occurrence of false alarms. Once the presence of a fault is detected, further fault identification analysis is carried out through contribution plots to identify variables that are responsible for the faults, and ultimately isolate the faults. If the detected fault is an actuator fault, then depending on the composition that is deemed more valuable, the control system is restructured by switching to one-point control using the only remaining healthy actuator. If the detected fault is a composition sensor fault, then inferential control will be implemented to remove the need of the faulty composition sensor.

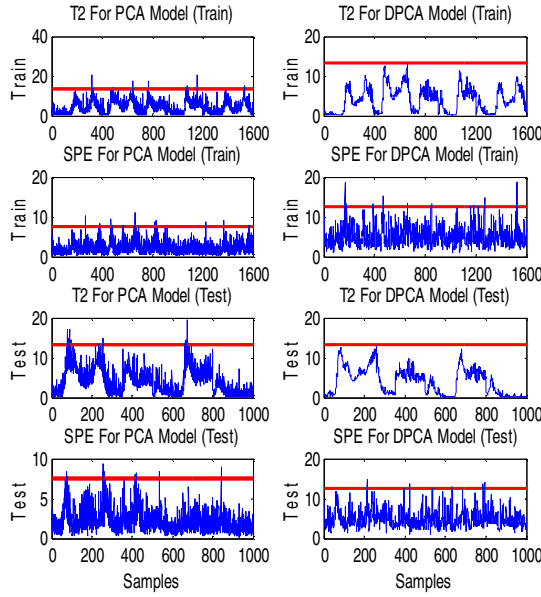


Figure 2. PCA and DPCA monitoring performance.

V. RESULTS AND DISCUSSIONS

Among the considered faults, F1 and F3 are low magnitude reflux flow actuator and steam flow actuator faults respectively, with disturbances introduced before and after faults introduction. F6 and F7 are F1 and F3 repeated, but with more profound disturbances. The intention was to investigate the effect of disturbances on fault propagation and detection, and to also check whether the alternative control strategy proposed will be able to handle the faults plus the disturbances. Figs. 3 and 4 show the T^2 and SPE plots for faults F1 to F5 and F6 to F9 respectively. Fig. 5 shows the contribution plots for F2, F4 and F5 while Fig. 6 presents that of F7, F8 and F9. Figs. 5 and 6 present the excess contributions of each variable to the larger than normal value of T^2 at the point of fault declaration. By excess contributions, we mean the difference between each variable contribution at the point of fault declaration and the variable average contribution to T^2 values under fault free conditions.

From the analysis of the T^2 and SPE plots shown in Figs. 3 and 4, the monitoring statistics for faults F2, F4, F5, F7, F8 and F9 exceeded their control limit at different times so these faults were detected. We allowed four sampling time (2 mins) to elapse after the statistical indices went over their limits before declaring faults in order to minimize possible declaration of false alarm. Faults F2 and F5 were detected 13 sampling times (6 mins 30 sec.) after fault introduction, on sample 763, while it took 115 sampling time (approximately 58 mins), on sample 865 for fault effect to manifest in F4 as presented in Fig. 3. Fault 7 (F7) was detected on sample 969, approximately 110 minutes after it was introduced on sample 750 as shown in Fig. 4. Note that faults F3 and F7 are exactly the same with the exception of disturbances introduced after the faults to investigate the effects of the disturbances on the faults propagation and detectability. A 10 percent increase in feed composition

disturbance was introduced on sample 900 in the case of F3 while the same magnitude of disturbance in feed flow rate was introduced in the case of F7, also on sample 900.

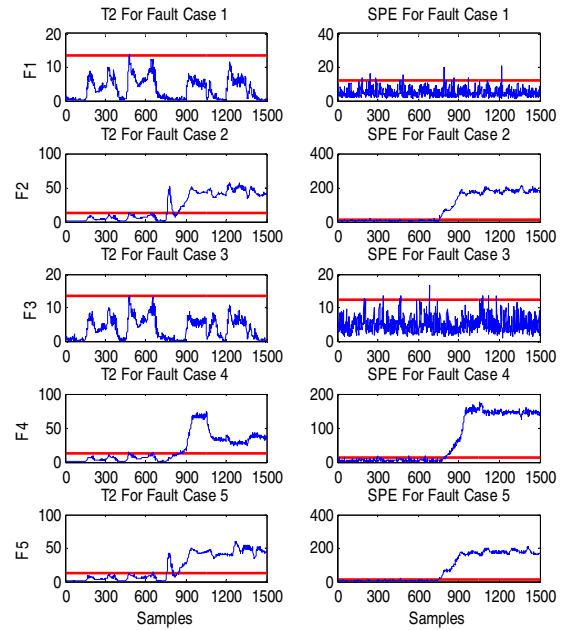


Figure 3. T^2 and SPE plots for fault cases F1 to F5

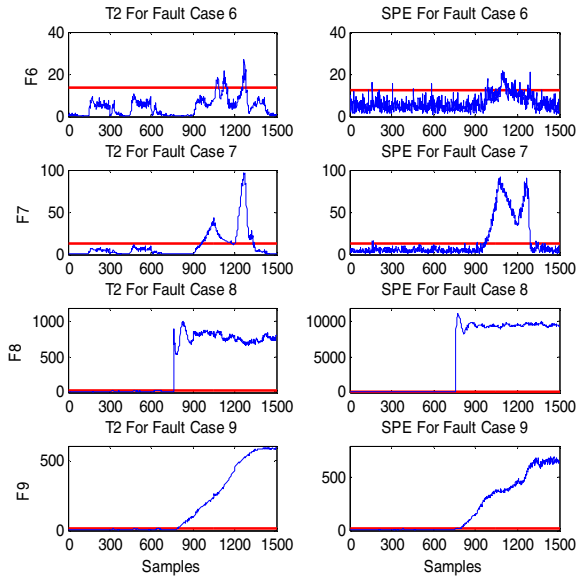


Figure 4. T^2 and SPE plots for fault cases F6 to F9

Basically, disturbances do affect faults propagation and detection in the column, and could amplify a rather minor undetected fault as shown in the case of F7 which was detected 69 sampling times (approx. 35 minutes) after the disturbance was introduced. Further analyses were conducted to identify the faults using contribution plots upon declaration of a fault. The T^2 contribution plots gave a more consistent indication of the variables responsible for the faults; hence only T^2 contribution plots were used for fault identification. In the case of F2 and F5 where reflux actuation faults were identified, the contribution

plot as shown in Fig. 5 identified the top composition and the top four tray temperatures (variables 1, 11, 12, 13 and 14, duplicated for the next 14 variables representing the previous sampling time) as the major contributors to the out-of-control situation.

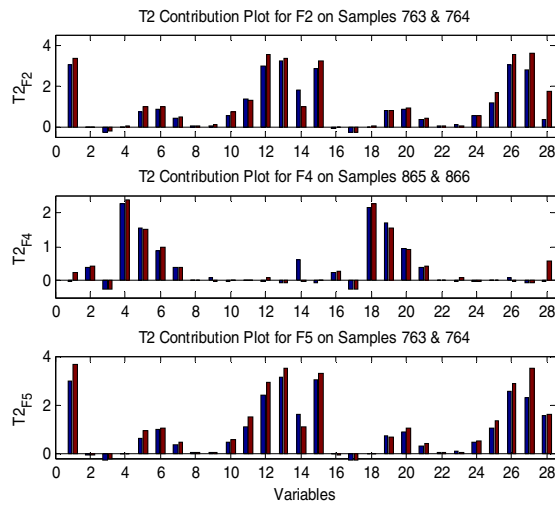


Figure 5. T^2 Contribution Plots for F2, F4 and F5

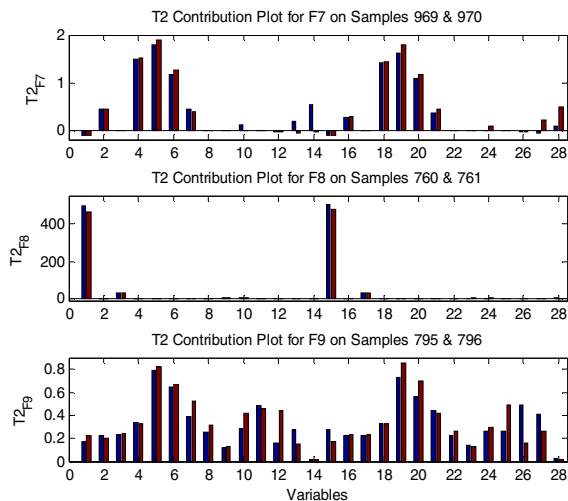


Figure 6. T^2 Contribution Plots for F7, F8 and F9.

Analysis of the T^2 contribution plots presented in Figs. 5 and 6 combined with the process knowledge aided the fault identification. For instance, when the reflux actuator fault occurred (stuck reflux valve), and after it has been detected, the contribution plot isolates variables indicative of the fault. The detected reflux actuator fault with reduced reflux flow caused the top tray temperature measurements to rise by certain percentage which ultimately led to the top composition drifting out of control. Hence, the contributions of these variables (top composition and the top tray temperatures) to the T^2 monitoring statistics increased significantly as presented in Fig. 5. The rise in the top tray temperature measurements as a consequence of reduced reflux flow is peculiar to the reflux actuation fault, which aided its isolation. Similarly, observing contribution plots for F4 and F7 as presented in Figs. 5 and 6; when steam actuator

fault occurred the bottom composition drifted out of control which also affected the steam controller output and the bottom tray temperatures. These effects manifest in the larger than average contributions of these variables to the T^2 values at the point of fault declaration and beyond. This was the pattern exploited in the faults identifications as different faults show different variable contributions to the T^2 values after occurrence of a fault.

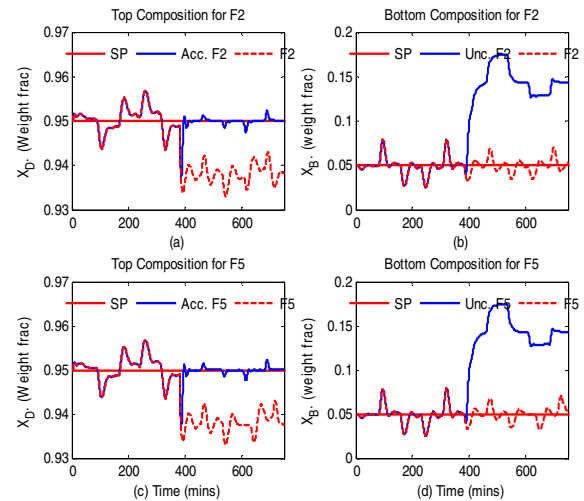


Figure 7. Responses of the top and bottom composition to F2 and F5 reflux actuator faults accommodation.

As mentioned in Table 2, F1, F2 and F6 are all reflux actuator faults of different magnitudes; while F3, F4 and F7 are steam actuator faults, also of different magnitudes. F5 is a combination of reflux and steam actuator faults, but with reflux actuator fault occurring first. Faults 8 and 9 (F8 and F9) are top and bottom compositions sensor faults which were also detected and identified respectively. Contribution plot for the top composition sensor fault (F8) shows the top composition (the sensor output) as the only variable responsible for the fault while the bottom composition sensor fault (F9) indicates that all the variables are responsible for the fault as presented in Fig. 6. As can be observed from Figs. 5 and 6, the faults show different variable contribution patterns which aided their isolation. There were no fault detected in F1, F3 and F6 due to the fact that only small changes were made to the values of the two actuators which were close to the nominal values of the two manipulated variables as shown in Table 1 under reflux and steam flow rates. The resulting values for the process variables were within normal conditions. Fault 6 (F6), a rather minor undetected fault in F1, was affected by the amplifying effect of the disturbance (increased in feed flow rate after sample 900) on its propagation which moved it to marginal stability.

Clearly, the conventional LV control strategy used for the column normal operation could not accommodate the faults. Hence, the control strategy in the column is restructured by switching to one-point control strategy where the only remaining healthy actuator, steam flow rate actuation in the case of F2 and F5 and reflux flow rate actuation in the case of F4 and F7 were used to

accommodate the faults and maintain the more valuable outputs of the two compositions within acceptable range while the other is uncontrolled. Steam flow actuation was immediately restructured and implemented to tolerate reflux valve faults, F2 and F5 by manipulating the steam flow rate to directly maintain the top composition at set point thereby tolerating reflux valve actuation faults in F2 and F5 as presented in Fig. 7. In the same vain, upon detection of steam flow actuation faults, F4 and F7, the column control structure was immediately switched to reflux valve one-point control by manipulating reflux flow rate to directly maintain the bottom composition at set point, thereby tolerating the steam flow actuation faults F4 and F7 as shown in Fig. 8. It is worth mentioning at this point that, the fault accommodation approach proposed in this work is sub-optimal as it is practically impossible to use one manipulated variable to maintain both top and bottom compositions at set points. The sub-optimal fault accommodation approach provides desirable performance and will be far more acceptable than shut-down. SP, Unc. and Acc. were used in Figs. 7 and 8 to represent set point, uncontrolled fault and accommodated fault respectively.

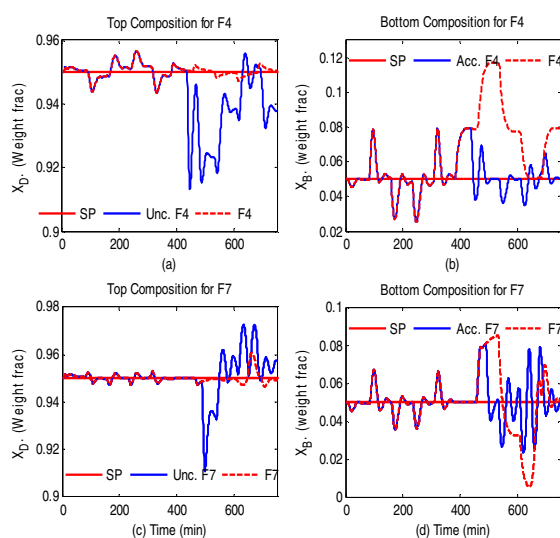


Figure 8. Responses of the top and bottom composition to F4 and F7 steam actuator faults accommodation

The effects of disturbances, feed flow rates and the feed compositions after the faults were well compensated for by the fault tolerating control approach as can be observed in Figs. 7 and 8. Sensor faults accommodation strategy is being investigated, and its detection and identification is only included in this work to demonstrate the effectiveness of DPCA – based FDD approach.

VI. CONCLUSIONS

This work investigates propagation, identification and accommodation of faults, in particular actuator faults, in distillation columns using comprehensive nonlinear simulation of a methanol-water separation column as an example. First, DPCA model is developed from normal process operation data. The DPCA effectively detected the faults and further diagnosis revealed the variables

responsible for different faults through contribution analysis. When the detected fault is identified as an actuator fault, one-point control strategy is implemented to directly control the more valued product composition leaving the other uncontrolled. The effect of disturbances on actuator fault propagation was also investigated. When the detected fault is identified as a composition sensor fault, then inferential control by-passing the faulty composition sensor needs to be implemented. This is under investigation and will be reported in the future. The effectiveness of the approach was demonstrated by the simulation results. Application of the approach on a more complex system, for instance, a crude distillation with several actuator and sensor faults is currently being investigated.

REFERENCES

- [1] Bureau of Labor Statistics, Occupational injuries and illnesses in the United States by Industry. Washington, DC: Government Printing Office, 1998.
- [2] McGraw-Hill Economics, Survey of investment in employee safety and health NY: McGraw-Hill Publishing Co., 1985.
- [3] National Safety Council, Injury facts 1999 Edition, Chicago: National Safety Council, 1999.
- [4] J. Love, Process automation handbook: a guide to theory and practice, Springer-Verlag, London, 2007.
- [5] T. E. Marlin, Process control: designing processes and control systems for dynamic performance, 2nd edn. McGraw Hill, 2000.
- [6] W. L. Luyben, Practical distillation control. Van Nostrand, 1992.
- [7] P. B. Deshpande, Distillation dynamics and control. ISA, Carolina, 1985.
- [8] J. Zhang and R. Agustriyanto, "Multivariate inferential feed-forward control," *Ind. Eng. Chem. Res.*, 2003, 42, 4186 – 4197.
- [9] M. T. Tham, F. Vagi, A. J. Morris and R. K. Wood, Online "Multivariable Adaptive Control of a Binary Distillation Column," *Can. J. Chem. Eng.* 1991, 69, 997-1009.
- [10] M. T. Tham, F. Vagi, A. J. Morris and R. K. Wood, "Multivariable and Multirate Self-Tuning Controls – A Distillation Column Case Study," *IEE Proc. Pt. D, Control Theory Appl.* 1991, 138, 9-24.
- [11] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A Review of Process Fault Detection and Diagnosis. Part I. Quantitative Model-Based Methods," *Computers and Chemical Engineering*, 2003a, 27(3), 293 – 311.
- [12] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri, "A Review of Process Fault Detection and Diagnosis. Part II. Qualitative Models and Search Strategies," *Computers and Chemical Engineering*, 2003b, 27(3), 313 – 326.
- [13] V. Venkatasubramanian, R. Rengaswamy, R. Yin, and S. N. Kavuri, "A Review of Process Fault Detection and Diagnosis. Part III. Process History Based Methods," *Computers and Chemical Engineering*, 2003c, 27(3), 327 – 346.
- [14] J. Zhang, "Improved on-line process faults diagnosis through information fusion in multiple neural networks," *Computers and Chem. Eng.*, 2006, 30, 558 – 571.

Fault Detection and Diagnosis for Operational Control Systems

Ning Sheng

Northeastern University

the State Key Laboratory of

Synthetical Automation for Process Industries

Shenyang, Liaoning, P R China 110819

Email: shengning2008@126.com

Hong Wang

University of Manchester

Control System Center

Manchester, U.K. M601QD

Email: hong.wang@manchester.ac.uk

Abstract—Operational control for complex industrial processes consists of two layers, namely the loop control layer and the operational layer. The former realizes the required loop control for each production unit and the later determines the best set-points to the control loops in the loop control layer. This paper formulates a general framework for the fault detection and diagnosis for such operational control systems, where it is assumed that the faults only take place in the loop control layers. For this purpose, a novel model is established for the representation of the fault detection and diagnosis. This is then followed by the development of fault detection and diagnosis algorithm, where fault diagnosis is realized using an adaptive fault diagnostic observer. A Lyapunov function is used to formulate the fault diagnosis algorithms where the system model uncertainties are also considered. Simulation has been carried out to show the effectiveness of the proposed method.

I. INTRODUCTION

Fault detection and diagnosis have become an important area of research and have been an integrated part of modern control systems for many industrial processes [1-2]. The developed methods include observer based approaches [3], system identification based methods and Principal component analysis (PCA) based techniques [4]. In recent years, fault detection and diagnosis for nonlinear systems has also become an increasingly important area, and examples include extended Kalman filtering based methods [5] and neural networks and fuzzy logic based approaches [6-8]. However, these existing methods are largely focussed on single control loops where actuator, systems and sensor faults are considered under an open loop structure. This differs from the situation for almost all the control systems in practice where closed loop operation has been widely employed. In this context, fault detection and diagnosis under closed loop system structure has also been the subject of study in recent years [9-10]. However, direct extension of the fault detection and diagnosis for open loop systems to closed loop systems is not straightforward [11-12]. Therefore, fault detection and diagnosis for closed loop systems is still a challenging issue at present. This is particularly true for complex industrial processes which operate in a multiple layer mode including loop control layer, operation control layer and operational management layer. Indeed, fault detection and diagnosis for operational layer is of particular important as this layer looks after the real-time plant-wide operation [13-18]. In this paper, a novel fault detection and diagnosis method will be proposed for the operational

layer which consists of loop control layer that employs many closed loop control systems.

II. OPERATIONAL CONTROL AND ITS MODEL FOR FAULT DETECTION AND DIAGNOSIS

Most of the complex industrial processes operate in a kind of 2D mode as shown in Fig. 1, where Most of the complex industrial processes operate in a kind of 2D mode as shown in Fig. 1, where Q_k are performance indexes that stand for the final product quality, consumption and costs of industrial processes. In Fig.1, subscripts min and max represent the range of the indexes which can be acquired a prior. Horizontally, there are normally a number of production units that are largely linked in a series to perform the required production along production lines. Vertically, there are a number of operation layers consisting of either one or several distributed control systems (DCS), planning and scheduling (PS) units and even manufacturing execution systems (MES). To simplify the representation, the operational optimal control can be illustrated into two layered control systems as shown in Fig.2, where a performance function J (e.g., energy or product quality) is required to be minimized via the selection of set-points to be applied to loop control layers. By minimizing this performance function, a number of optimal set-points can thus be obtained and used to the control system in the loop control layer. Upon receiving these set-points, the control loops in the loop control layer would manipulate the control variables in a closed loop manner so that the control outputs can follow the set-points.

The aims of process control systems have two objectives. On one hand, a proper controller needs to be selected to ensure the stability of control loops in the loop control layer stable. On the other hand, the system outputs $y_i (i = 1, \dots, n)$ should be made to track the optimized set-points $r_i^* (i = 1, \dots, n)$. In fact, the set-points $r_i^* (i = 1, \dots, n)$ are ideal values of the process control and the relationships between y_i (the system outputs) and r_i (ideal values) can be summarized as shown in Fig.2. When the operation condition varies, these set-points need to be tuned correspondingly.

III. MODEL REPRESENTATION

In this paper we will consider linear systems only for the control loops in the loop control layer. In this context, it can

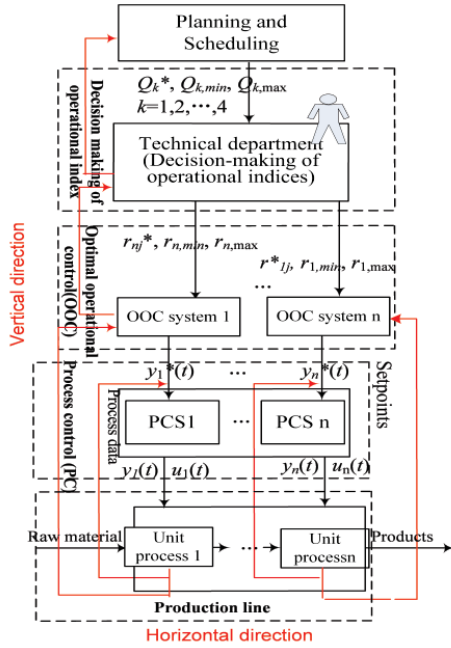


Fig. 1. The 2D operational mode for industrial processes.

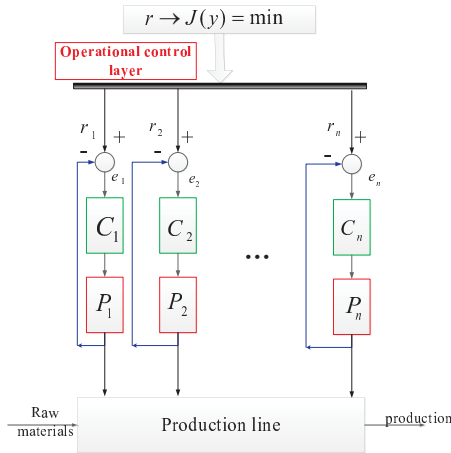


Fig. 2. Operational optimal control scheme for simple illustration.

be seen that for each control loops, its dynamic model can be expressed as

$$\begin{aligned} \dot{x}_i &= a_i x_i + b_i u_i \\ y_i &= c_i x_i, (i = 1, \dots, n) \end{aligned} \quad (1)$$

where $x_i \in R^{n_i}$ is the state vector of the i -th control loop, $u_i \in R^{m_i}$ is the control input of the i -th control loop, $y_i \in R^{m_i}$ is the measurable output of the i -th control loop, a_i, b_i and c_i are known parameter matrices, m_i and n_i are the known dimension of the system.

For the system of each loop, the tracking error can be defined as

$$e_i = r_i - y_i \quad (2)$$

The linear dynamic feedback controller of each loop system can be defined as follows

$$\begin{aligned} \dot{v}_i &= f_i v_i + g_i e_i + w_i y_i \\ u_i &= h_i v_i + D_i e_i \end{aligned} \quad (3)$$

where $v_i \in R^{p_i}$ is the state vector of the controller, f_i, g_i, h_i are controller parameter which are to be designed to ensure the system stable. In this paper we assume that such controller (3) has been designed already for each loop.

When there is no fault in the loop control layer, these tracking errors would converge to zero, and the actual output y_i would follow the pre-specified optimal set-points $r_i^*(i = 1, \dots, n)$. In this case the performance function $J(y_i) = J(r_i^*) = \min$ is at its minimum value and the whole system operates at its optimal status. However, when fault occurs in the loop control layer, the related loop tracking error $e_i \neq 0$ and as a result the actual performance function would satisfy $J(y_i) > J(r_i^*)$. In this case the set-points need to be re-selected so that the actual output y_i can still be made to follow the pre-specified optimal set-points $r_i^*(i = 1, \dots, n)$. As such, the new set-points applied to the control loops in the loop control layer should read

$$r_i = r_i^* + \Delta r_i (i = 1, \dots, n)$$

where Δr_i is the incremental tuned value of the set-points. For operational control, Δr_i is selected so that the actual output y_i can converge to its optimal set-point $r_i^*(i = 1, \dots, n)$. In this context, Δr_i can be regarded as the inputs to the operational control layer. Also, when fault occurs, equation (1) can be expressed as

$$\dot{x}_i = a_i x_i + b_i u_i + \theta_i d_i, \quad i = 1, \dots, n \quad (4)$$

where θ_i is the known parameter matrix and $d_i \in R^{q_i}$ is the fault vector of the i -th control loop that need to be detected and diagnosed, and is the dimension of the fault for the i -th control loop.

For each loop, define the closed loop state vector as

$$z_i = \begin{bmatrix} x_i \\ v_i \end{bmatrix} \in R^{n_i+p_i} \quad (5)$$

Then the following model representation for the fault detection and diagnosis of operational control system shown in Fig. 2 can be obtained to give

$$\begin{aligned} \dot{z}_i &= A_i z_i + B_i r_i + E_i d_i \\ e_i &= r_i - C_i z_i \\ i &= 1, 2, \dots, n \end{aligned} \quad (6)$$

where it has been denoted that

$$\begin{aligned} A_i &= \begin{bmatrix} a_i - b_i D_i c_i & b_i h_i \\ w_i c_i - g_i c_i & f_i \end{bmatrix} \in R^{(n_i+p_i) \times (n_i+p_i)}; B_i = \\ &= \begin{bmatrix} b_i D_i \\ g_i \end{bmatrix} \in R^{(n_i+p_i) \times m_i}; C_i = \begin{bmatrix} c_i \\ 0 \end{bmatrix} \in R^{m_i \times (n_i+p_i)} \\ E_i &= \begin{bmatrix} \theta_i \\ 0 \end{bmatrix} \in R^{(n_i+p_i) \times q_i} \end{aligned}$$

If $d_i = 0$, then one can simply choose $\Delta r_i = 0$. This would ensure that $e_i = r_i - y_i = 0$ and the whole operational control

system works well at its optimal status. However, if $d_i \neq 0$, one needs to choose a suitable Δr_i so as to ensure $y_i \rightarrow r_i^*$, which means that the performance function $J(y) = \min$.

The closed-loop system for the whole operational control can be further expressed as the following matrix form

$$\begin{aligned}\dot{z} &= Az + Br^* + Ed + B\Delta r \\ e &= Cz + \Delta r + r^*\end{aligned}\quad (7)$$

where

$$\begin{aligned}z &= \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} \in R^n, A = \text{diag}(A_i) \in R^{n \times n}, \\ B &= \text{diag}(B_i) \in R^{n \times m}, E = \text{diag}(E_i) \in R^{n \times q}, C = \\ &-\text{diag}(C_i) \in R^{m \times n}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in R^m, r^* = \begin{bmatrix} r_1^* \\ \vdots \\ r_n^* \end{bmatrix} \in \\ R^m, \Delta r &= \begin{bmatrix} \Delta r_1 \\ \vdots \\ \Delta r_n \end{bmatrix} \in R^m, r = r^* + \Delta r, d = \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} \in R^q, \\ n &= \sum_{i=1}^n n_i, m = \sum_{i=1}^n m_i, q = \sum_{i=1}^n q_i.\end{aligned}$$

In this regard, the purpose of fault detection and diagnosis is to use Δr and e to detect and diagnose d .

IV. FAULT DETECTION AND DIAGNOSIS ALGORITHM

A. Fault Detection

To detect the fault, the following detection observer is constructed [3]

$$\begin{aligned}\dot{\hat{z}} &= A\hat{z} + Br^* + B\Delta r + L(\hat{e} - e) \\ \hat{e} &= C\hat{z} + \Delta r + r^*\end{aligned}\quad (8)$$

where $\hat{z} \in \mathbb{R}^n$ is the state vector of the detection observer and $\hat{e} \in \mathbb{R}^m$ is the output vector of the observer. Since it has been assumed that the pair (A, C) is observable, L is a selected gain matrix such that $(A + LC)$ is a stable matrix. Define

$$\begin{aligned}\varepsilon &= \hat{z} - z \\ \varepsilon_0 &= \hat{e} - e = C\varepsilon\end{aligned}$$

Then the observation error equation can be obtained as

$$\begin{aligned}\dot{\varepsilon} &= \dot{\hat{z}} - \dot{z} \\ &= A(\hat{z} - z) - Ed + LC(\hat{z} - z) \\ &= (A + LC)\varepsilon - Ed\end{aligned}\quad (9)$$

As a result, the fault detection can be readily carried out as follows

$$\begin{aligned}\|\varepsilon_0(t)\| &= \|C\varepsilon(t)\| < \lambda; \text{no fault occurs} \\ \|\varepsilon_0(t_m)\| &= \|C\varepsilon(t_m)\| \geq \lambda; \text{fault has occurred}\end{aligned}\quad (10)$$

where λ is a pre-specified threshold, which can be selected as small as possible, t_m is the time when a fault occurs. If no

fault occurs, $d = 0$, the second term in (9) vanished, Since the system is stable, $\lim_{t \rightarrow \infty} \varepsilon(t) = 0$ and $\lim_{t \rightarrow \infty} \varepsilon_0(t) = 0$; If fault has occurred, $d \neq 0$, from (9), it can be seen that $\lim_{t \rightarrow \infty} \varepsilon(t) \neq 0$ and $\lim_{t \rightarrow \infty} \varepsilon_0(t) \neq 0$.

B. Fault Diagnosis

Once the fault is detected, fault diagnosis can be carried out so that a direct estimation to d can be obtained. For this purpose, denote the following observer

$$\dot{\hat{z}} = A\hat{z} + Br^* + E\hat{d} + B\Delta r + LC(\hat{z} - z) \quad (11)$$

where $\hat{e} = C\hat{z} + \Delta r + r^*$. To formulate the required fault diagnosis algorithm, we denote

$$\varepsilon = \hat{z} - z \quad (12)$$

Then it can be formulated that

$$\begin{aligned}\dot{\varepsilon} &= \dot{\hat{z}} - \dot{z} \\ &= (A + LC)\varepsilon + E(\hat{d} - d)\end{aligned}\quad (13)$$

Assuming that d is an unknown constant, then by choosing the following Lyapunov function

$$V = \varepsilon^T P \varepsilon + (\hat{d} - d)^T \Gamma^{-1} (\hat{d} - d) \quad (14)$$

the derivative of the Lyapunov function is given by

$$\begin{aligned}\dot{V} &= \varepsilon^T ((A + LC)^T P + P(A + LC)) \varepsilon \\ &\quad + 2(\hat{d} - d)^T E^T P \varepsilon \\ &\quad + 2(\hat{d} - d)^T \Gamma^{-1} \dot{\hat{d}}\end{aligned}\quad (15)$$

Let

$$(A + LC)^T P + P(A + LC) = -Q \quad (16)$$

$$C = -E^T P \quad (17)$$

Then the adaptive diagnostic law can be formulated using [3] to give

$$\begin{aligned}\dot{\hat{d}} &= -\Gamma E^T P \varepsilon = \Gamma C \varepsilon = \Gamma C(\hat{z} - z) \\ &= \Gamma C\hat{z} - \Gamma(r^* + \Delta r - e)\end{aligned}\quad (18)$$

It can be shown that when (18) is used, the following inequality should hold

$$\dot{V} = -\varepsilon^T Q \varepsilon \leq 0 \quad (19)$$

This indicates that if the error is not zero the Lyapunov function will continue to strictly decrease. As a result, the above diagnostic law can ensure that

$$\lim_{t \rightarrow \infty} \varepsilon(t) = 0$$

V. FAULT DETECTION AND DIAGNOSIS IN PRESENCE OF A MODEL UNCERTAINTY

In real process operation, there are always some uncertainties in the system model. Therefore in this section we will consider how the above adaptive diagnostic algorithm can be modified to copy with model uncertainties. In this context, the following operational control model can be written

$$\begin{aligned} \dot{z} &= Az + B(r^* + \Delta r) + Ed + \Delta(z, \Delta r) \\ e &= \begin{bmatrix} r_1^* + \Delta r - y_1 \\ r_2^* + \Delta r - y_2 \\ \vdots \\ r_n^* + \Delta r - y_n \end{bmatrix} \end{aligned} \quad (20)$$

where Δ is the model uncertainties for the system in Fig. 2, and is assumed to satisfy $\|\Delta(z, \Delta r)\| \leq M$ with a known upper bound M .

In this case one can still design a fault detection algorithm similar to 7. Since it has been assumed that the pair (A, C) is observable, L is again a selected gain matrix such that $(A - LC)$ is a stable matrix.

Define

$$\begin{aligned} \varepsilon &= \hat{z} - z \\ \varepsilon_0 &= \hat{e} - e = C\varepsilon \end{aligned} \quad (21)$$

Then the observation error equation can be expressed again as

$$\begin{aligned} \dot{\varepsilon} &= \dot{\hat{z}} - \dot{z} \\ &= (A + LC)\varepsilon - Ed - \Delta \end{aligned} \quad (22)$$

If no fault occurs, $d = 0$, the second term in (22) vanishes, this means that the observation error satisfies

$$\begin{aligned} \dot{\varepsilon} &= (A + LC)\varepsilon - \Delta \\ \varepsilon_0 &= C\varepsilon \\ \|\varepsilon_0\| &\leq \max_{\omega \geq 0} \|C(j\omega I - A - LC)^{-1}\| \end{aligned} \quad (23)$$

Since $\|\Delta(z, \Delta r)\| \leq M$, it can be further obtained that [19-20]. This can be calculated because all the matrices involved in (23) are known. Therefore, the fault detection can be readily carried out as follows

$$\begin{aligned} \|\varepsilon_0(t)\| &= \|\hat{e} - e\| \leq \lambda; \text{ no fault occurs} \\ \|\varepsilon_0(t_m)\| &= \|\hat{e} - e\| > \lambda; \text{ fault has occurred} \end{aligned} \quad (24)$$

where λ is the threshold defined by (23), t_m is the time when a fault occurs.

Once the fault is detected, fault diagnosis can again be carried out. For this purpose, we select the following observer as in equation 11

$$\dot{\hat{z}} = A\hat{z} + Br^* + E\hat{d} + B\Delta r + LC(\hat{z} - z) \quad (25)$$

where it has been denoted that

$$\hat{e} = C\hat{z} + \Delta r$$

By again denoting

$$\varepsilon = \hat{z} - z$$

$$\dot{\varepsilon} = (A + LC)\varepsilon + E(\hat{d} - d) - \Delta \quad (26)$$

For this error dynamics, by assuming again that d is a constant vector, we can choose the following Lyapunov function

$$V = \varepsilon^T P \varepsilon + (\hat{d} - d)^T \Gamma^{-1} (\hat{d} - d) \quad (27)$$

It can be calculated as

$$\dot{V} \leq -aV + b$$

where

$$\begin{aligned} a &= \min\left\{\beta, \frac{\sigma_d}{\lambda_{\max}(\Gamma^{-1})}\right\} \\ b &= \sigma_d \|d_0 - d\|^2 + \frac{1}{\gamma} M^2 \end{aligned}$$

This means that the adaptive diagnostic rule is convergent [21-22].

VI. SIMULATION STUDIES

To illustrate the effectiveness of the proposed algorithms, the following two layer operational control system is considered where the two control loops have the plant given by

$$G_1(s) = \frac{1}{5s+1}, G_2(s) = \frac{1}{4s+2}$$

respectively. Assuming that PI controllers are used for these two control loops with parameters given by $K_{P1} = 5, K_{I1} = 1, K_{P2} = 1.5, K_{I2} = 0.8$, respectively. Moreover these two control loops are both subjected to input fault with $\theta_1 = \theta_2 = 1$. For such system it can be seen that the parameters can be expressed as follows

$h_1 = 1, D_1 = 0, f_1 = -1, g_1 = 1, w_1 = 1, h_2 = 1, D_2 = 0, f_2 = -0.375, g_2 = 0.8, w_2 = 0.75, c_1 = c_2 = 1$. As a result, the whole system can be expressed as

$$\begin{aligned} \dot{z} &= Az + Br^* + Ed + B\Delta r + \Delta(z, \Delta r) \\ e &= Cz + \Delta r \end{aligned} \quad (28)$$

where the parameter matrices are formulated using PI parameters to give

$$\begin{aligned} A &= \begin{bmatrix} -0.2 & 0.2 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -0.5 & 0.25 \\ 0 & 0 & -0.05 & -0.375 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0.8 \end{bmatrix}, \quad E = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}, \\ r^* &= \begin{bmatrix} 28 \\ 66 \end{bmatrix}, \quad \Delta r = 0 \end{aligned}$$

For this system, two cases as described in sections 4 and 5 will be considered, where fault detection and diagnosis algorithms for the systems with and without model uncertainties will both be tested via simulation studies. The simulation results are therefore presented in the following.

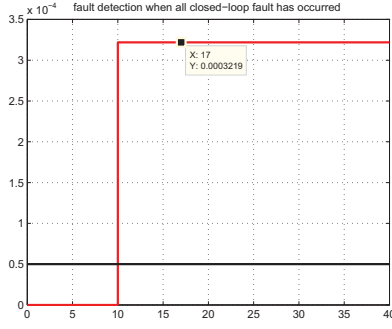


Fig. 3. The detection $\|\varepsilon_0\|$ and threshold λ (black line) when both fault occur.

A. Simulation Results for Systems without Model Uncertainties($\Delta(z, \Delta r) = 0$)

In this case the parameter matrices in the adaptive diagnostic rule are selected as follows

$$\Gamma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, Q = \begin{bmatrix} 0.8 & 0 & 0 & 0.05 \\ 0 & 4 & 0.25 & 1.375 \\ 0 & 0.25 & 3 & -0.05 \\ 0.05 & 1.375 & 0.05 & 1.5 \end{bmatrix}$$

Based upon the analysis in section 4, the following matrices can be readily calculated to give

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 \end{bmatrix}, L = \begin{bmatrix} 0.2 & -1 \\ 0.05 & 0.05 \\ 1 & 1 \\ 0.1 & 0.1 \end{bmatrix}$$

Using these parameter matrices, three cases will be studied as described in the following, where the fault detection threshold is set to $\lambda = 0.5 \times 10^{-4}$. In this case the faults happens at 10 seconds of the simulation with fault magnitude as $d_1 = 5$ and $d_2 = 8$ respectively. The simulation results are shown in figures 3 -5, where the fault detection signal is displayed in figure 3 and it has the magnitude of $\|\varepsilon_0\| = 0.0003219 > \lambda = 0.00005$. In this figure the red colour line stands for the detection signal and the black line is the threshold. It can be seen from this figure that after the actual faults occur at 10 seconds the detection signal can effectively detect the faults. To diagnose which and where the faults have occurred, the fault diagnosis algorithm obtained in section 4 is used to this example and the results are shown in figures 4 and 5. From these figures it can be seen that the estimated faults represented by red lines can converge to their actual values, respectively. This shows the effectiveness of the proposed algorithm.

B. Simulation Results for Systems Subjected to Model Uncertainties

In this case we assume that the model uncertainty of the system is given by

$$\Delta(z, \Delta r) = \begin{bmatrix} -0.001 \sin(z_1) \\ 0 \\ -0.002 \cos(0.5z_3) \\ 0 \end{bmatrix}$$

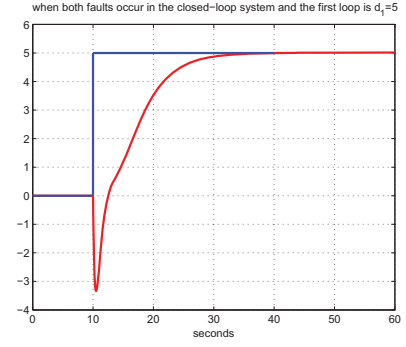


Fig. 4. The diagnosis of first closed-loop when both fault occur.

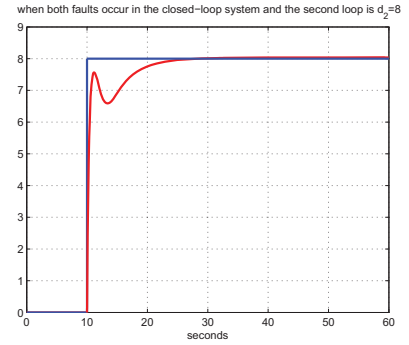


Fig. 5. The diagnosis of second closed-loop when both fault occur.

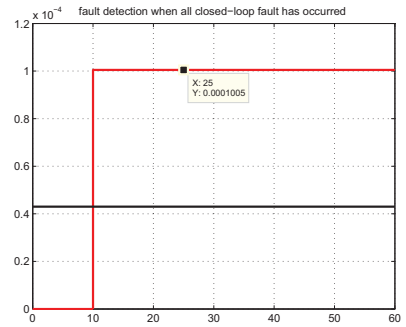


Fig. 6. The detection $\|\varepsilon_0\|$ (red line) and threshold λ (black line) when both fault occur.

where it can be shown that such a model uncertainty satisfies

$$\|\Delta(z, \Delta r)\| \leq \sqrt{5} \times 10^{-3} = M$$

This leads to the threshold of $\lambda \approx 0.43 \times 10^{-4}$. For the systems subjected to model uncertainties, we will still consider two cases in the following.

In this situation faults occur in both control loops with their values given by $d_1 = 1$ and $d_2 = 2$, respectively, at the 10 seconds time instant along the simulation phase. The simulation results are shown in Fig. 6–Fig. 8, where figure 6 shows the results of the fault detection and Fig. 7–Fig. 8 display the fault diagnosis results.

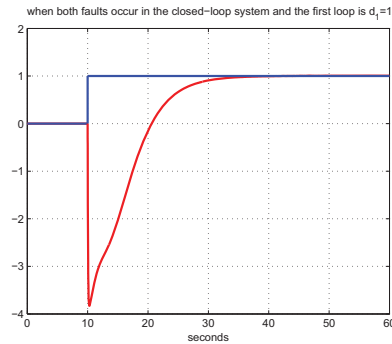


Fig. 7. The diagnosis of first closed-loop when both fault occur.

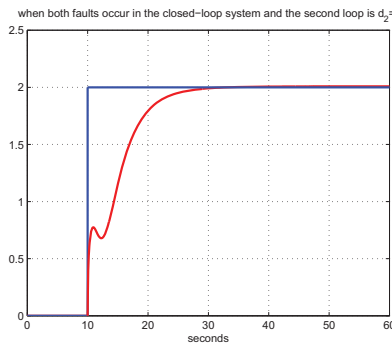


Fig. 8. The diagnosis of second closed-loop when both fault occur.

VII. CONCLUSION

In line with 2D operational structure of general complex industrial processes, in this paper a novel fault detection and diagnosis algorithm has been developed. The system considered is a two-layered structure which contains loop control layer and the operational control layer. A new model representation for such operational control systems has been proposed where the changes to the set-points to the loop control layer are regarded as the inputs and the outputs are those of loop control layer. Adaptive observer based fault detection and diagnosis algorithms are formulated using Lyapunov stability analysis and a simulated example has been included to show the effectiveness of the proposed methods. Encouraging results have been obtained.

ACKNOWLEDGMENT

The authors would like to thank the National Science Foundation of China(61290323,61333007,614730646), the IAPI Fundamental Research Funds(2013ZCX02-09),and the Fundamental Research Funds for the Central Universities(N130508002, N130108001).

REFERENCES

- [1] R. K. Mehra, J. Peschon, "An innovation approach to fault detection and diagnosis in dynamics", *Automatica*, vol. 7, pp. 637–640, 1971.
- [2] M. Iri, K. Aoki, E. O. Shima, et al., "An algorithm for diagnosis of system failures in the chemical process", *Computers and Chemical Engineering*, vol. 3, no. (1/4), pp. 489–493, 1979.

- [3] H. Wang and S. Daley, "Actuator fault diagnosis: an adaptive observer-based technique", *IEEE Trans on automatic control*, vol. 41, no. 7, pp. 1073–1078, 1996.
- [4] M. Misra, S.J. Qin, H. Yue, C. Ling, "Multivariate process monitoring and fault identification using multi-scale PCA", *Comput. Chem. Eng.*, vol. 26, pp. 1281–1293, 2002.
- [5] R. Li, J. H. Olson, "Fault detection and diagnosis in a closed-loop nonlinear distillation process: application of extended Kalman filters", *Industrial and Engineering Chemistry Research*, vol. 30, no.5, pp. 898–908, 1991.
- [6] E. Tian, D. Yue, "Reliable H1-filter design for T-S fuzzy model-based networked control systems with random sensor failure", *International Journal of Robust and Nonlinear Control*, vol. 23, no. 1, pp. 15–32, 2013.
- [7] X. Du, D. Liu, Y. Li, Y. Xiong, X. Wang, "Application of adaptive neuro-fuzzy inference system in power system fault recurrence", *Power System Technology*, vol. 30, no. 6, pp. 82–87, 2006.
- [8] Y. Zhang, X. Wang, F. Wang, "Fault accommodation for nonlinear systems using fuzzy adaptive sliding control", *Control and Decision*, vol. 20, no. 4, pp. 408–412, 2005.
- [9] H. H. Niemann, J. Stoustrup, "Robust fault detection in open loop vs. closed loop", In: *Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, California, USA: IEEE*, pp. 4496–4497, 1997.
- [10] Y. Xing, H. Wu, X. Wang, Z. Li, "Survey of fault diagnosis and fault-tolerance control technology for space craft", *Journal of Astronautics*, vol. 24, no. 3, pp. 221–226, 2003.
- [11] J. Wei, Z. Cen, R. Jian, "Sensor fault-tolerant observer applied in satellite attitude control", *Journal of Systems Engineering and Electronics*, vol. 23, no. 1, pp. 99–107, 2012.
- [12] S. M. M. Alavi, R. I. Zamanabadi, M. J. Hayes, "Robust fault detection and isolation technique for single-input/single-output closed-loop control systems that exhibit actuator and sensor faults", *IET Control Theory and Applications*, vol. 11, no. 2, pp. 951–965, 2008.
- [13] R. Nath, and Z. Alzein, "On-line dynamic optimization of olefins plants", *Computers and Chemical Engineering*, vol. 24, pp. 533–538, 2000.
- [14] J. L. Ding, T. Y. Chai, H. Wang, "Knowledge-based plant-wide dynamic operation of mineral processing under uncertainty", *IEEE Trans on Industry Informatics*, vol. 8, no. 4, pp. 849C–859, 2012.
- [15] J. Hasikos, H. Sarimveis, P. L. Zervas, N. C. Markatos, "Operational optimization and real-time control of fuel-cell systems", *Journal of Power Sources*, vol. 193, pp. 258–268, 2009.
- [16] P. Zhou, T. Y. Chai, J. Sun, "Intelligence-based supervisory control for optimal operation of a DCS-controlled grinding system", *IEEE Transactions on Control Systems Technology*, vol. 21, no. 1, pp. 162C–175, 2013.
- [17] P. Tatjewski, "Advanced control and on-line process optimization in multilayer structures", *Annual Reviews in Control*, vol. 32, pp. 71–85, 2008.
- [18] J. Jaschke, S. Skogestad, "NCO tracking and self-optimizing control in the context of real-time optimization", *Journal of Process Control*, vol. 21, pp. 1047–1416, 2011.
- [19] R. J. Patton, P. Frank and R. Clark, "Fault diagnosis in dynamic systems: theory and application", *Prentice Hall, Englewood Cliffs, NJ*, 1989.
- [20] P.M. Frank, "Enhancement of robustness in observer-based fault detection", *Int. J. Control*, vol. 59, pp. 955–981, 1994.
- [21] S. S. Ge and C. Wang, "Direct Adaptive NN Control of a Class of Nonlinear Systems", *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 214–221, 2002.
- [22] S. Tong, Y. Li, P. Shi, "Fuzzy adaptive backstepping robust control for SISO nonlinear system with dynamic uncertainties", *Information Sciences*, vol. 179, pp. 1319–1332, 2009.

Improved Kernel Canonical Variate Analysis for Process Monitoring

Raphael T. Samuel

Oil and Gas Engineering Centre,
School of Energy, Environment and Agrifood (SEEA),
Cranfield University,
Cranfield, Bedford, MK43 0AL, UK.
Email: r.t.samuel@cranfield.ac.uk

Yi Cao

Oil and Gas Engineering Centre,
School of Energy, Environment and Agrifood (SEEA),
Cranfield University,
Cranfield, Bedford MK43 0AL, UK.
Email: y.cao@cranfield.ac.uk

Abstract—This paper proposes a kernel canonical variate analysis (KCVA) approach for process fault detection. The technique employs the kernel principle to map the original process observations to a high dimensional feature space on which canonical variate analysis is performed. The aim is to obtain an effective monitoring technique that accounts for non-linearity and process dynamics simultaneously. The kernel principle accounts for non-linearity while the CVA accounts for serial correlations widely encountered in dynamic processes. The kernel CVA algorithm proposed in this work is based on QR decomposition in order to avoid singularity problems associated with kernel matrices which require a regularisation step. The technique is evaluated using the Tennessee Eastman Challenge process. Tests show the effectiveness of the proposed kernel CVA approach.

I. INTRODUCTION

Using multivariate statistical techniques to monitor chemical plants have gained much research interest in the last few decades. These techniques depend mainly on process history data and are therefore relatively easier to employ in large scale processes (i.e. processes involving several dozens or higher number of measured variables) compared to the classical approaches based on rigorous process models derived from first principles. Process plants generate large amounts of data from measuring several variables during normal operations. This is possible due to advancement in instrumentation and automation technology. The data acquired are easily stored and/or explored to extract useful information about the process. Improvement in data analysis applications and increase in computer power have also contributed immensely in providing the stage for data-driven techniques to thrive. Two early examples of multivariate statistical methods which are very widely used are Principal Component Analysis (PCA) [1] and Canonical Correlation Analysis (CCA) [2]. Both PCA and CCA are eigenvalue problems, however, PCA is used in cases involving a single collection of variables while CCA is used when considering two sets of variables.

Canonical correlation analysis attempts to find the existing relations between two multivariate data sets. This is achieved by obtaining linear combinations of each of the original sets of variables and determining the pairwise correlations of the linear combinations of the two sets of variables. The linear combinations are called canonical variates while the

pairwise correlations are known as canonical correlations. The strength of the association between the two sets of variables is measured by the canonical correlations. If correlation is considered to be the main determinant of information in the original two blocks of variables, then CCA can be used to obtain a reduced dimensional set of variables from the original data sets by discarding the canonical variate pairs with very low correlations. However, both PCA and CCA are static and linear techniques. They are therefore deficient in capturing relations in data rich in dynamic and non-linear characteristics.

Many complex chemical industry processes exhibit both non-linear and dynamic behaviour. Therefore, to effectively monitor such processes, the techniques employed are expected to capture these characteristics. Otherwise abnormal process conditions may be detected long after they have occurred or may not be detected at all. Both of these situations can compromise process safety, operational efficiency and consistent product quality, which are extremely important in chemical process industries. Ineffective monitoring can also lead to less than optimal maintenance practices leading to frequent equipment break down, longer downtimes and higher operational cost.

Several methods have been proposed to address either non-linearity or process dynamics separately but not many studies have addressed tackling both properties simultaneously. One of the few studies reported in the literature which address both of these characteristics was conducted by Choi and Lee [3]. They proposed the dynamic kernel PCA approach and tested it on a simulated non-linear process as well as a wastewater treatment process. This approach employed a kernel function to capture the non-linear relations and a time lag-data extension of the original observations to describe the dynamics of the process. They reported that the method provided better monitoring evidenced in lower missing alarms and smaller detection times compared to PCA and KPCCA. The kernel methods are the preferred techniques for capturing non-linear relations compared to methods based on neural network and principal curves because they do not require solving a non-linear optimization problem. However, in the DKPCA technique proposed by Choi and Lee, the kernel approach was combined with an extension of the PCA (dynamic PCA)

which has limitation in capturing process dynamics [4]. This is likely to limit the performance of this technique especially in faults that are not easily detectable. In other words, although the approach has a good technique for describing non-linear relations, the method employed for accounting for process dynamics, being an extension of a linear algorithm (i.e. PCA), leaves room for improvement.

Canonical variate analysis (CVA) is a state-space based technique which is widely reported as an appropriate methodology for monitoring dynamic processes [5]–[8]. Like the CCA, the CVA finds relations between two sets of variables but the two sets of variables are obtained from expanding an observation at a given time instant to p past and f future measurements, in order to account for serial correlations.

To improve the monitoring of non-linear dynamic processes, Odiowei and Cao [4] proposed the CVA with KDE technique. In their work, the CVA was associated with kernel density-based upper control limits derived from the estimated probability density functions of the monitoring indices instead of determining the control limits based on the Gaussian assumption. Nevertheless, this approach does not directly address non-linear problems in a dynamic process. Considering the successful application of kernel methods in several application domains and the CVA technique in describing non-linear and dynamic behaviour respectively, a combination of these two approaches should make an appropriate scheme for describing non-linear and dynamic relations simultaneously in data-driven process monitoring. However, not much is reported on kernel CVA in the literature even though many studies involving kernel extensions to the CCA exist [9]–[11]. A management system based on kernel CVA to monitor and diagnose smart homes is reported by Giantomassi and others in [12] but application of this technique in the chemical process industry is not well investigated. Also, the work mentioned above does not provide a comparison of the technique with other approaches. This makes it difficult to assess how its performance compares with other known techniques. Furthermore, since the past and future kernel matrices generated are singular, regularisation of these matrices is needed to perform the matrix inversion step required to implement their CVA algorithm.

The objective of this paper is to implement KCVA using QR decomposition to preclude the need for regularising the kernel matrices generated and to investigate the performance of the proposed approach in monitoring a chemical process. The paper also provides a comparison of the effectiveness of the technique with the kernel CVA approach based on the regularisation of singular kernel matrices. Assessment of both techniques was done by applying them to simulation data obtained from the Tennessee Eastman benchmark process.

The rest of the paper is organised as follows: Section II summarises the KCVA procedure adopted. Section III shows how to compute the upper control limits of the monitoring statistics using the KDE method. The proposed kernel CVA based process monitoring procedure is presented in Section IV. Section V describes the application to the Tennessee Eastman process while conclusions reached are presented in Section VI.

II. KERNEL CANONICAL VARIATE ANALYSIS

The idea of kernel CVA is to extract state variables that also capture non-linear characteristics in the observed data using non-linear kernel transformation and CVA. A brief description of the kernel CVA technique adopted is given in this section. Detailed discussion including mathematical procedure on non-linear mapping based on a kernel function and the CVA algorithm can be found in [4], [13]–[15].

To account for time correlations, each observation vector \mathbf{x} is expanded at a given time point t to obtain information from the past (p) and future (f) measurements each containing d variables using (1):

$$\mathbf{x}_{(p,t)} = \begin{bmatrix} \mathbf{x}_{(t-1)} \\ \mathbf{x}_{(t-2)} \\ \vdots \\ \mathbf{x}_{(t-p)} \end{bmatrix} \in \mathbb{R}^{dp} \quad \text{and} \quad \mathbf{x}_{(f,t)} = \begin{bmatrix} \mathbf{x}_{(t)} \\ \mathbf{x}_{(t+1)} \\ \vdots \\ \mathbf{x}_{(t+f-1)} \end{bmatrix} \in \mathbb{R}^{df} \quad (1)$$

The various components are mean-centred as follows:

$$\hat{\mathbf{x}}_{(p,t)} = \mathbf{x}_{(p,t)} - \bar{\mathbf{x}}_{(p,t)} \quad \text{and} \quad \hat{\mathbf{x}}_{(f,t)} = \mathbf{x}_{(f,t)} - \bar{\mathbf{x}}_{(f,t)} \quad (2)$$

where $\bar{\mathbf{x}}_{(p,t)}$ and $\bar{\mathbf{x}}_{(f,t)}$ are the sample means of $\mathbf{x}_{(p,t)}$ and $\mathbf{x}_{(f,t)}$ respectively. The past and future vectors are then arranged together in columns to obtain the corresponding past and future matrices, \mathbf{X}_p and \mathbf{X}_f respectively.

$$\mathbf{X}_p = [\hat{\mathbf{x}}_{(p,p+1)}, \hat{\mathbf{x}}_{(p,p+2)}, \dots, \hat{\mathbf{x}}_{(p,p+M)}] \in \mathbb{R}^{dp \times M} \quad (3)$$

$$\mathbf{X}_f = [\hat{\mathbf{x}}_{(f,p+1)}, \hat{\mathbf{x}}_{(f,p+2)}, \dots, \hat{\mathbf{x}}_{(f,p+M)}] \in \mathbb{R}^{df \times M} \quad (4)$$

where the columns of the truncated Hankel matrices for N observations is $M = N - f - p + 1$.

To apply the kernel principle, non-linear mappings, Φ_1 and Φ_2 are used to map \mathbb{R}^{dp} and \mathbb{R}^{df} into a high dimensional feature space, $\Phi_1 : \mathbb{R}^{dp} \rightarrow F$ and $\Phi_2 : \mathbb{R}^{df} \rightarrow F$ respectively. Kernel matrices (\mathbf{K}_p and \mathbf{K}_f) are obtained using the kernel trick, ([13], [14]):

$$\mathbf{K}_p = \langle \Phi_1(\mathbf{X}_p), \Phi_1(\mathbf{X}_p) \rangle, \quad (5)$$

$$\mathbf{K}_f = \langle \Phi_2(\mathbf{X}_f), \Phi_2(\mathbf{X}_f) \rangle \quad (6)$$

where the elements of these kernel matrices are defined as

$$(\mathbf{K}_p)_{ji} = \langle \Phi_1(\hat{\mathbf{x}}_{(p,p+j)}), \Phi_1(\hat{\mathbf{x}}_{(p,p+i)}) \rangle$$

$$(\mathbf{K}_f)_{ji} = \langle \Phi_2(\hat{\mathbf{x}}_{(f,p+j)}), \Phi_2(\hat{\mathbf{x}}_{(f,p+i)}) \rangle$$

for all $j, i = 1, \dots, M$. These kernel matrices are mean-centred as follows:

$$\mathbf{K}_{cp} = \mathbf{K}_p - \mathbf{B}\mathbf{K}_p - \mathbf{K}_p\mathbf{B} + \mathbf{B}\mathbf{K}_p\mathbf{B} \quad (7)$$

$$\mathbf{K}_{cf} = \mathbf{K}_f - \mathbf{B}\mathbf{K}_f - \mathbf{K}_f\mathbf{B} + \mathbf{B}\mathbf{K}_f\mathbf{B} \quad (8)$$

where \mathbf{K}_{cp} and \mathbf{K}_{cf} are the past and future mean-centred kernel matrices, \mathbf{B} is an $M \times M$ matrix in which each element is equal to $\frac{1}{M}$.

Kernel CVA seeks to find weights which make the linear combinations of \mathbf{K}_{cp} and \mathbf{K}_{cf} have maximal correlations.

Since kernel matrices are ill-conditioned and suffer computational instabilities, a regularisation step is normally needed so that the matrix inversion required by the CVA algorithm can be carried out. However, such a regularisation step reduces the accuracy of the model which makes the monitoring performance poor. In this paper, the mean-centred past and future kernel matrices were factorised using QR decomposition as follows:

$$\mathbf{K}_{cp} = \mathbf{Q}_p \mathbf{R}_p \quad \text{and} \quad \mathbf{K}_{cf} = \mathbf{Q}_f \mathbf{R}_f, \quad (9)$$

where \mathbf{Q}_p and \mathbf{Q}_f are orthogonal matrices and \mathbf{R}_p and \mathbf{R}_f are upper triangular matrices. Though \mathbf{K}_{cp} and \mathbf{K}_{cf} are not full rank, they were managed by using the MATLAB backslash operator which makes them equivalent to pseudo-inverses. The product of the orthogonal matrix pair was computed and canonical variates were obtained by performing singular value decomposition (SVD):

$$\mathbf{W} = \mathbf{Q}_f^T \mathbf{Q}_p = \mathbf{U} \mathbf{S} \mathbf{V}^T, \quad (10)$$

where T denotes transpose, \mathbf{U} and \mathbf{V} are orthogonal matrices, while \mathbf{S} is a diagonal matrix whose entries on the main diagonal (singular values) show the degree of correlation between pairs of \mathbf{U} and \mathbf{V} . This procedure precludes the computational problems associated with obtaining the scaled Hankel matrix for performing SVD when covariance and cross-covariance matrices are used in a KCVA-based methodology as proposed in [12] and is a major strength of the proposed approach.

The normalised left and right singular vectors, (\mathbf{U}^*) and (\mathbf{V}^*) respectively are obtained using the following:

$$\mathbf{U}^* = \mathbf{R}_f^+ \mathbf{U} (M-1)^{\frac{1}{2}} \quad \text{and} \quad \mathbf{V}^* = \mathbf{R}_p^+ \mathbf{V} (M-1)^{\frac{1}{2}}, \quad (11)$$

where the superscript in \mathbf{R}_f^+ and \mathbf{R}_p^+ represent pseudo-inverse. Sorting the normalised singular values and the columns of the singular vectors associated with them in descending order makes \mathbf{V}_n^* (i.e. the first n columns of \mathbf{V}^*), the most dominant pairwise correlations with those of \mathbf{U}^* . Thus, the transformation matrices for determining the n -dimensional state variables and residuals are obtained as:

$$\mathbf{J} = \mathbf{V}_n^* \in \mathbb{R}^{M \times n} \quad \text{and} \quad \mathbf{L} = (\mathbf{I} - \mathbf{J} \mathbf{J}^T) \in \mathbb{R}^{M \times M} \quad (12)$$

The state space \mathbf{Z} and residual space \mathbf{E} are computed using (13):

$$\mathbf{Z} = \mathbf{J} \cdot \mathbf{K}_{cp} \in \mathbb{R}^{n \times M} \quad \text{and} \quad \mathbf{E} = \mathbf{L} \cdot \mathbf{K}_{cp} \in \mathbb{R}^{M \times M} \quad (13)$$

Hotellings T^2 and the Q statistic or squared prediction error (SPE) are also used in kernel CVA as the monitoring statistics. The Hotellings T^2 monitors the changes in the state space while the Q statistic monitors the changes in the residual space. They are determined using (14)

$$T_k^2 = \sum_{i=1}^n z_{i,k}^2 \quad \text{and} \quad Q_k = \sum_{i=1}^M e_{i,k}^2, \quad (14)$$

where n is the number of states retained, $z_{i,k}$ and $e_{i,k}$ are $(i, k)^{\text{th}}$ the entries of \mathbf{Z} and \mathbf{E} matrices respectively.

III. COMPUTATION OF UPPER CONTROL LIMITS

To correct the Gaussian assumption, the kernel density estimation (KDE) technique was used to estimate the probability density functions (PDF) of the monitoring indices. Upper control limits were computed from the estimated PDFs instead of using parametrically obtained control limits.

Given the probability density function $g(x)$, the probability of x to be less than c at a specified confidence level α is given by (15):

$$P(x < c) = \int_{-\infty}^c g(x) dx = \alpha \quad (15)$$

The control limits of the monitoring statistics (T^2 and Q) were determined using (16).

$$\int_{-\infty}^{T_\alpha^2} p(T^2) dT^2 = \alpha \quad \text{and} \quad \int_{-\infty}^{Q_\alpha} p(Q) dQ = \alpha \quad (16)$$

A more comprehensive account of the KDE technique including the importance of selecting the bandwidth H and methods of obtaining an optimum value can be found in [16] and [17].

IV. FAULT DETECTION PROCEDURE FOR KERNEL CANONICAL VARIATE ANALYSIS

Similar to other multivariate statistical process monitoring methodologies, the fault detection strategy of kernel CVA involves two phases: off-line training and on-line monitoring or testing. The off-line training phase involves development of the process model, calculation of the monitoring indices and their upper control limits using the normal operation data. Conversely, on-line monitoring involves computing the monitoring indices using faulty or test data and comparing their values with the control limits obtained in the off-line training phase to determine the status of the process. The steps involved in the proposed kernel CVA technique for the training and monitoring phases are outlined below:

A. Off-line Training

- 1) Obtain observation vector.
- 2) Expand observation vector at each time point t to obtain information from the past (p) and future (f) measurements using (1).
- 3) Form kernel matrices of the past and future measurements.
- 4) Mean-centre the past and future kernel matrices. Factorise the mean-centred past and future kernel matrices using QR decomposition to obtain pairs of upper triangular and orthogonal matrices.
- 5) Compute the product of the orthogonal matrix pair from step 4 and perform singular value decomposition. Normalise the canonical coefficients.
- 6) Determine states and residuals.
- 7) Compute monitoring indices T^2 and Q at each time point as the sum of the squared state variables and residuals respectively.

B. On-Line Monitoring

- 1) Acquire test data and define past and future matrices and arrange data similar to training data.
- 2) Form kernel matrices of the past and future measurements using the same function and parameters used in the training stage and mean-centre.
- 3) Calculate states and residuals of test data.
- 4) Compute T^2 and Q of test data.
- 5) Monitor process by comparing value of T^2 and Q against their control limits. A fault is detected if both monitoring indices exceed their control limits.

V. APPLICATION STUDY

In this section kernel CVA via QR decomposition and kernel CVA with regularisation were applied to the Tennessee Eastman (TE) process for monitoring performance evaluation. Faults 3, 9, and 15 of the TE process were considered for the application study. Faults 3 and 9 are step change and random variation in D feed temperature respectively, while Fault 15 is condenser cooling water sticking. These faults are usually more difficult to detect in the TE process.

A. Tennessee Eastman Process

The TE process is a simulation of a real industrial plant manifesting both non-linear and dynamic properties. It is widely used as a benchmark process for evaluating process monitoring and control approaches [18]. It is made up of five major units: separator, compressor, reactor, stripper and condenser, and eight components coded A to H. A schematic diagram of the processes is presented in Figure 1.

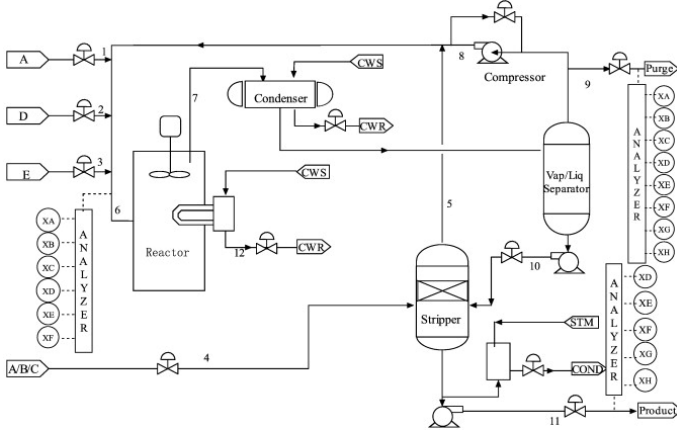


Fig. 1: Schematic diagramme of TE process

The TE process consists of 960 samples and 52 variables which include 22 continuous variables, 19 composition measurements sampled by 3 composition analysers and 12 manipulated variables. All the 22 continuous variables and 11 manipulated variables were used in this study. Variables 1 to 22 in Table I are measurement variables while 23 to 33 are manipulated variables. The agitation speed of the reactor's

TABLE I: Monitoring variables for the TE process

No.	Description	No.	Description
1	A feed (stream 1)	18	Stripper temperature
2	D feed (stream 2)	19	Stripper stream flow
3	E feed (stream 3)	20	Compressor work
4	Total feed (stream 4)	21	Reactor cooling water outlet temp
5	Recycle flow (stream 8)	22	Separator cooling water outlet temp
6	Reactor feed rate (stream 6)	23	D feed flow (stream 2)
7	Reactor pressure	24	E feed flow (stream 3)
8	Reactor level	25	A feed flow (stream 1)
9	Reactor temperature	26	Total feed flow (stream 4)
10	Purge rate	27	Compressor recycle valve
11	Separator temperature	28	Purge valve (stream 9)
12	Separator level	29	Separator pot liquid flow (stream 10)
13	Separator pressure	30	Stripper liquid product flow
14	Separator under flow	31	D Stripper stream valve
15	Stripper level	32	Reactor cooling water flow
16	Stripper pressure	33	Condenser cooling water flow
17	Stripper under flow		

stirrer (the 12th manipulated variable) was not included because it is constant. Details of the 33 variables are shown in Table I.

B. Parameters Selection

Some key parameters were determined to optimise the approach. These include the kernel used and its bandwidth, the length of lag used, and the number of states retained. The radial basis kernel which is a common choice in previous studies [14], [19] was used in this paper. The value of the kernel parameter c can be determined using cross validation or by using the relation $c = Wn\sigma^2$, where W is a constant which is dependent on the data being used, n and σ^2 are the dimension and variance of the input space respectively [14], [20]. The latter method was adopted in this study with a value of $c = 20$.

The length of lag represents the number of past or future observations that correlate significantly with an observation at a particular time point. For the CVA algorithm, the optimal number of lags can be obtained by considering the summed squares of all process measurements [4]. Consequently, a lag of 15 was adopted.

According to Negiz and Cinar [21], the states to retain can be selected based on the number of dominant singular values. Fig. 2 shows the normalised singular values of the training data. It can be observed from Fig. 2 that the singular values decrease very slowly. Therefore, choosing how many states to retain based on dominant singular values will not give a realistic model [4]. Furthermore, the number of states retained does not really matter in this case because the monitoring indices were used jointly for fault detection due to their complementary nature. This means that fault detection is acknowledged if either the T^2 or Q statistic detects a fault since detectable process variation may not always occur in both the model space and the residual space at the same time. Consequently, 16 states were retained to curtail the false alarm rate.

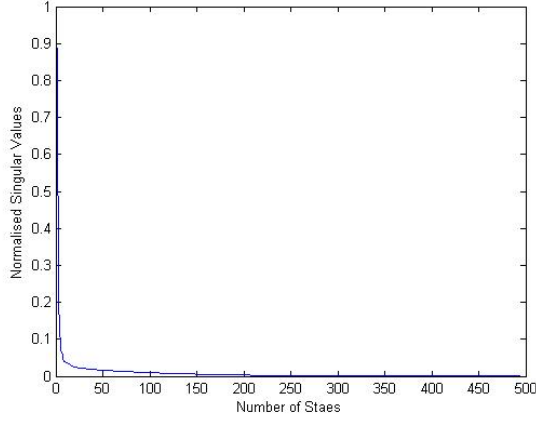


Fig. 2: Normalised singular values of training data

C. Results and Discussion

Monitoring performance was based on fault detection rates (FDR), false alarm rate (FAR), and detection time. FDR is the percentage of faulty observations identified correctly. It was computed as

$$FDR = \frac{\theta_{fc}}{\theta_{tf}} \times 100 \quad (17)$$

where θ_{fc} denotes the number of fault samples identified correctly and θ_{tf} is the total number of fault samples. FAR is the percentage of observations identified as abnormal under normal operating mode. It was calculated as,

$$FAR = \frac{\theta_{nf}}{\theta_{tn}} \times 100 \quad (18)$$

where θ_{nf} is number of normal observations reported as faulty and θ_{tn} represent the total number of normal observations. Detection delay was computed as the amount of time that passes before a fault is detected after it has occurred.

Table IIa shows the rate of fault detection for Faults 3, 9 and 15 using the QR decomposition approach while Tables IIb, IIc, and IId show results obtained at different values of regularisation for the same faults. The detection rates at a regularisation value of 10^{-2} were the lowest (51.25, 78.13, and 85.38 percent) for Faults 3, 9 and 15 respectively. At a very small regularization value of 10^{-8} , the detection rates improved but the values were still lower than the results obtained via QR decomposition. Also, the detection delay for the QR-based approach for all three faults was 15 seconds while the corresponding rates for the technique based on regularisation were 54/45, 84/63, and 54/45 seconds for Faults 3, 9 and 15 respectively for the worst and best detection time delays. In all faults considered, the QR based detection times were better than the best detection times obtained via the regularisation approach. The regularisation approach also had higher FARs (which makes it relatively poorer) except for the smallest regularisation value.

Fig. 3 shows the monitoring statistics for Fault 15. The poor monitoring performance arising from choosing a poor

TABLE II: Detection performance (a) KCVA with QR decomposition (Faults 3, 9 and 15), (b) KCVA with regularisation (Fault 3), (c) KCVA with regularisation (Fault 9), and (d) KCVA with regularisation (Fault 15)

(a)			
	Fault 3	Fault 9	Fault 15
FDR (%)	98.25	97.50	98.25
FAR	0.0382	0.0382	0.0382
Detection delay, s	15	15	15
(b)			
Regularisation value	10^{-2}	10^{-5}	10^{-8}
FDR (%)	51.25	98.13	98.13
FAR	0.0458	0.0840	0.0076
Detection delay, s	54	45	45
(c)			
Regularisation value	10^{-2}	10^{-5}	10^{-8}
FDR (%)	78.13	97.38	97.38
FAR	0.0458	0.0840	0
Detection delay, s	84	63	63
(d)			
Regularisation value	10^{-2}	10^{-5}	10^{-8}
FDR (%)	85.38	98.13	98.13
FAR	0.0458	0.0840	0
Detection delay, s	54	45	45

regularisation value is shown in Fig 3b. It can be seen that the monitoring index (the solid signal) did not fully go above the control limit (dash-dot horizontal line) most of the time which shows that fault detection performance was poor.

VI. CONCLUSIONS

The existing kernel CVA technique is improved in this paper for detecting process faults. Problems of non-linearity and dynamism in processes were accounted for simultaneously by employing the kernel principle followed by the CVA technique. In the proposed method the product matrix obtained from the past and future kernel matrices, on which singular value decomposition was performed at the CVA stage, was obtained via QR decomposition to avoid singularity problems associated with kernel matrices, such that there was no need to carry out regularisation of the generated kernel data. Results obtained from applying this technique to the Tennessee Eastman process were compared with results based on different values of regularisation. The results show that the proposed technique outperformed the KCVA based on regularisation in both monitoring rate and the time taken to detect faults. This supports the effectiveness of the proposed method in enhancing process monitoring performance. Avoiding the need to determine an optimum regularisation value reduces the parameters required for implementing kernel-based CVA by

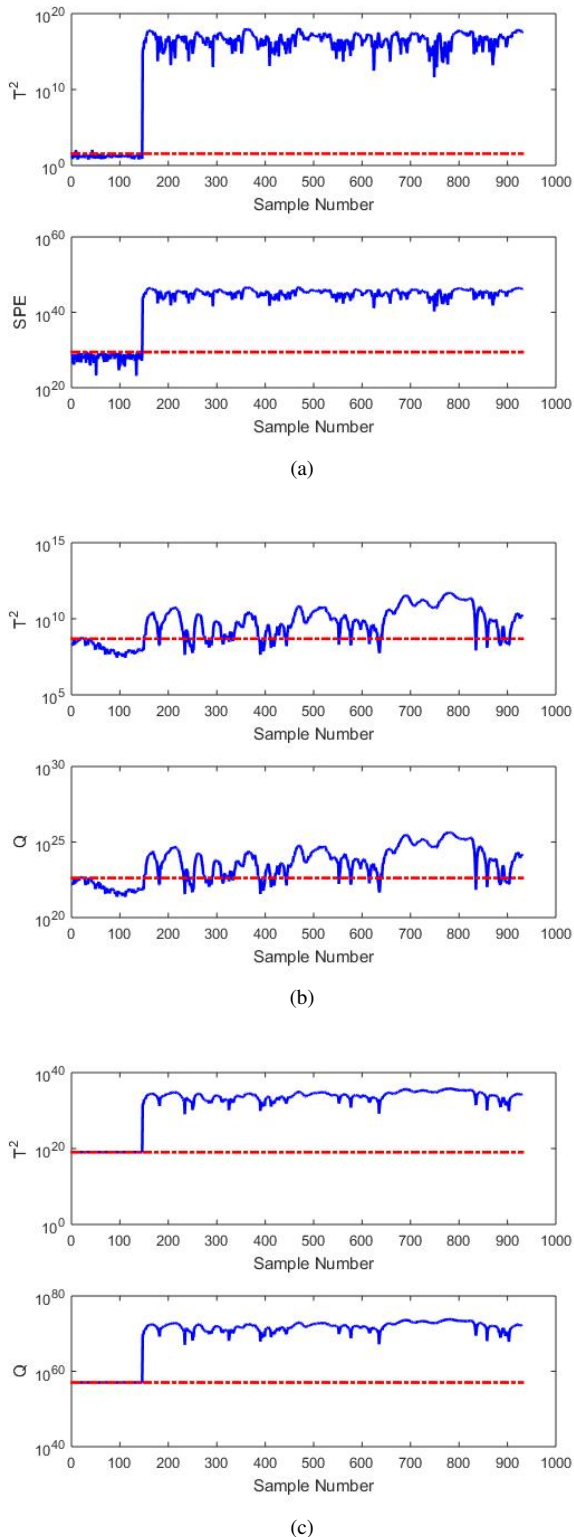


Fig. 3: Monitoring statistics of Fault 15. (a) KCVA with QR, (b) KCVA with regularisation (10^{-2}), (c) KCVA with regularisation (10^{-8})

one. This is desired because a poorly chosen regularisation parameter gives poor monitoring results.

REFERENCES

- [1] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 498–520, 1933.
- [2] —, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 321–377, 1936.
- [3] S. W. Choi and I.-B. Lee, "Nonlinear dynamic process monitoring based on dynamic kernel pca," *Chemical Engineering Science*, vol. 59, pp. 5897–5908, 2004.
- [4] P.-E. P. Odiowei and Y. Cao, "Nonlinear dynamic process monitoring using canonical variate analysis and kernel density estimation," *IEEE Transaction on Industrial Informatics*, vol. 6, no. 1, pp. 36–44, 2010a.
- [5] E. L. Russell, L. H. Chiang, and R. D. Braatz, "Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 51, pp. 81–93, 2000.
- [6] B. C. Juricek, D. E. Seborg, and W. E. Larimore, "Fault detection using canonical variate analysis," *Industrial & Engineering Chemistry Research*, vol. 43, pp. 458–474, 2004.
- [7] B. Jiang, X. Zhu, D. Huang, J. A. Paulson, and R. D. Braatz, "A combined canonical variate analysis and fisher discriminant analysis (cva-fda) approach for fault diagnosis," *Computers and Chemical Engineering*, vol. 77, pp. 1–9, 2015.
- [8] B. Jiang, D. Huang, X. Zhu, F. Yang, and R. D. Braatz, "Canonical variate analysis-based contributions for fault identification," *Journal of Process*, vol. 26, pp. 17–25, 2015.
- [9] X. Zhu, Z. Huang, H. T. Shen, J. Cheng, and C. Xu, "Dimensionality reduction by mixed kernel canonical correlation analysis," *Pattern Recognition*, vol. 45, no. 8, pp. 3003–3016, 2012.
- [10] S.-Y. Huang, M.-H. Lee, and C. K. Hsiao, "Nonlinear measures of association with kernel canonical correlation analysis and applications," *Journal of Statistical Planning and Inference*, vol. 139, pp. 2162–2174, 2009.
- [11] S. Tan, F. Wang, Y. Chang, W. Chen, and J. Xu, "Fault detection and diagnosis of nonlinear processes based on kernel ica-kcca," in *Chinese Control and Decision Conference*. Xuzhou: IEEE, May 2010, pp. 3869–3874.
- [12] A. Giantomassi, F. Ferracuti, S. Iarlori, S. Longhi, A. Fonti, and G. Comodi, "Kernel canonical variate analysis based management system for monitoring and diagnosing smart homes," in *2014 International Joint Conference on Neural Networks, IJCNN 2014, Beijing, China, July 6-11, 2014*, 2014, pp. 1432–1439. [Online]. Available: <http://dx.doi.org/10.1109/IJCNN.2014.6889821>
- [13] B. Scholkopf, A. Smola, and K. Muller, "Non-linear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1399, 1998.
- [14] J. M. Lee, C. K. Yoo, S. W. Choi, P. A. Vanrolleghem, and I. B. Lee, "Nonlinear process monitoring using kernel principal component analysis," *Chemical and Engineering Science*, vol. 59, pp. 223–234, 2004.
- [15] R. T. Samuel and Y. Cao, "Fault detection in a multivariate process based on kernel pca and kernel density estimation," in *International Conference on Automation and Computing*. Cranfield, UK: IEEE, September 2014.
- [16] X.-W. Chen, "An improved branch and bound algorithm for feature selection," *Pattern Recognition Letters*, vol. 24, pp. 1925–1933, 2003.
- [17] L. Liang, "Multivariate statistical process monitoring using kernel density estimation," *Development in Chemical Engineering and Mineral Processing*, vol. 13, pp. 185–192, 2005.
- [18] L. Chiang, E. Russell, and R. Braatz, *Fault Detection and Diagnosis in Industrial Systems*. London: Springer, 2001.
- [19] G. Stefatos and A. B. Hamza, "Statistical process control using pca," in *Proceedings of 15th Mediterranean Conference on Control and Automation*. Athens, Greece: IEEE, July 2007, pp. 1418–1423.
- [20] S. Mika, B. Scholkopf, A. Smola, K. R. Muller, M. Scholz, and G. Ratsch, "Kernel pca and de-noising in feature spaces," in *Advances in Neural Information Processing System*, vol. 11, 1999, pp. 536–542.
- [21] A. Negiz and A. Cinarl, "Monitoring of multivariable dynamic processes and sensor auditing," *Journal of Process Control*, vol. 8, no. 5-6, pp. 375–380, 1998.

Solution to Failure Detection of Closed-loop Systems and Application to IC Engines

Lan-Xiang Zhu¹, Feng Yu¹,

¹ School of Electronic Information, Changchun Architecture & Civil Engineering College, Changchun, China.

Ding-Wen Yu²,

² School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao, China.

A.M.S. Ertiame³, D. L. Yu³

³ Control Group, School of Engineering, Liverpool John Moores University, Liverpool, U.K.

Corresponding author: d.yu@ljmu.ac.uk

Abstract—When a neural network model is trained to predict system output, the prediction error can be used as residual to report fault. Most existing research uses system open-loop input/output data, while the trained model is used to detect fault when the system runs under closed-loop control. This paper analyses the drawback of the training data acquisition and proposes new data acquisition method, so that the model accuracy is greatly improved. In addition, detection of the sensor fault, which is involved in the closed-loop, is discussed and the simulation for detecting such sensor faults are conducted. The new scheme is assessed and validated by being applied to the automotive engine air path to detect some simulated faults. The simulation results show that the developed method is effective and the residual is more sensitive to the faults.

Keywords—Fault detection; automotive engines; closed-loop fault detection; independent RBF model.

I. INTRODUCTION

Fault detection and isolation (FDI) for automotive engines have been investigated for over two decades, but real applications with on-board FDI for dynamic faults (not the faults occurred during the steady state) have seldomly reported. There are two main reasons. One is that the false alarm rate is too high to be acceptable; the other is that most methods are too complicated and need accurate mathematical model, which is not practical to implement.

The typical work that has done before is briefly introduced here. Gertler [1-2] and his co-workers attempted using linear Parity space method to detect simulated faults. Simulated faults and experiments have been done. But the results were not very satisfactory due to the severe nonlinearities in engine dynamics. Reference [3] introduced an electro-mechanical position servo, used in the speed control of large diesel engines, as a benchmark for model based fault detection and identification. Isermann [4] proposed model-based fault-detection and diagnosis methods for some technical processes. The goal was to generate several symptoms indicating the difference between nominal and faulty status. On-line sensor fault detection, isolation, and accommodation in automotive engines have been studied by Capriglione, [5-7]. Their paper described the hybrid solution, based on artificial neural networks (ANN), but their methods used ANN just as classifiers rather than dynamic models. Therefore, the application of the method would not be straight forward and need adjustment for each individual engine.

It is noticed in the literature that most reported research on FDI of open-loop internal combustion engines. And the FDI methods developed on the open-loop engines will find difficulties and problems when they are tested with the engines under closed-loop control. In further, these problems also appear when these methods were tried to be applied to real engine systems. The main differences between open-loop engine and engine under closed-loop control are lies on the following points.

Firstly, the neural network model of the engine air path needs to be trained with engine input/output data. In order to excite the engine dynamics on the whole range of frequency, for open-loop engine a random amplitude sequence (RAS) is used. However, as the input signal for engine under closed-loop control is the controller output and cannot be designed, the persistent exciting will not be achieved. One method is to superimpose a pseudorandom binary sequence (PRBS) onto the control signal. To excite nonlinear dynamics of a nonlinear dynamic system such as engines, both amplitude and frequency of the excitation signal should be considered for their wide range. As the diverged amplitude can be achieved by the variable amplitude of the control signal, while the diverged frequency achieved by the PRBS, the above stated signal will be used in this research.

Secondly, the RAS signal applied to set-point of the closed-loop system has been considered in this research, and then the generated control signal is applied to the engine for data excitation. This is from the fact that in practice the disturbance, the throttle angle change, is often a step change, which can be simulated by a RAS signal. On other hand, the control signal would be a filtered signal (as the controller could be seen as a filter) and lost the deep edge of the step signal, so the high frequency part would be lost. However, in practice it is just this type of signal dominated the operation process, and therefore, the training of the NN model with these data would be reasonable.

Thirdly, the output of some sensors may be used as feedback signal in the closed-loop control. This will change the signal pattern completely. When a sensor fault occurs, the sensor output is feedback to change the error signal, so that the engine output is regulated to a value such that the sensor output nearly equivalent to the set-point. This makes the deviation of the sensor output from nominal value quickly vanishes, and consequently the detection on this signal would become more difficult.

In this paper, a new fault detection and isolation method

will be implemented by using mean value engine model when this model is under close loop control system. An independent radial basis function neural network (RBFNN) model is used to model a dynamic system. Feed-back (FB) and feed-forward (FF) control methods will be applied to the mean value engine model. Firstly, fuel injection value corresponding to angle position value will be defined by using mean value engine model (MVEM). Secondly, PID controller will be used in the feed-back to adjust the difference between the requested and the actual air fuel ratio by compensating the feed-forward controller output. Thirdly RBF neural network models will be used to predict MVEM outputs. The model prediction error is used to generate residual for fault detection. Fault isolation will not be discussed in this paper due to the limited space. The proposed method is applied to air path dynamics of the IC engines using a well-know benchmark simulation model: MVEM [8]. Different types of fault were simulated on the MVEM, and then the input/output data were used to detect and isolate these faults.

II. ENGINE DYNAMICS

In this work, the fault detection method is developed and evaluated based on the MVEM [8] of which a block diagram is shown in Fig.1.

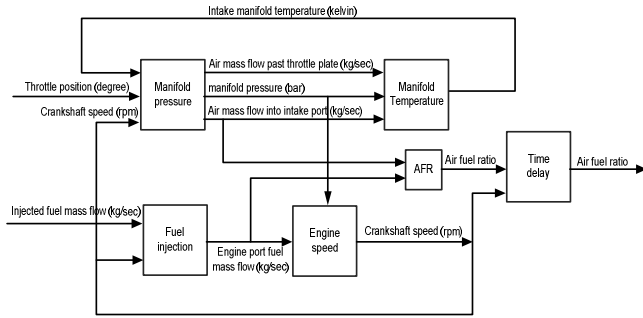


Fig.1 Mean value engine model

A. Manifold Filling Dynamics

It includes two nonlinear differential equations: one for the manifold pressure and the other for the manifold temperature. The manifold pressure is described as

$$\dot{p}_i = \frac{RT_i}{V_i} (-\dot{m}_{at} + \dot{m}_{ap} + \dot{m}_{EGR}) \quad (1)$$

Where \dot{p}_i is absolute manifold pressure (bar), \dot{m}_{at} is air mass flow past throttle plate (kg/sec), \dot{m}_{ap} is air mass flow into intake port (kg/sec), \dot{m}_{EGR} is EGR mass flow (kg/sec), T_i is intake manifold temperature in Kelvin, V_i is (manifold & port passage) volume (m^3) and R is gas constant (287×10^{-5}). The manifold temperature dynamics are described by the differential equation (Hendricks, et al, 2000)

$$\dot{T}_i = \frac{R T_i}{P_i V_i} [-\dot{m}_{ap}(k-1) T_i + \dot{m}_{at}(k T_a - T_i) + \dot{m}_{EGR}(k T_{EGR} - T_i)] \quad (2)$$

B. Crankshaft Speed Dynamic

The crankshaft dynamics is derived using conservation of rotational energy on the crankshaft (Hendricks, et al, 2000).

$$\dot{n} = -\frac{1}{I_n} (P_f(p_i, n) + P_p(p_i, n) + P_b(n)) + \frac{1}{I_n} H_u \eta_i(p_i, n, \lambda) \dot{m}_f(t - \Delta\tau_d) \quad (3)$$

Both the friction power P_f and the pumping power P_p are related with the manifold pressure p_i and the crankshaft speed n . The load power P_b is a function of the crankshaft speed n only. The volumetric efficiency η_i is a function of the manifold pressure p_i , the crankshaft speed n and the air/fuel ratio AFR . Where I is the scaled moment of inertia of the engine and its load and where the mean injection/torque time delay has been taken into account with the variable $\Delta\tau_d$.

C. Fuel injection dynamics

According to the identification experiments with an SI engine carried out by Hendricks et al. (Hendricks, et al, 2000), the fuel flow dynamics could be described as

$$\dot{m}_{ff} = \frac{1}{\tau_f} (-\dot{m}_{ff} + X_f \dot{m}_{fi}) \quad (4)$$

$$\dot{m}_{fv} = (1 - X_f) \dot{m}_{fi} \quad (5)$$

$$\dot{m}_f = \dot{m}_{fv} + \dot{m}_{ff} \quad (6)$$

Where \dot{m}_{ff} is fuel film mass flow, \dot{m}_{fi} is injected fuel mass flow, \dot{m}_{fv} is fuel vapor mass flow. The model is based on keeping track of the fuel mass flow. The parameters in the model are the time constant τ_f for fuel evaporation, and the proportion X_f of the fuel which is deposited on the intake manifold or close to the intake valves.

III. CLOSED-LOOP CONTROL DESIGN

Fig.2 shows the block diagram for the automatic control loop for the MVEM including feed-forward and feed-back controllers. Where the MVEM control input u is the injected fuel mass \dot{m}_{fi} and the disturbance input ϕ is the throttle angle position. The feed-forward controller that correlates the steady state value between the MVEM control input \dot{m}_{fi} and the disturbance ϕ will be used in the feed-forward path. In order to achieve better transient response, feed-forward and feed-back controller will be designed as following.

A. FF+FB control design

The feed-forward controller will be implemented by look-up table configuration. The data of this table were determined from the MVEM by using simulation test. The working range for the throttle angle has been given to the MVEM starting from 20 to 60 degree by step 5 degree in order to cover 9 cases. Secondly, the gain k has been changed for each case to adjust the air fuel ratio equal to 14.7. Finally, the suitable corresponding injected fuel mass can be determined for each throttle angle value by using equation 7 (The data in the look-up-table is omitted here for simplicity).

$$\dot{m}_{fi} = k \times \phi \quad (7)$$

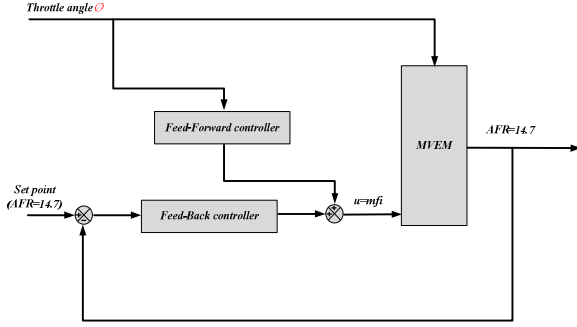


Fig.2 feed-forward plus feedback control of the MVEM

The feedback control used a PI controller that is designed using the trial and error method. The design procedure is not presented here due to the limited space.

B. Evaluation

A new set of square signal as shown in Fig.3 was used as the throttle angle. The range of this excitation signals was bounded between 20 and 60 degrees. This almost covers the whole throttle angle position in normal operation condition. The AFR response of the engine under the developed FF+FB control is displayed in Fig.4. From Fig.12 it can be seen that the obtained results are very accurate and the AFR equal to 14.7 at the steady state and is acceptable at the transient state. The other outputs of MVEM are crankshaft speed, manifold pressure and temperature. All these outputs will be used for the RBF model training.

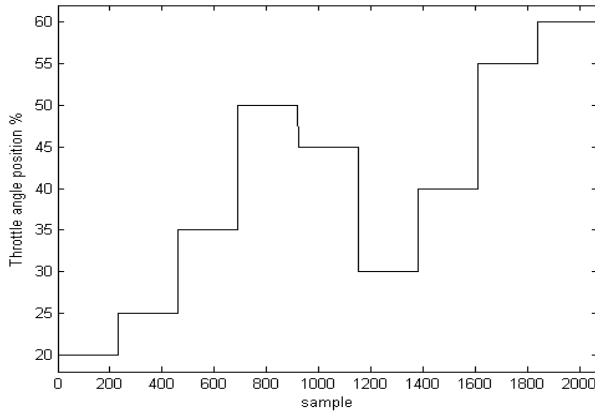


Fig.3 Change of throttle angle

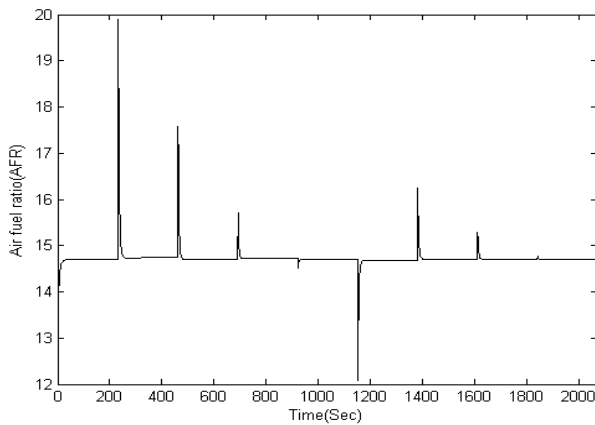


Fig.4 AFR response under FF+FB control

IV. RESIDUAL GENERATION

A. Engine modelling using RBF

A radial basis function (RBF) neural network is used to model the SI engines, because the weights of RBF are linearly related to the objective function, so that any linear optimisation algorithm can be used to train the network weights and the training is very fast.

When the Gaussian function is chosen, the output of the network can be calculated according to (8)-(9).

$$y = W * \phi \quad (8)$$

where W is the weight matrix, $\phi \in \mathbb{R}^{n_h}$ is the hidden layer output vector with its i^{th} entry given as

$$\phi_i = e^{-\frac{\|x-c_i\|^2}{\sigma^2}}, \quad i = 1, \dots, n_h \quad (9)$$

where x is the input vector, c is the centre matrix, and σ is the width of the Gaussian function. Here centre and width are chosen using the K-means clustering method, while the weights are trained using the Recursive Least Squares algorithm.

There are two different modes of modelling a dynamic system using neural networks, one is the dependent mode and the other is the independent mode. The structure of the independent mode is displayed in Fig.5.

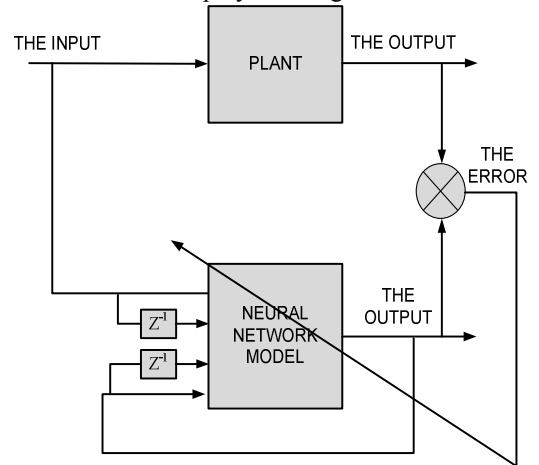


Fig. 5 The structure of independent model

In this mode the model output, instead of process output, is fed-back as part of model input in the independent mode. Then, such a model can run independently of the process.

The first step in the engine modelling by using RBFNN is the generation of a suitable training data set. As the training data will influence the accuracy of the neural network modelling performance, the objective of experiment design on training data is to make the measured data become maximally informative, subject to constraints that may be at hand. As mentioned above, a set of random amplitude signals (RAS) were designed for the throttle angle position to obtain a representative set of input data. The sample time of 0.1 sec was used.

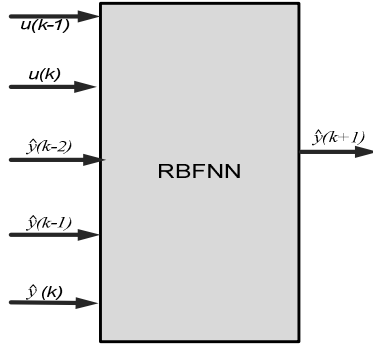


Fig.6 RBFNN model structure

The second step is to determine the input variables of the RBF model. The SI engine to be modeled has two input variables: throttle angle and the output of the PID controller which is fuel flow rate, and four outputs: air manifold temperature, air manifold pressure, crank shaft rotary speed and air fuel ratio. According to the dynamics in (1)-(6) and modeling trials, the network input that generated the smallest modeling errors was selected as first order for process input and second order for process output as shown in Fig.6.

As selected above, the RBF model has 16 inputs and 4 outputs. The hidden layer nodes have been selected as 15. Before the training, 15 centres were chosen using the K-means clustering algorithm, and the width σ was chosen using the p -nearest-neighbours algorithm. All Gaussian functions in the 15 hidden layer nodes used the same width. For training the weights W the recursive least squares algorithm (Zhai, *et al.*, 2007) was applied and the following initial values were used: $\mu = 0.98$, $w(0) = 1.0 \times 10^{-6} \times U_{(nh \times 4)}$, $P(0) = 1.0 \times 10^8 \times I_{(nh)}$, where μ is the forgetting factor, I is an identity matrix and U is the matrix with all element unity, n_h is the number of hidden layer nodes.

Totally a data set with 2070 samples was collected from the MVEM. Before training and testing, the raw data is scaled linearly into the range of [0 1]. Figs.7~10 show the model validation results of the 500 samples in the test data set. It can be seen that there is a good match between the two outputs with a very small error, in general. The modelling error of the test data set is smaller than the training data set (not shown). The mean absolute error (MAE) index is used to evaluate the modelling effects. For this model the MAE values of crankshaft speed, manifold pressure, manifold temperature and air fuel ratio are 0.0038, 0.0942, 0.0027 and 0.0029 respectively.

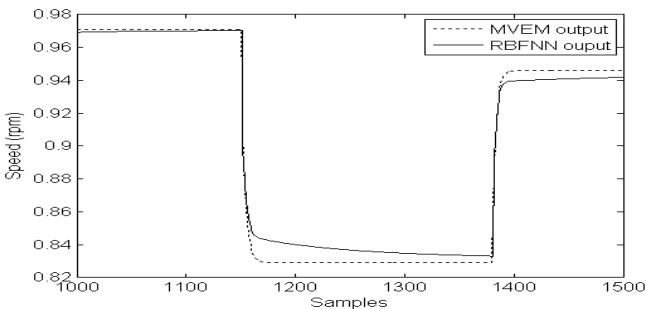


Fig.7 Model and engine outputs for crankshaft speed

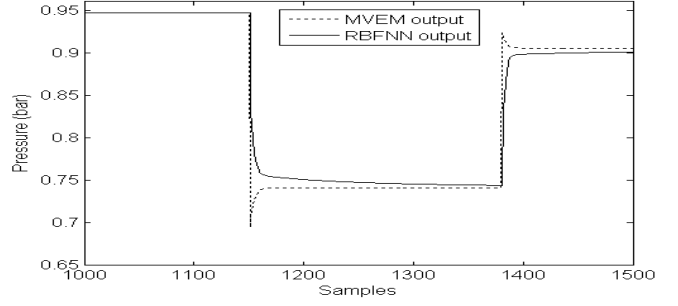


Fig.8 Model and engine outputs for air manifold pressure

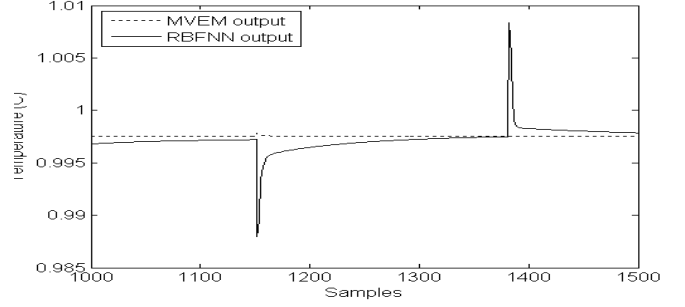


Fig.9 Model and engine outputs for air manifold temperature

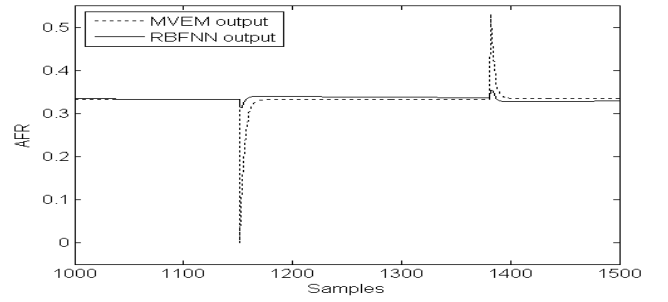


Fig.10 Model and engine outputs for AFR

B. Residual Generation

The residual is generated in the following way. Firstly, an independent neural network model is trained with engine data collected from the engine without any fault, which is called healthy condition. Then, the model is used in parallel to the engine in on-line mode to predict engine output. The modelling error between the engine output and model prediction will be used as residual signal. Thus, if no fault occurs in the engine system, the residual is just modelling error caused by noise and model-plant mismatch. When any fault occurs, the engine output will be affected and will deviate from the nominal values, while the model prediction will not be affected by the fault. So, the residual will have a significant deviation from zero caused by faults. The proposed strategy can be applied to different systems to detect dynamic faults. In this study, the residual \mathcal{E} is generated as the sum-squared filtered modelling error as follows,

$$e(k) = [y(k) - \hat{y}(k)] \quad (10)$$

$$e_{\text{filtered}}(k) = \mu e_{\text{filtered}}(k-1) + (1-\mu)e(k) \quad (11)$$

$$\mathcal{E}(k) = \sqrt{e_{\text{filtered}}^T e_{\text{filtered}}} \quad (12)$$

V. APPLICATION TO ENGINES

A. Simulating Faults

Before the developed method is tested on a real engine with real faults, it was tested in this research on the nonlinear simulation of SI engines, the MVEM with different faults simulated on it. One component fault, one actuator fault and four sensor faults with different levels of intensity have been investigated as practical examples of spark ignition (SI) engine faults. The component fault is air leakage in the intake manifold. The actuator fault is a malfunction of the fuel injector. The four sensor faults are malfunction of the intake manifold pressure sensor, manifold temperature sensor, crank shaft speed sensor and air fuel ratio sensor. Details of the simulation of these faults are described as follows.

To collect the engine data subjected to the air leakage fault, equation (1) of the manifold pressure is modified to equation (13):

$$\dot{p}_i = \frac{RT_i}{V_i} (-\dot{m}_{at} + \dot{m}_{ap} + \dot{m}_{EGR} - \Delta l) \quad (13)$$

where \dot{p}_i is the absolute manifold pressure (bar), \dot{m}_{at} is the air mass flow rate past throttle plate (kg/sec), \dot{m}_{ap} is the air mass flow rate into the intake port (kg/sec), \dot{m}_{EGR} is the EGR mass flow rate (kg/sec). The added term Δl is used to simulate the leakage from the air manifold, which is subtracted to increase the air outflow from the intake manifold. $\Delta l = 0$ represents no air leak in the intake manifold. The air leakage level is simulated as 10% of total air intake in the intake manifold. This fault occurs from the sample number 1450~1500 in the faulty data as shown in Fig.11, and was simulated by changing the Simulink model of the MVEM.

For SI engines, the target is to achieve an air–fuel mixture with a ratio of 14.7 kg air to 1 kg fuel. This means the normal value of air fuel ratio is 14.7. Because any mixture less than 14.7 to 1 is considered to be a rich mixture, any more than 14.7 to 1 is a lean mixture. Lean mixture causes the efficiency of the engine reduced, while rich mixture will cause emission increased. The fuel injector is controlled by the controller with correct amount of fuel. If the fuel injector has any fault the injected fuel amount will not be correct and affect the air/fuel ratio. Here, the malfunction of the fuel inject is simulated by reducing the injected fuel amount of 15% of the total fuel mass flow rate between the sample number 1750 and 1800 as shown in Fig.11. This fault is also simulated by changing the Simulink model of the MVEM.

The four sensor faults considered are for the crankshaft speed, manifold pressure, manifold temperature and air fuel ratio sensors respectively. 10%, 15%, 10% & 20% changes to these sensors are simulated in the MVEM Simulink model with a switch to control its on/off for each fault. These faults are simulated from sample numbers 250 to 300, 550 to 600, 850 to 900 and 1150 to 1200 respectively, as displayed in Fig.11.

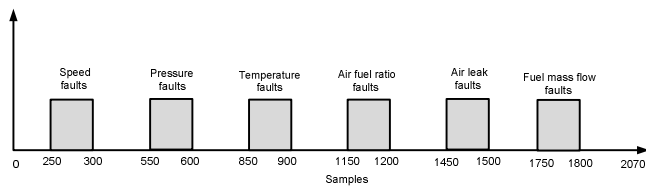


Fig.11 Time distribution of the simulated faults

B. Fault detection

The MVEM Simulink model with all 6 faults simulated was run to test the faults, while the throttle angle at different values between 20° and 60° for all the fault conditions. The sample time is chosen as 0.02 sec which is the same as that used for engine control.

Fig.12 to Fig.15 show the simulation results for the fault detection of the engine under the closed-loop control. Each figure shows model prediction error of one or two variables. The residuals are calculated from the filtered model prediction errors and are displayed in Fig.16. It is evident that all the 6 simulated faults have been clearly detected, provided not more than one fault will occur simultaneously. Also, there is not a false alarm though the model prediction error is not negligible during the transient states.

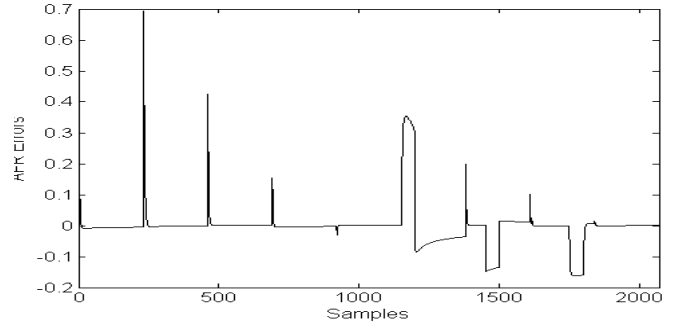


Fig.12 Error of air fuel ratio

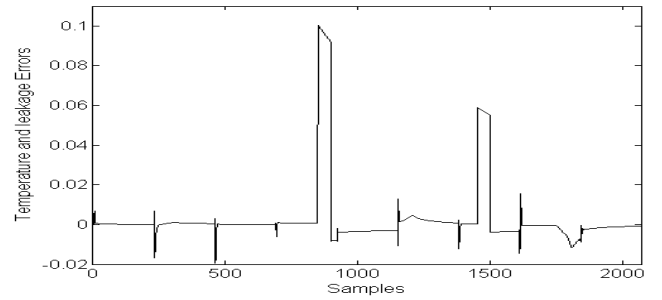


Fig.13 air manifold temperature and leakage

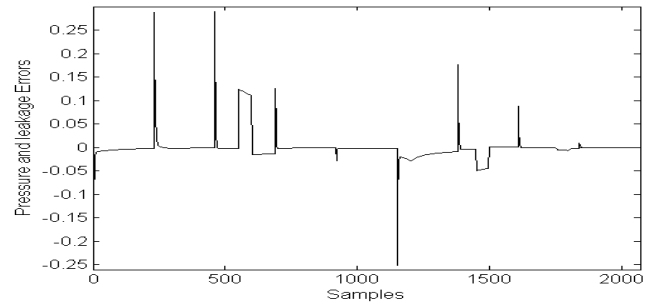


Fig.14 Error of air manifold pressure and leakage

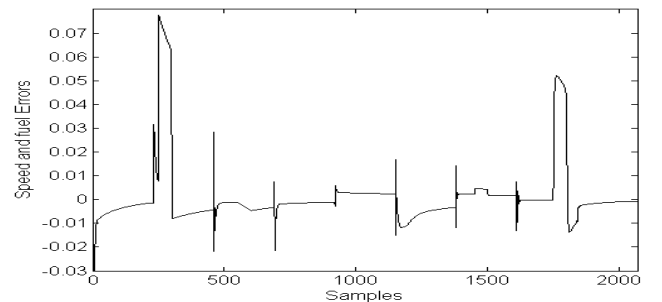


Fig.15 Error of engine speed and fuel

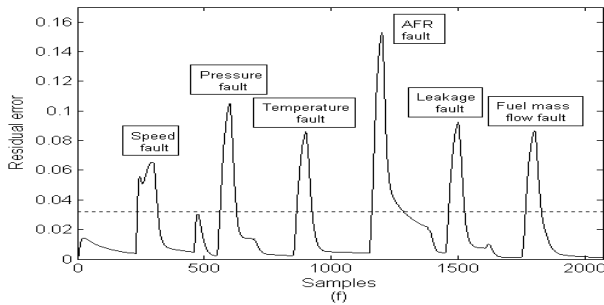


Fig.16 Residuals for all faults

VI. CONCLUSIONS

Fault detection for closed-loop control system is investigated using neural network model. The training data is acquired when the system is excited by a control variable superimposed with an RAS sequence. In this way the model is more accurate than using the data collected with normal methods. The developed method is evaluated with application to an SI engine nonlinear model. When 6 faults are simulated on the MVEM Simulink model, all the faults are detected with zero false alarm rate and zero fault miss rate. The developed method is possible to be further developed and applied to real engines.

REFERENCES

- [1] Gertler J., Costin M., Fang X., Hira R., Kowalczyk Z. and Luo Q., (1993), Model based on board fault detection and diagnosis for automotive engines, *Control Engineering Practice*, Vol. 1, No.1, pp 3-17.
- [2] Gertler J., Costin M., Fang X., Kowalczyk Z., Kunwer M., and Monajemy R., (1995), Model based diagnosis for automotive engines – algorithm development and testing on a production vehicle, *IEEE Trans. on Control Systems Technology*, Vol. 3, No.1, pp 61-69.
- [3] Blanke, M. and Patton, R.J. (1995). Industrial actuator benchmark for fault detection and isolation, *Control Engineering Practice*, Vol.3, No. 12, pp.1727-17307.
- [4] Isermann, R. (2005). Model-based fault-detection and diagnosis status and applications, *Annual Reviews in Control*, Vol.29, pp.71-85.
- [5] Capriglione, D., Liguori, C., Pianese, C. and Pietrosanto, A. (2003). On-line sensor fault detection, isolation, and accommodation in automotive engines, *IEEE Trans. on instrumentation and measurement*, Vol. 52, No. 4, pp.1182-1189.
- [6] Capriglione, D., Liguori, C., Pianese, C. and Pietrosanto, A. (2004), Analytical redundancy for sensor fault isolation and accommodation in public transportation vehicles, *IEEE Trans. on Instrumentation and Measurement*, Vol. 53, No.4, pp. 993-999.
- [7] Capriglione, D., Liguori, C. and Pietrosanto, A. (2007), Real – Time Implementation of IFDIA Scheme in Automotive Systems, *IEEE Trans. on Instrumentation and Measurement*, Vol. 56, No.3, pp. 824-830.
- [8] Hendricks, E., Engler, D. and Fam, M. (2000). A Generic Mean Value Engine Model for Spark Ignition Engines.

An Adaptive Time-frequency Filtering Algorithm for Multi-component LFM Signals based on Generalized S-transform

Dianwei Wang¹, Jing Wang², Ying Liu¹, Zhijie Xu²

¹School of Telecommunication and Information Engineering, Xi'an University of Posts and Telecommunications, Chang'an West Street, Xi'an, China, 710121.

²School of Computing and Engineering, University of Huddersfield, Queensgate, Huddersfield, UK, HD1 3DH.

D.Wang@hud.ac.uk

Abstract—Recent studies show that Cohen class bilinear time-frequency distribution methods do not have satisfactory denoising performance when analyzing multi-component LFM signals. This paper has constructed a new adaptive time-frequency filtering factor and has proposed an adaptive time-frequency filtering algorithm based on generalized S-transform. Firstly, the time-frequency distribution is obtained by transforming the time domain signals to time-frequency domain by using generalized S-transform, which is followed by calculating instantaneous frequency based on the phase information from the time-frequency distribution. Secondly, the time-frequency distribution regions occupied by clustered energy of effective signal are identified through time-frequency region extraction method and all time-frequency distribution spectrum out of the regions are removed. Thirdly, a novel TF filtering factor is constructed by the time-frequency concentration characteristic to restrain the random noise components in the regions of effective signal. Finally, the filtered signals are retrieved by using inverse generalized S-transform. Simulation results demonstrate that the proposed filtering algorithm has satisfactory performances for signal denoising which most features of original signal can be remained.

Keywords: *time-frequency filtering; multi-component LFM signal; time-frequency concentration characteristic; generalized S-transform*

I. INTRODUCTION

Linear Frequency Modulation (LFM) signal is a kind of special non-stationary signals and has been widely used in the fields of radar, sonar, communication and seismic prospecting, *etc.* [1]. The signals captured in practice are always interfered seriously by inner thermal noise, clutter and electromagnetic interference (EMI), *etc.*. It is necessary to add effective filtering mechanism before using those signals [2]. The general filtering algorithms took signals' stationarity as premise, which are only suitable for analyzing stationary signals [3]. However, the LFM signals are typical nonstationary signals with time-varied frequency characteristic. New filtering approach should be developed[4].

Time-frequency (TF) analysis theory can reveal the rules of frequency varied with temporal changes and has been widely used in nonstationary signal processing and denoising applications [5]. Commonly-used TF filtering methods include Wigner-Ville distribution (WVD), wavelet and the fractional Fourier transform (FRFT), *etc.*

While WVD is disturbed by the interference of cross terms which limits its application to multicomponent nonstationary signal processing[6]; the performance of wavelet denoising algorithms are highly related to the setting of wavelet coefficients value, which is difficult to maintain under real-world settings [7]; the FRFT requires statistical characteristic of noised signal, which is usually not applicable for most engineering applications [8].

For solving above mentioned problems, Stockwell et al. [9] had proposed S-transform, which combines the characteristics of short time Fourier transform (STFT) and the wavelet. The S-transform is a linear transform and has no cross term artifacts, which is a main superiority compared to the Cohen class bilinear TFD. With these excellent properties, S-transform has become an innovative tool in TF analysis [10] and has been applied for many fields, such as seismic signal processing, biomedical signal processing, and radar signal processing. Recently, generalized S-transform was proposed based on introducing additional parameters to adjust the window function. For example, Pinnegar proposed a TF filtering algorithm based on S-transform for nonstationary signal processing. Although the denoising performance has been proven, this TF filtering method cannot remove background noise[11]. Chen et al. proposed a generalized S-transform and applied it to signal extracting and denoising, but it is worth noting that the method is suffered from the TF spectrum amplitude redundancy from effective signal through using fixed TF filtering factor to dispose the overall TFD [12]; George presented a two steps TF filtering method based on S-transform but it only had good performance on high signal noise ratio (SNR) signals and the denoising effect is limited by manually selected weight function [13]. Li proposed a time-frequency filtering method based on generalized S-transform and applied it to machinery fault diagnosis. However, it is not suitable for low SNR signal [14].

In this paper, we constructed a new adaptive TF filtering factor and proposed a new adaptive time-frequency algorithm for multi-component LFM signals based on the generalized S-transform. Firstly, the time domain signals are transformed to time-frequency domain by generalized S-transform and the TFD distribution is obtained. Secondly, the time-frequency distribution regions occupied by clustered energy of effective signal are identified through time-frequency region extraction method and all time-frequency distribution spectrum out

of the regions are removed. Thirdly, a novel TF filtering factor is constructed by the time-frequency concentration characteristic to restrain the random noise components in the regions of effective signal. Finally, the time domain filtered signals are retrieved by inverse generalized S-transform. Simulation results showed that the proposed algorithm had satisfactory performances for denoising and most features from original signal can be remained. The algorithm provides a new way for multi-component LFM signals processing and denoising.

II. BASIC THEORY OF GENERALIZED S-TRANSFORM

S-transform and its invert transform of a time series $h(t)$ derived by Stockwell, et al. in [9] are defined as:

$$S(\tau, f) = \int_{-\infty}^{\infty} h(t)w(t-\tau, f)e^{-i2\pi ft} dt \quad (1)$$

$$h(t) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} S(\tau, f) d\tau \right\} e^{i2\pi ft} df \quad (2)$$

where $S(\tau, f)$ is the S-transform of $h(t)$, f is the frequency and τ is the center of window function $w(t-\tau, f)$. The most commonly used $w(t-\tau, f)$ is the Gaussian window defined as:

$$w(t-\tau, f) = \frac{|f|}{\sqrt{2\pi}} e^{-\frac{f^2(\tau-t)^2}{2}} \quad (3)$$

It is worth noting that the window should be normalized:

$$\int_{-\infty}^{\infty} w(t, f) dt = 1 \quad \forall f \in IR \quad (4)$$

It is derived in [9] that the S-transform and the Fourier transform are inverse between each other. So $h(t)$ is exactly recoverable from the S-transform, and the S-transform can be calculated by using fast Fourier transform (FFT) directly. Moreover, compared with other bilinear transform, the S-transform is a linear transform which can avoid cross-terms such as the Wigner-Ville distribution.

To improve the concentration of S-transform, Pinnegar, Chen and other scholars introduced adjust parameters and reconstruct Gaussian window functions to formulate generalized S-transform, which allows the transform adaptively verifies the window functions base on the distribution features of frequency, therefore it is more effective in practice [11-12]. The generalized S-transform is defined as:

$$S(\tau, f) = \int_{-\infty}^{\infty} h(t) \frac{|\lambda_{GS}| |f|^\gamma}{\sqrt{2\pi}} e^{-\frac{\lambda_{GS}^2 f^{2\gamma} (\tau-t)^2}{2}} e^{-i2\pi ft} dt \quad (5)$$

$\gamma \in [1/2, 3/2]$

where λ_{GS} and γ are adjustment parameters. When γ is defined, the window width of the generalized S-transform broadened when λ_{GS} increasing. In order to get higher time resolution, narrower window function should be chosen but the frequency resolution would be decreased at the same time due to the Heisenberg uncertainty theory. If $\lambda_{GS} = \gamma = 1$, The generalized S-transform converts to the S-transform. The following figure 1 is the WVD of a two-

component LFM signal, and figure 2 is TFD of generalized S-transform. It is shown by figure 1 and figure 2 that the WVD of the two-component LFM signal has the best TF concentration characteristic but is seriously disturbed by cross term, however the TFD of generalized S-transform has very good TF concentration characteristic and free from cross term.

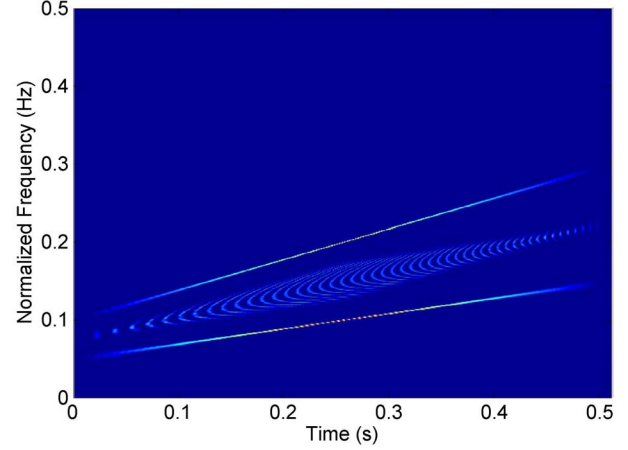


Figure 1. The WVD of the LFM signal

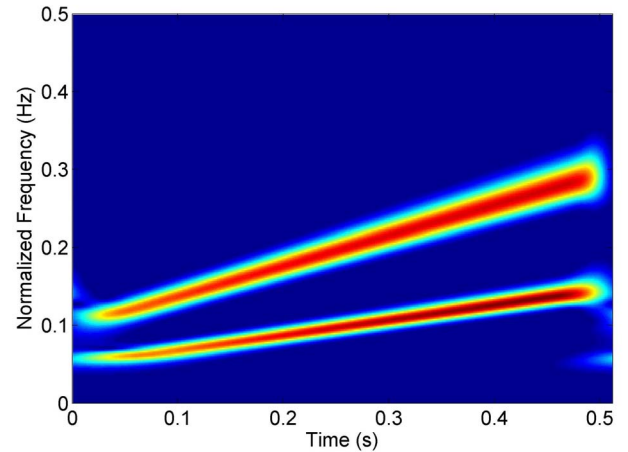


Figure 2. The generalized S-transform of the LFM signal

III. THEORY AND PROCESS OF ADAPTIVE TF FILTERING

A. The Solution of Instantaneous Frequency (IF)

The S-transform has phase information, therefore we can solve the IF from the TFD directly and reduce computation cost significantly.

Considering a multi-component LFM signal $x(t)$ as:

$$x(t) = s(t) + n(t) \quad (6)$$

where $s(t)$ is the effective signal component of original multi-component LFM signal, $n(t)$ is the noise. The translation result of $x(t)$ by generalized S-transform is $GS(\tau, f)$, which can be expressed as complex TF spectrum form:

$$GS(\tau, f) = |E_{GST}(\tau, f)| \exp[i\theta_{GST}(\tau, f)] \quad (7)$$

where $E_{GST}(\tau, f)$ is instantaneous amplitude and $\theta_{GST}(\tau, f)$ is instantaneous phase, which meet following conditions:

$$E_{GST}(\tau, f) = \sqrt{\{\text{Re}[GS(\tau, f)]\}^2 + \{\text{Im}[GS(\tau, f)]\}^2} \quad (8)$$

$$\theta_{GST}(\tau, f) = \arctan \left\{ \frac{\text{Im}[GS(\tau, f)]}{\text{Re}[GS(\tau, f)]} \right\} \quad (9)$$

IF derived from the phase information above can be expressed as

$$IF_{GST}(\tau, f) = \frac{\partial \theta_{GST}(\tau, f)}{\partial \tau} \quad (10)$$

The process of (7) - (10) is the solution process based on generalized S-transform. IF can be solved from the TFD of generalized S-transform straightforward and the computation amounts can be reduced significantly.

B. The Extraction of TF Regions of Effective Signal

The TFD of effective signal concentrated on finite regions around the IF commonly due to the finite support and edge characteristics, however the TFD of noise randomly scattered in the overall TF plane. Referenced to the extraction method proposed in [15], the TFD regions of effective signal can be identified on the TF plane.

Assuming a TFD of multi-component LFM signal $x(t)$ is $P_x(t, f)$, which contains clustered energy of effective signal, noise and interference (the generalized S-transform has no interference term). On the whole TF plane, the TF regions of effective signal are finite around the IF. Setting up a suitable energy density threshold α , the i th area occupied by TF regions of effective signal can be expressed as:

$$R_i(t, f) = \begin{cases} P_s(t, f), & P_s(t, f) \geq \alpha \\ 0, & P_s(t, f) < \alpha \end{cases} \quad (t, f) \in R \quad (11)$$

Based on the TFD characteristic, at the position of the IF, the TFD emerge an energy peak and the amplitude of $P_s(t, f)$ has local maximum value. In this paper, we used a region growing method to extract the TF regions containing effective signal on the TF plane. At first, we set the IF position point as seed. Then the seed is expanded in each direction. If the current value of $P_s(t, f)$ is greater than the threshold α but less than other values of IF at the same time, or multiple TF regions are overlapped, it should be combined these regions as a merged region. It is clear that the size of the TF regions is sensitive to the threshold of α . In this paper, an image processing method was adopted for choosing the threshold and the details can be referred to [15].

Considering that the finite TFD regions of effective signal containing the most of the energy from effective signal, and the TFD of noise is randomly scattered across the overall TF plane (only few part of them are distributed inside the TFD regions), if we remove all the TFD spectrum outside of these TFD regions, most energy of noise can be eliminated, through which a rough filtering

processes can be implemented. However, there is few part of noise's TFD regions can overlap with TFD regions of effective signal, it is necessary to add further refine filtering processes to the TFD regions. In this paper, we construct an adaptive time-frequency factor for this purpose.

C. Design of Adaptive TF Filtering Factor

It is illustrated by (11) that the filter effect is depended on the TF factor, so it is important to construct a better performed TF factor. In this paper we construct a new TF factor defined as:

$$F_f(n, k) = \frac{|P_x(n, k)|}{\xi + \max(|P_x(n, k)|)} \quad (12)$$

where $|P_x(n, k)|$ is the amplitude value of any point's TFD on the TF plane, $\max(|P_x(n, k)|)$ is the maximum amplitude value of TFD of the TF region occupied by effective signal, which corresponds to the maximum amplitude value of TF energy peak region of effective signal, ξ is an adjustable parameter of TF factor with the range $\xi \subseteq [0, 1]$. According to (12) it is shown that the limitation of filter factor's output value is 1.

The TFD of signal after being filtered is shown in Equation 13:

$$P_s(n, k) = GS[x(n)]F_f(n, k) \quad (13)$$

where $GS[x(n)]$ is the generalized S-transform of signal $x(t)$, $P_s(n, k)$ is the TFD of filtered signal after two steps TF filtering operation, $F_f(n, k)$ is the time-frequency factor. If we transform the TFD of $P_s(n, k)$ into time domain by the convert generalized S-transform, the final effective signal $\tilde{s}(n)$ after being filtered in time domain can be received as following:

$$\tilde{s}(t) = GS^{-1}[P_s(n, k)] \quad (14)$$

According to (12) it is shown that the ξ has impact on the value of TF factor which determines the performance of the TF filtering algorithm. It is highlighted by [15] that the time-frequency concentration of any point on the TF plane can gradually increase its value when approaching the IF point, which actually holds the maximum value of TF plane. So we can construct the adjustable parameter ξ by calculating the normalized distance throughout the point in the TF regions covered by effective signal with the IF point, the specific steps are as follows:

Firstly, the IF points of each TF region occupied by effective on the discrete TF plane is calculated by the method described by (10) and defined as $IF_i(n, k)$, where $i = 1, 2, \dots$ is the sequence number of TF regions;

Secondly, the distance between any (n, k) of $P_x(n, k)$ and the IF point $IF_i(n, k)$ is normalized and marked as $r(n, k)$, which is a set of distance

$\{r(n, k), n = 1, 2, \dots, N; k = 1, 2, \dots, N\}$ after this normalization.

Finally: Construct the adjustable parameter by referring the normalized set of distance as follows:

$$\xi = r_N(n, k) \quad (15)$$

According to (12) and (15), it is shown that the TFD points with higher TF concentration (nearer to the IF points) have smaller attenuation by using the TF factor, and the points with lower TF concentration (farther to the IF points) have greater attenuation on the contrary, so the proposed algorithm can suppress or remove noise and also remain most features of original signal.

IV. SIMULATION RELUS AND PERFORMANCE ANALYSIS

A. Simulation results

In order to verify the performance of noise immunity of this method, a set of two-component LFM signal has been designed with the duration time is 0.5s and the number of sampling points is 512. The initial normalized frequencies of each component are 0.05 and 0.15 and the modulation slopes are 100 and 120 respectively. The relative amplitudes of each component are equal. The waveform of the pure signal is shown in figure 3, and its TFD transformed by generalized S-transform ($p=1, \lambda_{gs}=0.5$, the same below) is illustrated in figure 4. A zero-mean Gaussian white noise is added to the simulation signal and the signal to noise ratio (SNR) is set to -5dB by selecting a suitable variance of the Gaussian white noise, and the waveform in time domain and TFD are shown in figure 5 and figure 6 respectively.

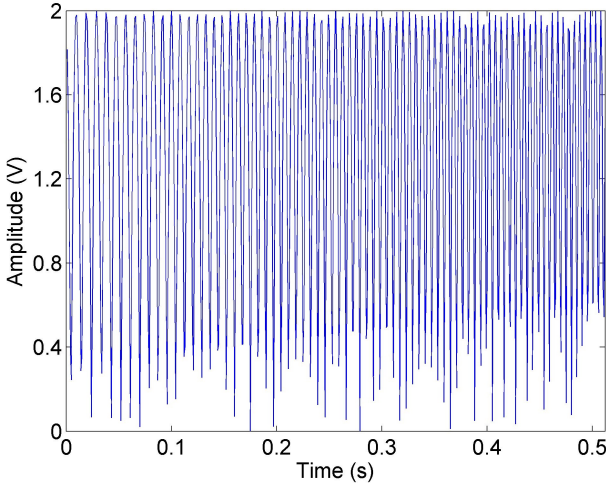


Figure 3. Waveform of the original LFM signal

It can be seen from the figure 3 that the time domain waveform of signal $x(t)$ is cluttered and irregular, and the signal-to-noise ratio (SNR) of $x(t)$ is -6.62(dB), therefore it is difficult to detect the actual signal and extract the desired feature. Figure 4 illustrated that the noise is distributed in the whole TF plane randomly and all the TFD of different components is interfered severely due to the noise. While both the basic shape and the localization characteristic of the TFD of each effective component are still remained,

which provide theoretical basis for using time-frequency filtering theory.

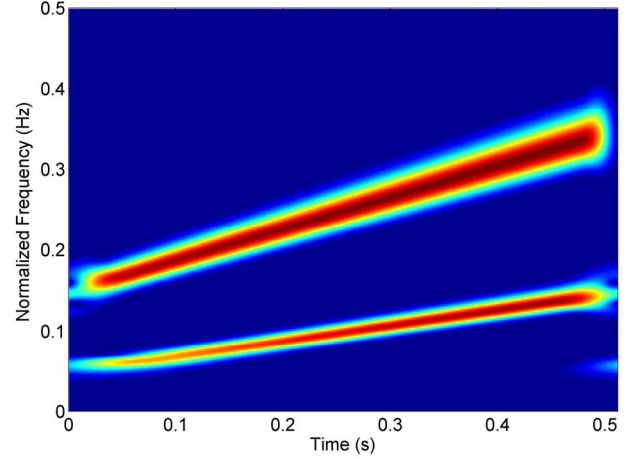


Figure 4. The generalized S-transform TFD of the original LFM signal

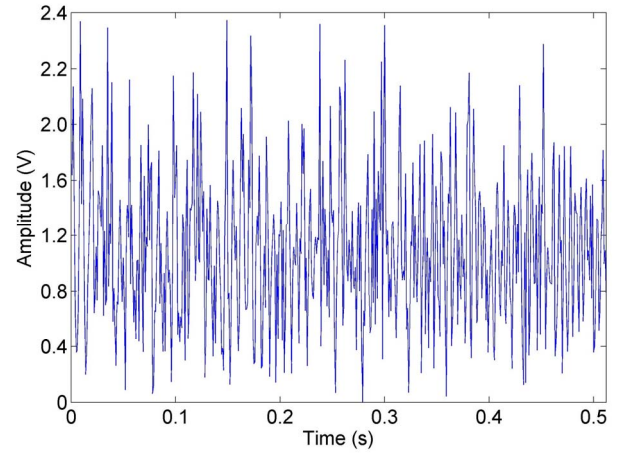


Figure 5. Waveform of the signal with noise

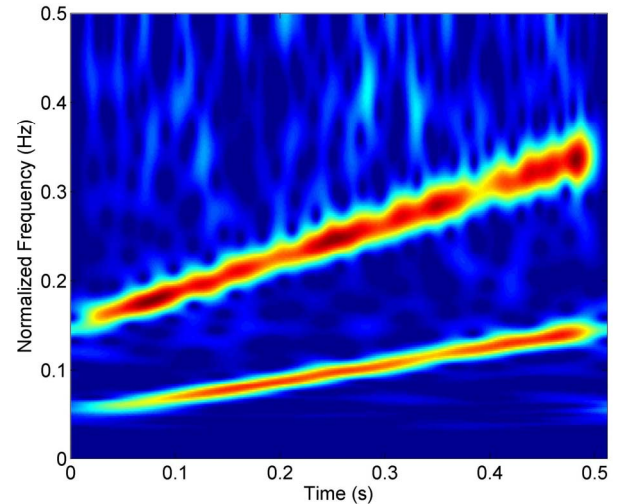


Figure 6. The generalized S-transform TFD of the LFM signal with noise

Figure 7 and figure 8 are the waveform and its TFD of the final result of signal processed by the TF filtering method proposed in this paper. The SNR of the final filtered signal rises to 3.67(dB) so the elevation amount is reach to 10.29. Furthermore both the figure 7 and 8 indicate that the waveform and its TFD of final filtered signal are very similar to the original LFM signal.

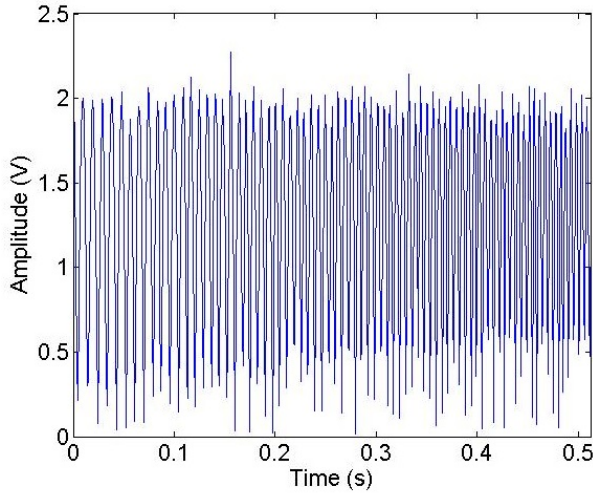


Figure 7. The waveform of filtered signal

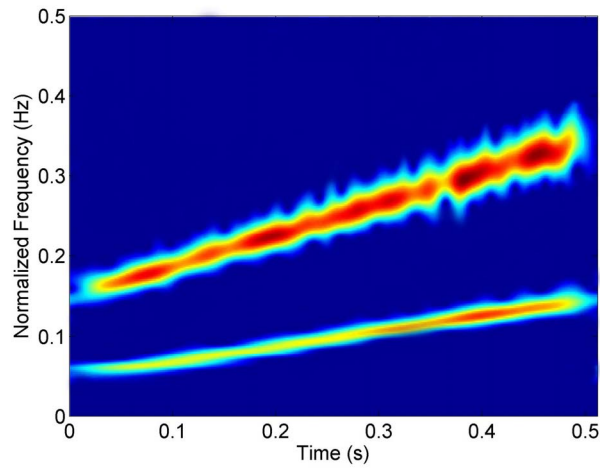


Figure 8. The TFD of filtered signal

It can be seen that the proposed adaptive TF filtering algorithm can remove the noise effectively and keep most of the characteristic of original signal. The method provides a new way for multi-component LFM signal processing.

B. Performance Analysis

To comparatively analyze the performance of the TF filtering algorithm proposed in this paper with other algorithms, another two classic TF filtering algorithm presented in [13] and [14] are selected, and two important evaluation indicators, SNR rising amount and Mean Square Error (MSE) of the original and the filtered signal are adopted in this section. In the following analysis and comparison process, the computer simulation instance and the main parameters adopted are same as section IV.A. Figure 9 shows the curves of output filtered signal's SNRo (dB) varies with the original signal's SNRi (dB).

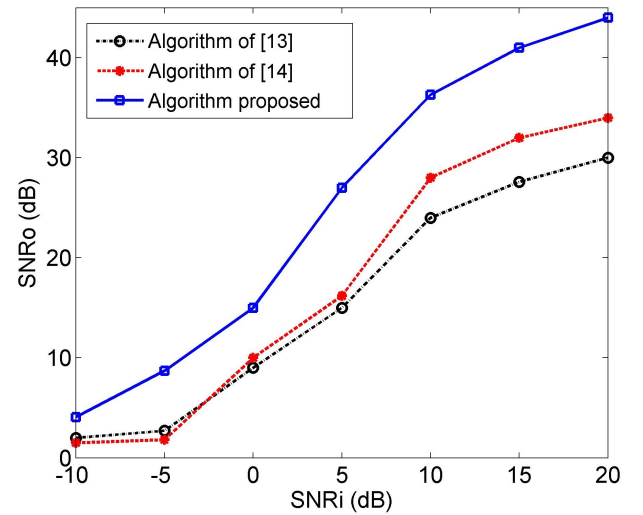


Figure 9. The variation curves of SNRo with SNRi

Figure 9 shows that the novel algorithm proposed in this paper has better SNR rising performance than others, especially in the low SNR circumstance, mainly because it remove the TFD of the noise outside the regions of effective components entirely and construct a new TF filtering factor by using the TF concentration characteristic, which can restrain the TFD spectrum of noise in the TF regions occupied by effective signal.

Another important evaluation indicator is the bias of the original and the filtered signal which reflects the size of the distortion. The smaller distortion means the filtered signal can remain more original information. As introduced in [16], the MSE rule is adopted in this section. Table 1 is the different MSE values processed by these three algorithms in different SNR conditions.

Table 1 shows that the MSE of the TF filtering algorithm proposed in this paper is smaller than others in each SNR condition, which has proven the effectiveness of proposed method.

TABLE 1. THE MSE OF THE ORIGINAL AND THE FILTERED SIGNAL IN DIFFERENT SNR

SNR/dB	MSE		
	Algorithm of [13]	Algorithm of [14]	Algorithm Proposed
-5	0.1606	0.2038	0.1315
0	0.1173	0.1951	0.0997
5	0.0985	0.1294	0.0706
10	0.0417	0.0655	0.0191
15	0.0109	0.0471	0.0083
20	0.0072	0.0143	0.0009

V. CONCLUSION AND DISCUSSION

This paper presented an adaptive TF filtering method based on generalized S-transform and applied the algorithm to the denoising processing for multi-component LFM signals. Simulation results illustrated that the method proposed in this paper had satisfactory performance in SNR rising and features remaining against other popular approaches, which also offers a new ideas to nonstationary

signal processing. However if the SNR of the signals is too low, then the TFD regions of effective signal will be difficult to extracted, and the denoising performance of the proposed TF filtering algorithm will be degraded. Further researches should be made on how to develop new TF regions extraction method and construct more effective TF filtering factor to solve such problems existing.

ACKNOWLEDGMENT

This work is supported by the 2014 Special Foundational Research Project for Intensifying the Police with Science and Technology of Ministry of Public Security, People's Republic of China (Fund No. 2014GABJC024, 2014GABJC022, 2014GABJC023), 2014 Special Scientific Research Project of the Education Department of Shaanxi Province (Fund No. 14JK1680) and 2015 The Natural Science Foundation Research Project of Shaanxi Province (Fund No.2015JM6350).

REFERENCES

- [1] Huiyan Hao, "Multi component LFM signal detection and parameter estimation based on EEMD-FRFT," *Optik - International Journal for Light and Electron Optics*, vol.124, pp. 6093-6096, December 2013.
- [2] Ding L F, Geng F L. Principle of radar, 4th ed., Xi'an: Publishing Houses of Xidian University, 2009.
- [3] Samantaray S.R., Dash P.K. "Pattern recognition based digital relaying for advanced series compensated line," *International Journals of Electrical Power Energy System*, vol.30, pp.102-112, February 2008.
- [4] Shibin Wang, Xuefeng Chen, Yan Wang, et al., "Nonlinear squeezing time-frequency transform for weak signal detection," *Signal Processing*, vol.113, pp.195-210, August 2015.
- [5] Zhang X. D., Bao Z. Analysis and processing of nonstationary signal. Beijing: Tsinghua University Press. 1998.
- [6] G. Chen, J.Chen, G.M.Dong. "Chirplet Wigner-Ville distribution for time-frequency representation and its application", vol.41, pp.1-13, December 2013.
- [7] Caio F.F.C. Cunha, André T. Carvalho, Mariane R. Petraglia, Antonio C.S. Lima, "A new wavelet selection method for partial discharge denoising," vol.125, pp. 184-195, August 2015.
- [8] Soo-Chang Pei, Jian-Jiun Ding, "Fractional Fourier Transform, Wigner Distribution, and Filter Design for Stationary and Nonstationary Random Processes," *IEEE Transactions on Signal Processing*, vol.58, pp.4079-4092, August 2010.
- [9] Stockwell R.G. Mansinhal L, Lowe R.P, "Localization of the complex spectrum: the S-transform," *IEEE Transaction on Signal Processing*, vol.44, pp. 998-1001, April 1996.
- [10] S. Roopa n, S.V.Narasimhan, "S-transform based on analytic discrete cosine transform for time-frequency analysis," *Signal Processing*, vol.105, pp.207-215, December 2014.
- [11] M. Pinnegarcr, "Time-local spectral analysis for non-stationary time series: the S-transform for noisy signal," *Fluctuation and Noise Letters*, vol.3, pp.357-364, September 2003.
- [12] Chen X. H., He Z. H., Huang J. D, "Generalized S transform and its time-frequency filtering," *Signal Processing*, vol.24, pp.28-31, January 2008.
- [13] George N V, Sahu S S, Mansinha L, et al., "Time Localised Band Filtering Using Modified S-Transform," *Proc. International Conference on Signal Processing Systems*, Singapore, pp.42-46, May 2009.
- [14] X. W. Li, J. H. Yang, M. Li and J. W. Xu, "A time-frequency filtering method based on generalized S transform and its application in machinery fault diagnosis," *Applied Mechanics and Materials*, vol.157-158, February 2012.
- [15] Tian Guangming, Chen Guangyu, "Signal estimation based on time-frequency filtering specified by time-frequency regions occupied by clustered energy," *Signal Processing*, vol.20, pp.263-267, June 2004.
- [16] Zhang Y., Wang X. Q., Peng Y. N, "Adaptive center weighted modified trimmed mean filter," *Journal of Tsinghua University (SCI & Tech)*, vol.39, pp.76-78, September 1999.

Applying Feedback to Stock Trading: Exploring A New Field of Research

T C Yang, Z G Li and Y N Shu
Department of Engineering and Design
University of Sussex
Brighton, UK
t.c.yang@sussex.ac.uk

Abstract— Nowadays, feedback control is everywhere and affects everything. Recently, a novel idea of applying feedback control to stock trading is proposed. Some interesting results are published in recent IEEE/IFAC control conferences. Different from current statistical or artificial neural network based approaches, this opens a new research field in analytical strategies for stock trading. In this paper we follow the scheme of “Simultaneous Long-Short feedback control” and carry out a new study. In our study, instead of one feedback gain for both long and short trading, different trading gains are applied. Based on recent stock prices over one year (from 2014/3/27 to 2015/3/26), we show our selected simulation results for two stocks. One is the Apple Incorporated stock. The trend of its price over the year is increasing. The other is the NASDAQ-100 index and the trend of its price is decreasing. A number of new research topics are also proposed in this paper.

Keywords- Feedback control, Stock trading

I. INTRODUCTION

The evolution of feedback control from an ancient technology to a modern field is a fascinating microcosm of the growth of the modern technological society. Nowadays, feedback control is everywhere and affects everything [1], from generation and distribution of electricity, telecommunication, process control, steering of ships, control of vehicles and airplanes, operation of production and inventory systems, regulation of packet flows in the Internet, modelling and control of economic systems, to disrupting the feedback of harmful biological pathways that cause disease. Recently, a novel idea of applying feedback control to stock trading is studied and some interesting results published [2-11]. Although such an idea is not completely new, the recent pinioning work □ see [2], the first paper in this field and presented in the 2008 IFAC World conference □ was attributed to Prof Barmish, B. Ross (The author of a well-known book: “New Tools for Robustness of Linear Systems” published in 1993) and his team. Their work lays a foundation for many future researches yet to follow. In this paper, in Section 2 we introduce Simultaneous Long-Short (SLS) feedback control. Among the all work presented by Prof Barmish’s team [2-11], this is a key stock trading scheme. In Section 3, we present our work summarized in the

abstract and propose a number of new research topics. A brief conclusion is given in Section 4.

The work presented in [2-11] and here is an interdisciplinary study crossing the areas of control and finance. People in control community may not familiar with some financial terms. For some background knowledge, the interested readers are suggested to refer to the Wikipedia or a book [12] (There were three ACC and CDC tutorial sessions on this subject [13-15]).

II. SIMULTANEOUS LONG-SHORT (SLS) FEEDBACK CONTROL

II.1 Notations and the Idealized Market [2]

In consistent with and quoted from [4], in this paper $p(t)$ is used for the stock price, $I(t)$ for the amount invested with $I(t) < 0$ being a short sale, $g(t)$ and $V(t)$ for the cumulative trading profit or loss on $[0, t]$ and account value respectively. When speaking of a “short sale” above, we mean the following: When the trader has negative investment $I(t) < 0$, this means stock is borrowed from the broker and is immediately sold in the market in the hope that the price will decline. When such a decline occurs, this short seller can realize a profit by buying back the stock and returning the borrowed shares to the broker. Alternatively, if the stock price increases, the short seller may choose to “cover” the trade, return the borrowed stock to the broker and take a loss. The idealized market is characterized by a number of assumptions:

(A-1) Continuous and Costless Trading: It is assumed that the trader can react instantaneously to observed price variations with zero transaction cost; i.e., no brokerage commissions or fees. That is, the amount invested $I(t)$ can be continuously updated as price changes occur. Motivation for this assumption is derived in part from the world of high-frequency trading; e.g., with the help of programmed trading algorithms, flash traders working for hedge funds can execute literally thousands of trades per second with minimal brokerage costs. In fact, even the small trader using a high-speed internet connection can easily execute many trades per minute. This assumption is also made in the celebrated Black-Scholes model; e.g., see [16].

(A-2) Continuously Differentiable Prices: The stock price $p(t)$ is continuously differentiable on $[0, T]$, the time interval of interest. It should be noted that this is the most serious assumption differentiating the idealized market from a real market. That is, this assumption rules out the possibility that price gaps may occur following various real market events such as earnings announcements or major news. In contrast to most of the finance literature, instead of making predictions based on a geometric Brownian motion model for price, $p(t)$ is treated in this paper as an uncertain external input against which the aim is to robustify the trading gain $g(t)$.

(A-3) Perfect Liquidity: It is assumed that the trader faces no gap between a stock's bid and ask prices. That is, orders are filled instantaneously at the market price $p(t)$. We do not view this assumption as serious in the sense that stocks trading large volume on major exchanges typically have bid-ask spreads which are small fractions of a percent.

(A-4) Trader as a Price-Taker: It is assumed that the trader is not trading sufficiently large blocks of stock so as to have an influence on the price. Note that this assumption would be faulty in the case of a large hedge or mutual fund. For example, when a hedge fund dumps millions of shares onto the market, the stock price typically declines during the course of the transaction.

(A-5) Interest Rate and Margin: It is assumed that the trader accrues interest on any uninvested account funds at the risk-free rate of return r . However, when "extra" funds are brought into the account via a short sale, consistent with the standard practice of brokers, if these funds are "held aside" as cash, no interest is accrued. If the trader opts to use this cash to obtain leverage via purchase of additional stock, margin charges accrue at interest rate m . Another way that margin charges result is when a trader has $I(t) > V(t)$. That is, the trader is essentially being given a loan by the broker and charged margin interest rate m for the use of the funds. To avoid distracting technical details regarding the way brokers "mark to market" to calculate margin charges, the following model used in this paper is a simplification on the way margin accounts are typically handled: When $|I(t)| > V(t)$, margin charges are compounded at interest rate m . Note that the absolute value used for $I(t)$ takes care of $I(t) < 0$ when a short is involved. Finally, for simplicity, we assume that both interest and margin rates are the same. That is, $m = r$. This is a type of efficient market assumption. In practice, it would usually be the case that $m > r$ with the difference $m - r$ being a function of the size of the trader. For example, a large brokerage house trading its own portfolio would have virtually no spread between these two interest rates. With this assumption, we have a very simple equation summarizing interest accruals and margin charges accruals. That is, over time interval $[0, t]$, the accumulated interest $i(t)$ is:

$$i(t) = \int_0^t (V(\tau) - |I(\tau)|) d\tau \quad (1)$$

$i(t) < 0$ represents margin interest owed.

(A-5-S) In this paper, in order to concentrate on the fundamentals of the trading algorithm, we make a special assumption

$$m = r = 0. \quad (2)$$

(A-6) Simplified Collateral Requirement: In a brokerage account, associated with the granting of margin is a collateral requirement on securities. For example, if the account value is V , some clients are allowed to carry $2V$ in equities before forced liquidation of assets occur, larger clients may have larger upper bounds, etc. More generally, our model assumes $\gamma \geq 1$ is specified as part of the trading scenario. Then, we only allow instantaneous investments

$$|I(t)| \leq \gamma V(t) \quad (3)$$

for satisfaction of collateral requirements.

II.2 Dynamics and State Equations [4]

Consider an infinitesimal time increment dt over which both the trading gain g and the account value V are updated. Letting dp be the corresponding stock price increment, the corresponding incremental trading gain is simply the percentage change in price multiplied by the amount invested. Hence,

$$dg = \frac{dp}{p} I \quad (4)$$

During this same time period, the incremental change in the account value is the sum of the contributions from both stock and idle or borrowed cash. That is,

$$dV = dg + r(V - |I|)dt \quad (5)$$

with the starting point above, for the trading profit or loss, the differential equation is

$$\frac{dg}{dt} = \frac{1}{p} \frac{dp}{dt} I(t) \quad (6)$$

and the correspond account value equation is

$$\frac{dV}{dt} = \frac{dg}{dt} + r(V - |I(t)|) \quad (7)$$

with these equations having initial conditions

$$\begin{aligned} V_0 &= V(0) \geq I(0) = I_0 \\ g(0) &= 0 \end{aligned} \quad (8)$$

As noted before, for the differential equations above, we view the price variation $p(t)$ as an external input. The investment $I(t)$ plays the role of the controller and is yet to be specified.

Using notation:

$$\rho(t) \equiv \frac{1}{p} \frac{dp}{dt} \quad (9)$$

and substituting (9) and (6) to (7):

$$\frac{dV}{dt} = \rho(t)I(t) + r(V(t) - |I(t)|) \quad (10)$$

For this equation, consider time intervals of two types.

Type 1 Intervals: On such an interval $[t_1, t_2]$ we have $|\rho(t)| \geq r$. Taking the controller to be of the form

$$I(t) = \gamma^* V(t) \operatorname{sgn} \rho(t) \quad (11)$$

with $0 \leq \gamma^* \leq \gamma$, the resulting account value equation is

$$\frac{dV}{dt} = [\gamma^* |\rho(t)| + r(1 - \gamma^*)] V(t) \quad (12)$$

and the associated endpoint solution is readily calculated to be

$$V(t_2) = \exp\left(r(1 - \gamma^*)(t_2 - t_1) + \gamma^* \int_{t_1}^{t_2} |\rho(\tau)| d\tau\right) V(t_1) \quad (13)$$

Type 2 Intervals: On such an interval, $[t_1, t_2]$, we have $|\rho(t)| < r$. In this case, taking the controller to be $I(t) \equiv 0$, the account value equation degenerates to

$$\frac{dV}{dt} = rV(t) \quad (14)$$

with endpoint solution given by

$$V(t_2) = \exp(r(t_2 - t_1)) V(t_1) \quad (15)$$

From the analysis for the two types of intervals above, it follows that if the idealized market trader uses , the inequality

$$r(1 - \gamma^*)(t_2 - t_1) + \gamma^* \int_{t_1}^{t_2} |\rho(\tau)| d\tau \geq 0 \quad (16)$$

is satisfied and it follows that

$$V(t_2) \geq V(t_1) \quad (17)$$

The cumulative trading profit or loss $g(t)$ is considered as the system output and a static feedback control law of the form $I = f(g)$ with f being a continuous function is used in [4]. That is, the amount invested $I(t)$ in the stock is modulated as a function of the trading profits or losses $g(t)$ accrued over $[0, t]$. In the sequel, the focus is on time-invariant linear feedback controls

$$f(g) = I_0 + Kg \quad (18)$$

with $I_0 = I(0)$ being the initial investment.

II.3 Simultaneous Long-Short (SLS) Linear Feedback Control [4]

To establish the main result, first to construct a controller which is a superposition of two linear feedbacks as described by Eq. (18), one being a long trade with $I_0, K > 0$ and the other being a short trade with $I_0, K < 0$. These trades can be viewed as running simultaneously in parallel. The amount invested in the long trade is $I_L(t)$ and the amount invested in the short trade is $I_S(t)$. Hence, the net overall investment is

$$I(t) = I_L(t) + I_S(t) \quad (19)$$

and, as time evolves, the relative amounts in each of these trades will change. It may well be the case that one of these two trades will become “dominant” as time evolves. For example, in a raging bull market, one would expect to see get large $I_L(t)$ and $I_S(t)$ tending to zero. With $K > 0$ and $I_0 > 0$ fixed, the two feedback controllers are defined by

$$I_L(t) = I_0 + Kg_L(t) \quad (20)$$

$$I_S(t) = -I_0 - Kg_S(t) \quad (21)$$

where $g_L(t)$ and $g_S(t)$ are the trading gains or losses for the long and short trades respectively. Hence, the overall investment and trading gains for the combined trade are

$$I(t) = K(g_L(t) - g_S(t)) \quad (22)$$

$$g(t) = g_L(t) + g_S(t) \quad (23)$$

which begins at $I(0) = I_0$ and $g(0) = 0$. Refer to Eq. (4) and (20), the individual trades satisfy the differential equation

$$\frac{dg_L}{dt} = \rho(t)(I_0 + Kg_L) \quad (24)$$

$$\frac{dg_S}{dt} = -\rho(t)(I_0 + Kg_S) \quad (25)$$

with initial conditions $g_L(0) = 0$ and $g_S(0) = 0$.

The setup above leads to many results which are consistent with common sense. For example, with $K > 0$ and price $p(t)$ increasing, will increase and $|I_S(t)|$ will decrease. That is, the trader becomes “net long”. The question then arises whether the trading gains from the long position will be sufficient to offset losses from the short leading to a net profit. The *Arbitrage Theorem* [4] to follow answers this question and others in the affirmative provided the so-called “adequate resource condition” below [4] is satisfied. Given that an idealized

market is assumed, this framework should be viewed as one which shows us the limits of state feedback control.

Adequate Resource Condition [4]: In the theorem follow, it is assumed that the combination of initial account value $V(0) = V_0$, feedback gain K and constant and prices $p(t)$ are such that the collateral requirement $|I(t)| \leq \gamma V(t)$ is assured over the time interval of interest. Equivalently, if the investment $I(t)$ demands more resources than are currently available in the account, the trader has the ability to respond to a margin call by bringing in more funds.

Arbitrage Theorem (see [4] for the proof): At all times $t \geq 0$, assume the adequate resource condition $I(t) \equiv 0$ is satisfied. Then, the Simultaneous Long-Short static linear feedback controller leads to trading profit

$$g(t) = \frac{I_0}{K} \left[\left(\frac{p(t)}{p(0)} \right)^K + \left(\frac{p(t)}{p(0)} \right)^{-K} - 2 \right] \quad (26)$$

satisfying $g(t) > 0$ for all non-zero price variations.

II.4 Practical Implementation of SLS Controller [4]

As emphasized in [2-11], the use of prices $p(t)$ which are continuously differentiable is an idealization. In real markets, charts of prices at discrete times can appear highly non-differentiable and discontinuous. This raises questions about the efficacy of the static feedback SLS controller in real-world markets. Motivated by the fact that the SLS controller performs well in idealized markets, it becomes a candidate for implementation and back-testing in real markets. Hence, now consider that trading occurs at discrete times and note that the inter-sample time can be either small such as one minute for a high-frequency trader or large such a one day for a mutual fund.

Let $p(k)$, $V(k)$, $I(k)$ and $g(k)$ denote the discrete-time counterparts of $p(t)$, $V(t)$, $I(t)$ and $g(t)$ respectively, introducing the one-period percentage change in stock price

$$\rho(k) = \frac{p(k+1) - p(k)}{p(k)} \quad (27)$$

we consider various cases for the discrete-time model.

Simplest Case: The simplest scenario occurs when we assume no controller reset (see next section) and that long and short investments $I_L(k)$ and $I_S(k)$ maintain their proper sign; i.e., $I_L(t) \geq 0$, $I_S(t) \leq 0$ whereas $\rho(k)$ is assured by the dynamics in continuous time, in the discrete-time case, a large value of might lead to an undesirable sign reversal. If, in addition we assume no collateral requirements (say ρ is large), it follows from the continuous-time analysis that suitable dynamic update equations are:

$$\begin{aligned} I_L(k+1) &= (1 + K\rho(k))I_L(k) \\ I_S(k+1) &= (1 - K\rho(k))I_S(k) \\ I(k+1) &= I_L(k+1) + I_S(k+1) \\ g_L(k+1) &= g_L(k) + \rho(k)I_L(k) \\ g_S(k+1) &= g_S(k) + \rho(k)I_S(k) \\ g(k+1) &= g_L(k+1) + g_S(k+1) \\ V(k+1) &= V(k) + g(k) + r(V(k) - |I(k)|) \end{aligned} \quad (28)$$

with r now denoting the one-period risk-free rate of return.

More General Case: To handle the sign restriction conditions on $I_L(k)$ and $I_S(k)$, the update equations are modified to:

$$\begin{aligned} I_L(k+1) &= \max\{(1 + K\rho(k))I_L(k), 0\} \\ I_S(k+1) &= \min\{(1 - K\rho(k))I_S(k), 0\} \end{aligned} \quad (29)$$

and then build in the account collateral requirement by modifying the total investment to be

$$I(k+1) = \min\{|I_L(k+1)| + |I_S(k+1)|, \gamma V(k)\} \quad (30)$$

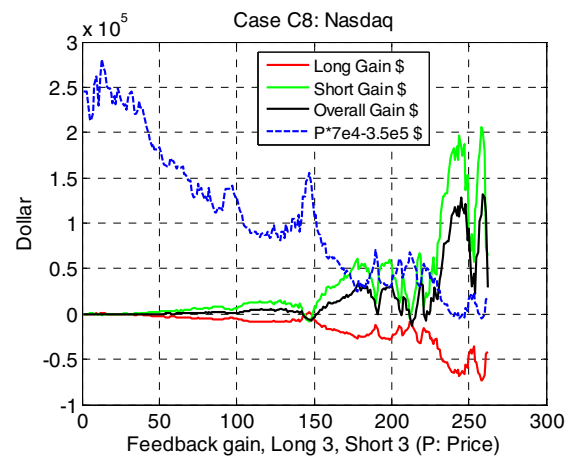
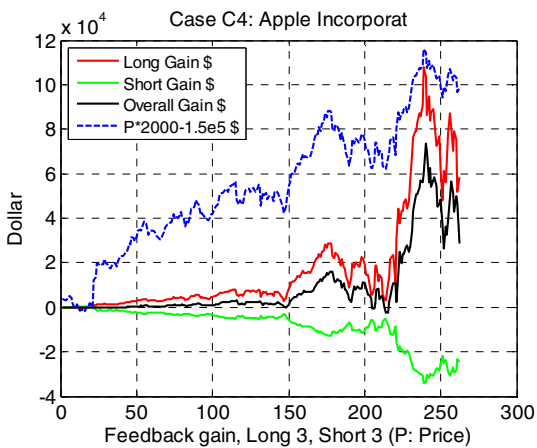
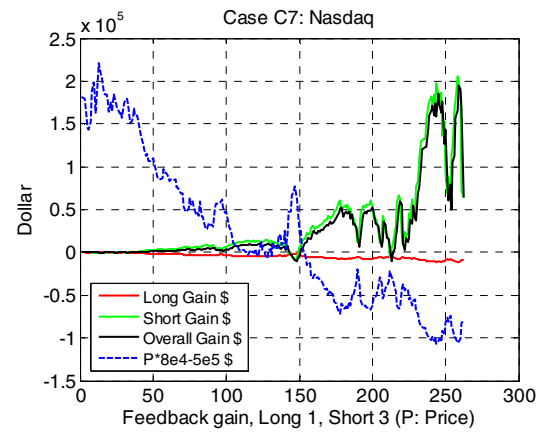
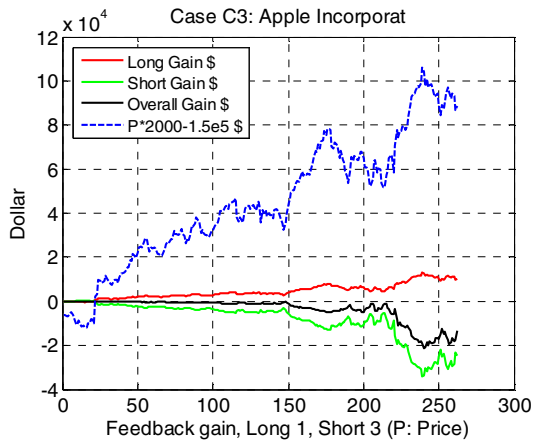
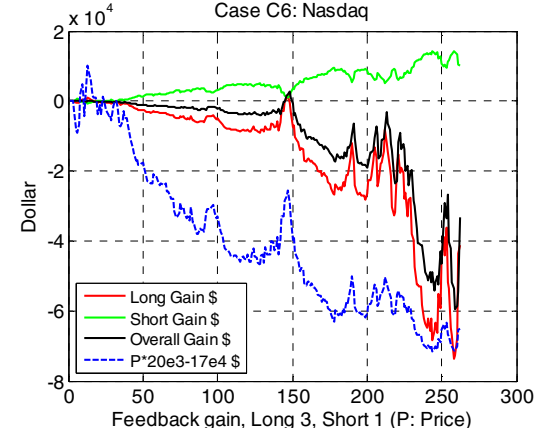
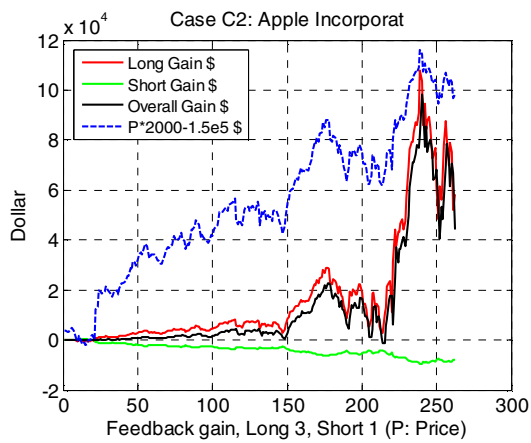
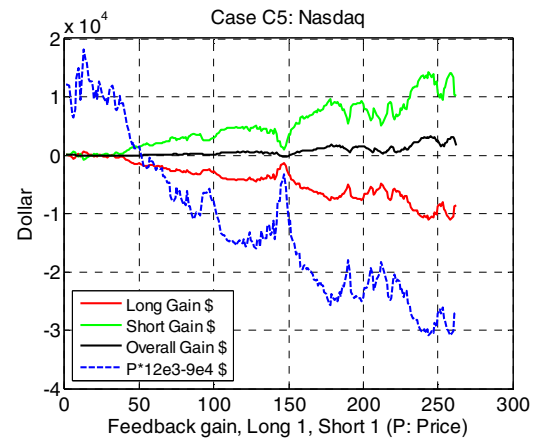
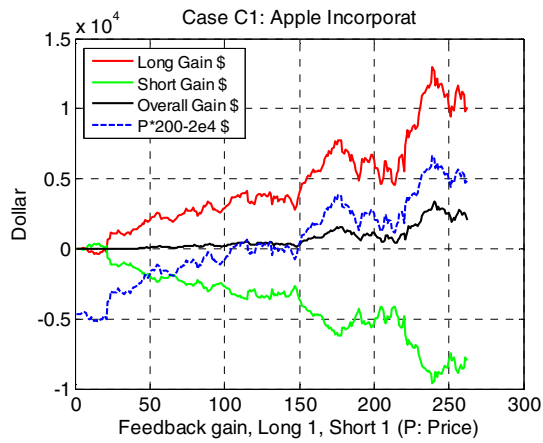
Clearly, as already suggested in [4], how to choose the feedback gain K in (28) is an issue to be further investigated.

III. SELECTED SIMULATION RESULTS AND FUTURE WORK

Based on the known literatures [2-11], there is a great scope for further work. We have made a few attempts on different topics. The simplest new idea is to assign different gains to the long and short parts of Simultaneous Long-Short feedback control. We use K_L and K_S to replace an unique K in (28): using K_L for the long trading part and K_S for the short trading part. To limit the length of this paper, we presented some selected simulation results summarized in the table below. The corresponding eight plots are presented in the next page.

Case	K_L	K_S	Stock	Gain g
C1	1	1	A	2032.60 \$
C2	3	1	A	44246.02 \$
C3	1	3	A	-13525.42 \$
C4	3	3	A	28688.00 \$
C5	1	1	B	1626.98 \$
C6	3	1	B	-33132.42 \$
C7	1	3	B	63781.38 \$
C8	3	3	B	29021.99 \$

Stock A: Apple Incorporated, B: NASDAQ-100;
Data from 2014/3/27 to 2015/3/26



In the all plots, in order to compare the relationships between long gain g_L , short gain g_S , overall gain g and price change in p , re-scaled p is plotted using blue broken lines. Observation of these plots cannot easily lead to some useful suggestions. This is only part of our study so far. In the remain space of this paper, we discuss some possible future research.

(1) Further to the topic of how to choose the feedback gain K in (28), where K is a constant [4], to study not only different K_L and K_S values, but also how these values can be changed adaptively.

(2) Apart from the feedback control K , how to choose an optimized initial value I_0 ? In general, how to modify the feedback scheme when some assumptions are not fully valid in a real trading world?

(3) Contrast to simultaneous long and short trading, in financial trading literature, one of the key topics is “triggering mechanism” for entering or exiting a long or a short trade. The study of this topic can be combined with “model-free” feedback control scheme as proposed in [2-11].

(4) In [9], a simple gain feedback controller is extended to a PI controller. From our point of view, it would be more useful to study a PD controller. In process control, the main purpose of the I term in a PI controller is to eliminate the steady-state error, but it will introduce delay. The D term in a PD controller is to “predict”, which is also a useful function in stock trading.

Broadly speaking, the work led by Prof Barmish, B. Ross and the work presented here belong to technical analysis, for example see [17]. Technical analysis is an approach to predicting future price movements based on identifying patterns in prices, volume and other market statistics. Technical analysis usually proceeds by recording market activity in graphical form and then deducing the probable future trend from the pictured history. The premise is that prices exhibit various geometric regularities, which, once identified, inform the trader what is likely to happen next. This in turn allows the trader to run a profitable trading strategy. Prof Barmish, B. Ross’s work has opened a new promising field based on, not statistical models, but the principle of feedback control. Therefore, many issues addressed in [17] can be revisited.

One of the important indicators of the market is trading volumes of stocks. Financial academics and practitioners have long recognized that past trading volume may provide valuable information about a security. However, there is little agreement on how volume information should be handled and interpreted. Even less is known about how past trading volume interacts with past returns in the prediction of future stock returns. Stock returns and trading volume are jointly determined by the same market dynamics, and are inextricably linked in theory. Combined with feedback viewpoint, there are indeed many topics to be studied.

V. CONCLUSION

This paper introduces a relatively new branch of technical analysis for stock trading. It involves the application of classical control theoretic concepts to stock and option trading. The key is to formulate the trading law as a feedback control on the price sequence. Simulation results from real stock prices based on our initial work are briefly presented. More importantly, new research topics are proposed.

REFERENCES

- [1] K. J. Astrom and P. R. Kumar, "Control: A perspective," *Automatica*, vol. 50, pp. 3-43, Jan 2014.
- [2] B. R. Barmish, "On trading of equities: a robust control paradigm," *Proceedings of the IFAC World Congress, Seoul, Korea* vol. 1, pp. 1621-1626, 2008.
- [3] S. Iwarere and B. R. Barmish, "A Confidence Interval Triggering Method for Stock Trading Via Feedback Control," *2010 American Control Conference*, pp. 6910-6916, 2010.
- [4] B. R. Barmish, "On Performance Limits of Feedback Control-Based Stock Trading Strategies," *2011 American Control Conference*, pp. 3874-3879, 2011.
- [5] B. R. Barmish and J. A. Primbs, "On Arbitrage Possibilities Via Linear Feedback in an Idealized Brownian Motion Stock Market," *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pp. 2889-2894, 2011.
- [6] B. R. Barmish and J. A. Primbs, "On Market-Neutral Stock Trading Arbitrage Via Linear Feedback," *2012 American Control Conference*, pp. 3693-3698, 2012.
- [7] S. Malekpour and B. R. Barmish, "How Useful are Mean-Variance Considerations in Stock Trading via Feedback Control?," *2012 IEEE 51st Annual Conference on Decision and Control*, pp. 2110-2115, 2012.
- [8] S. Malekpour and B. R. Barmish, "A Drawdown Formula for Stock Trading Via Linear Feedback in a Market Governed by Brownian Motion," *2013 European Control Conference*, pp. 87-92, 2013.
- [9] S. Malekpour, J. A. Primbs, and B. R. Barmish, "On stock trading using a PI controller in an idealized market: The robust positive expectation property," *Decision and Control 2013 IEEE 52nd Annual Conference on*, pp. 1210-1216, 2013.
- [10] S. Iwarere and B. R. Barmish, "On Stock Trading Over a Lattice via Linear Feedback," in *IFAC World Congress*, 2014, pp. 7799-7804.
- [11] S. Malekpour and B. R. Barmish, "The Conservative Expected Value: A New Measure with Motivation from Stock Trading via Feedback," in *IFAC World Congress*, 2014, pp. 8719-8724.
- [12] C. T. E., J. F. Weston, and S. Kuldeep, "Financial Theory and Corporate Policy (4th Ed.)," *Addison-Wesley*, 2013.
- [13] J. A. Primbs and B. R. Barmish, "ACC 2011 Tutorial Session: An Introduction to Option Trading from a Control Perspective," in *2011 American Control Conference*, ed, 2011.
- [14] J. A. Primbs and B. R. Barmish, "ACC 2012 Tutorial Session: An Introduction to Hedged-Like Stock Trading from a Control Theoretic Point of View," *2012 American Control Conference (Acc)*, pp. 4496-4497, 2012.
- [15] B. R. Barmish, J. A. Primbs, S. Malekpour, and S. Warnick, "On the basics for simulation of feedback-based stock trading strategies: An invited tutorial session," *Decision and Control, 2013 IEEE 52nd Annual Conference on*, pp. 7181-7186, 2013.
- [16] F. Black and M. Scholes, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, vol. 81, pp. 637-654, 1973.
- [17] Edwards, Robert D.; Magee, John; Bassetti, W.H.C. "Technical Analysis of Stock Trends, 9th Edition. American Management Association, 2007.

Research into Big Data for Smart Grids

E C Eze, T C Yang and C R Chatwin

Department of Engineering and Design
University of Sussex
Brighton, UK
t.c.yang@sussex.ac.uk

D Yue

Research Institute of Advanced Technology
Nanjing University of Posts and Telecommunications
Nanjing, China
yued@njupt.edu.cn

H N Yu

School of Design, Engineering and Computing
Bournemouth University
Poole, UK
yuh@bournemouth.ac.uk

Abstract— Two the largest man-made large-scale systems were created and expanded in the last century: electrical power system and world-wide computer networks, the Internet. Entering the new century, both systems are facing enormous challenges. On the one hand, Smart Grids – modernization of electrical power systems to deal with energy shortage and environment issues – are being developed in accelerating pace. On the other hand, human beings are also facing with unprecedented challenges from collecting, storing, transferring, mining, processing and visualizing massive data: Big data. The advent of a smart grid – the overlay of advanced sensing, communications, and controls on the electric network – is transforming utilities and other power sector players into IT companies. As information technologies are embedded across the entire system – from power plants through transmission lines, substations, distribution circuits, meters, and every device in industrial and residential users. Utilities’ operational and information models will increasingly resemble those of telecom, Internet, or even financial trading companies. This will require a fundamentally new approach to interoperability, speed, and managing and making sense of vast new floods of data. Therefore, this paper is to follow this route to address Big data in future power systems. Specifically, in this paper, we report our initial work on predicting wind farm outputs based on historical wind speed data.

Keywords- *Big Data, Smart grid, Wind farm output prediction.*

I. INTRODUCTION

Evolving technologies in the energy and utilities industry, including smart meters, new IT technologies and smart grids, can provide companies with unprecedented capabilities for, to name a few, forecasting demand, shaping customer usage patterns, preventing outages, optimizing unit commitment and more. At the same time, these advances also generate

unprecedented data volume, speed and complexity, so-called Big data. Tracing back to the emerging of the new era of Big data, Google’s pioneering paper on Map-Reduce (MR) published in 2008 [1] was the trigger that led to lot of developments afterwards. While not a fundamental paper in terms of technology – the map-reduce paradigm was known in the parallel programming literature – it along with Apache Hadoop (the open source implementation of the MR paradigm) [2] enabled end users (not just scientists) to process large data sets on a cluster of nodes – a usability paradigm shift. Hadoop which comprises the MR implementation along with the Hadoop Distributed File System (HDFS) has now become the de-facto standard for data processing, with lot of industrial variations having their own Hadoop cluster installations. A “Forrester Wave Hadoop Report” [3] evaluated 13 enterprise Hadoop solution providers against 15 criteria. Providers were analyzed on their current offering, strategy and market presence. Clearly, in this very dynamic market to offer platforms for Big data processing, things are changing fast. One always needs to do his/her own survey to have a latest picture.

The rest of the paper is organized as follows. A typical Big data platform and three broad research themes are introduced in Section II. In order to have a good understanding of application requirements, Section III is fully devoted to operation of future power systems. Section IV is a summary of our initial work on predicting wind farm outputs based on historical wind speed data. A brief conclusion is made in Section V.

II. BIG DATA

Big Data is the growing challenge that organizations face. They deal with large and fast-growing sources of data or information that also present a complex range of analysis and use problems. Digital data production in many fields of human activity from science to enterprise is characterized by an exponential growth – for the past two years, 90% of the data in the world have been created. Big data technologies will become a new generation of technologies and architectures which is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time.

The four inter-related basic blocks of the Hadoop Big data platform are (1) systems and infrastructure management, (2) data management, (3) visualization and (4) the applications on top of the above three (Figure 1).

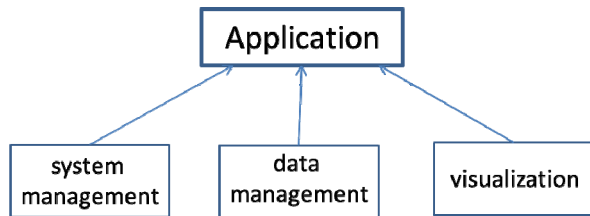


Figure 1: A typical Big data platform

The three broad research themes that are attracting a lot of research include:

(a). Storage, search and retrieval of Big-data

Trends in this direction includes use of coding techniques such as Erasure Coding to optimize storage space, performing coding on graphics processing units, context sensitive search as well as Iceberg queries on Big-data.

(b) Analytics on Big-data

Current trends include real-time analytics by using Twitter's Storm integrated with Esper or other such query tools, video analytics and ad-placement analytics as well as incremental analysis using Google's Dremel. Interesting startups include Parstream, Skytree, Palantir and Platfora.

(c) Computations on Big-data

Trends in this direction includes the moving beyond Map-Reduce paradigm using Google's Pregel and investigations of paradigms beyond the Map-Reduce for Big data. Interesting startups include Paradigm4, Black Sky, HPCC (which is exploring Hadoop alternatives) and YarcData (which has an alternative to the MR paradigm based on graph processing).

III. OPERATION OF FUTURE POWER SYSTEMS

A. Challenges facing

Energy is the life blood of modern civilizations and has enormously boosted the development of the world economy and human society. Today, the majority of energy consumed is from three main fossil fuels, coal, petroleum and natural gas. These three together supply more than 85% of the world's energy consumption. Nevertheless, due to the nonrenewable and non-environmentally-friendly features of fossil fuels, more and more society and environment problems are emerging: rapid increases of fuel prices, greenhouse gas emissions from fuel combustion, acid rain, and so on. To address these issues, non-polluting renewable energy resources, such as solar, wind and hydrogen, are extensively proposed as alternative resources to deal with the emerging energy crisis associated with fossil fuels. Yet aiming for higher shares of distributed renewable energy in end-users' energy consumptions, the current power system, which is suffering from transmission and distribution losses and vulnerable to power outages, is unlikely to meet challenges on system efficiency and stability when delivering renewable energies. To this end, upgrading the aging power system towards the smart grid is imperative by integrating efficient communication infrastructures with power systems for timely system monitoring and control [6]. Within the upgraded system, power equipments are interconnected to practice a brand new power management paradigm, that is, utilizing bidirectional information flows to drive bidirectional electricity flows. In this way, we can significantly mitigate impacts of variability and uncertainty of renewable resources and dramatically improve energy efficiency. As outlined in the abstract, such bidirectional information flows will generate Big data and this motivates the research into Big Data for Smart Grids. However, information collecting, storing, transferring, mining, processing and visualizing is not for information itself, but for the reliable and optimal operation of future power systems. To this end, the next subsections will outline some "hardware operation details" of a typical smart grid. We use the FREEDM system: Future Renewable Electric Energy Delivery and Management system [4] as an example which is envisioned to demonstrate the generation, distribution, storage, and management of renewable energy resources in a smart grid. The two Figures used, 2 and 3, are copied from the FREEDM publications [4-10].

B. Overview

Figure 2 gives an overview of a FREEDM system [4]. As shown in the figure, residential users are able to supply energy demands with distributed renewable energy generators installed in their houses, such as solar panels and wind turbines. These energy generators, together with residential energy storage facilities, like batteries and electrical vehicles, make

conventional energy customers to be energy providers by selling excess energy to the public. Accordingly, an Energy Internet is formed across interconnected users to exchange information and share energy in bidirectional manners.

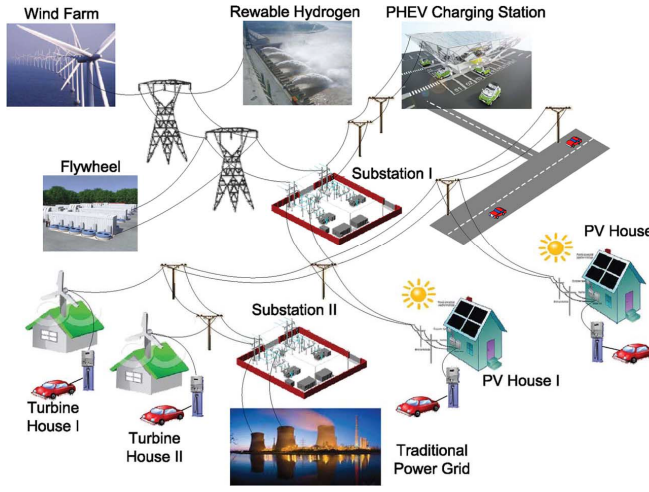


Figure 2: An overview of a FREEDM system

C. Power flow in a “FREEDM community”

Since a FREEDM zone is composed of several residential houses that are geographically close, it also indicates a “FREEDM community,” where residents operate renewable energy facilities coordinately for energy sharing. To manipulate a FREEDM zone, two equipments are crucial, including the Intelligent Energy Management (IEM) and the Intelligent Fault Management (IFM). The IEM aims to exploit real-time equipment monitoring for local energy managements, such as on-demand routing energy, interfacing various loads, and so on [4]. These energy management functions are achieved by an advanced power electronic equipment, the Solid State Transformer (SST), which is integrated to undertake responsibilities of electricity transforming with different outputs: direct current (dc) or alternating current (ac), at different voltage levels, and at different power quality levels. Apart from the underlying SST, an IEM also involves Distributed Grid Intelligence (DGI) as the software platform, which is more related to control operations in response to various system situations, such as electricity dispatching and feeder reconfigurations. The IFM is designed to identify and isolate unexpected faults for system stability maintenances, which is established on a Fault Isolation Device (FID) and also controlled by the DGI platform.

With IEMs and IFMs, residential renewable energy generators are prone to be organized in a “FREEDM Community”. Zone A is a 1MVAFREEDM system instance that entails that the total power of all loads in Zone A is 1MVA. At the entrance of the FREEDM zone, a 1 MVA IEM is installed as the interface

between 69 kV ac transmission lines and 12 kV ac distribution lines. In the FREEDM zone, the 12 kV distribution bus hooks multiple IFMs for line protections and five 20 kVA IEMs in a loop manner for energy sharing. Each 20 kVA IEM is mapped to a residential house to manage all renewable energy facilities at home, including loads, Distributed Energy Storage Devices (DESD) (i.e., batteries and electrical vehicles), and Distributed Renewable Energy Generators (DREG) (i.e., solar panels and wind turbines). To accommodate power demands of end users, the 20 kV IEM is equipped with two outputs, including 120 V ac and 380 V dc. 120 V ac is the most common voltage level of power supply for home appliances in North America. Yet 380 V dc is an emerging dc voltage standard dedicated to provide a dc output for data centers, uninterruptible power supply (UPS) and lightening applications towards a high energy efficiency. The “FREEDM community” may run in three different states, including self-sufficiency, charging, and discharging. As DREG and DESD equipments are employed to energize loads in the zone, an equilibrium point can be achieved when powers supplied by DREGs and DESDs are right equal to load consumptions. In that case, Zone A is in a self-sufficiency state and does not need any power from the grid. Accordingly, the charging state means that powers generated or stored in Zone A is not enough to supply its own loads, then energy of the grid will be introduced to supplement the energy shortage. As for the discharging state, it implies that powers inside Zone A are more than enough to supply other zones.

D. Communication infrastructure

Besides critical power electronic equipments, such as IEMs and IFMs, another important feature of FREEDM systems is an efficient communication infrastructure [4], which is responsible for delivering system related messages to confer accurate and timely system awareness to all FREEDM equipments towards efficient and intelligent system managements. For a thorough communication scenario analysis, a FREEDM zone is divided into four levels, including inter-zone level, FREEDM system level, SST level, and user level.

(1) *Inter-Zone Level.* Communications in the inter-zone level aim to establish connections between multiple power distribution systems for synergetic energy sharing. For example, when Zone A runs in a charging state, i.e. powers of DREGs and DESDs cannot satisfy load demands, the zone agent (1MVAIEM in ZoneA) needs to negotiate with neighbor zones to determine which zones it should buy energy from, regarding available powers and real-time prices.

(2) *FREEDM System Level.* Communications in this level are related to interactions of peer equipments, including the 1 MVA IEMs, five 20 kVA IEMs and

multiple IFMs. Towards a well-maintained zone, peer equipments exchange information frequently regarding real-time measures of powers, currents and voltages. Such information exchanges tend to be more frequent in a fault scenario, in which all peer equipments report states in a high rate for a fast and accurate fault positioning.

(3) *SST Level*. Communications in this level aim to make SST to “talk” with other equipments, including sending out running states, and receiving outside commands for real-time equipment monitoring and control. Thereby, it is more related to on-device communications towards an intelligent equipment.

(4) *User Level*. The user level involves more equipments, such as loads, DESDs, and DREGs, all of which need to share real-time information for optimized system states. For example, residents can leverage real-time electricity price information to determine how to use excess generated powers, charging DESDs or selling to neighbors.

E. Three main applications

(1) Monitoring grids: Utilities need to proactively identify abnormal conditions and take action to both prevent power delivery disruptions and optimize overall grid reliability. Distribution companies can improve both customer satisfaction and regulatory compliance by reducing the number and duration of power outages.

(2) Unit commitment Companies must optimize the scheduling of their generation assets, taking into account a broad range of constraints to generate an optimal solution. These considerations include cost, emissions, ability to use existing delivery infrastructure and other factors—for example, wind and solar energy sources may be heavily weather-dependent and intermittent, requiring analysis of large weather data sets to forecast output.

(3) Forecasting and scheduling loads: Accurate demand forecasting is essential to energy planning and trading. Companies must be able to predict when they can profitably sell excess power and when they need to hedge supply. They must determine when it is economically advantageous to buy, sell or trade power on the open market. By anticipating purchases well in advance, organizations are better able to obtain favorable prices.

IV. PREDICTING WIND FARM OUTPUTS

As a green and renewable energy resource, wind power generation has been growing rapidly around the world. Accurate short-term wind power forecasts with a prediction horizon from one hour to several days are critical to optimize the scheduling of wind farm maintenance, security and reserve for the grid. These have an impact on grid reliability and costs for market-based ancillary service. This optimization requires

manipulating large volume of historical data generated from various instruments – anemometers, pyrometers, rain gauge etc; in a wind farm as depicted in Figure 3. In this section, we report our initial work on predicting wind farm outputs based on historical wind speed data.

A. Data collection, processing and cleaning

Prediction is based on data recorded at various instruments and Probability Mass Functions (PMFs) are used to model the data. Here, probabilities \square are sorted and grouped to form \square . After which the process is repeated to produce \square and thus, further combined to have a generalized or uniform prediction. The size of \square is the size of the vector generated from Ω . $P_X(\square) = P(X = \square)$ and $P(w \square \Omega \text{ s.t } X(w) = \square)$. Ω produces \square that forms the output of the prediction. Here $P_X(\square)$ are non-negative $P_X(\square) \geq 0$ and $\sum_{x=0}^n P_X(\square) = 1$. To predict power output from a wind farm, data at 10 minutes interval from 2002 to 2006 generated from a wind farm near Boston Massachusetts was used. These data are from pyrometers (that measured internal and external temperature), anemometers of variable hub heights – 25, 30 and 45 meters that recorded wind speed and wind directions. Longitude, latitude and altitude were not considered. The data, after merging and joining, is about 1.5GB. For ease of presentation and visualization, in this paper only one year data was used and the results of five months were presented. A typical Big Data platform, Microsoft Azure (<http://azure.microsoft.com/>) was used for data storage and analysis.

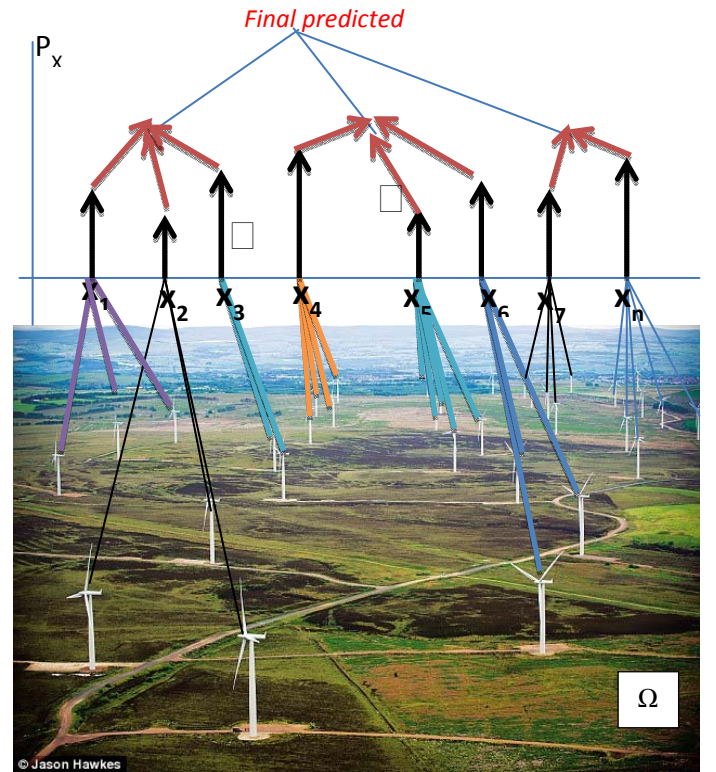


Figure 3: Mass-density-mode wind speed prediction.

Data processing and cleaning was carried out to remove missing values. The normalized data is plotted in Figure 4. This figure shows the diurnal and seasonal pattern. It also shows wind speed increases in height and that the recorded wind speed and direction has similar pattern irrespective of hub heights, though with significant lower wind speed in summer over winter. Figure 4 is generated from joining and merging selected dataset using Azure's capabilities. The observed correlation patterns are shown in Figure 5 a, b and c respectively on a randomly 5 months data sample at 25m (Figure 5 a), 30m (Figure 5 b) and 45m (Figure 5 c). Based on the merged data of figure 5, distribution of wind speed was analyzed as shown in figure 6.

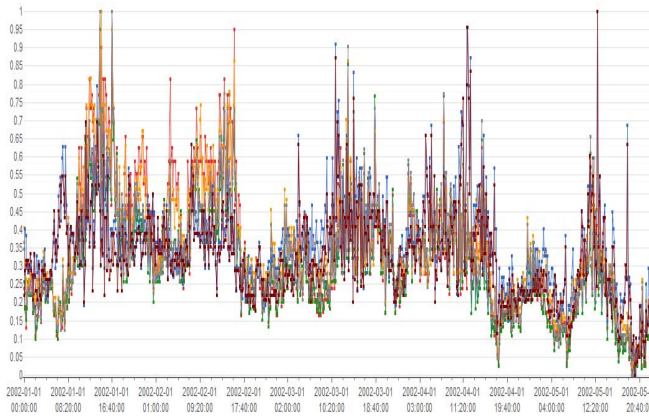


Figure 4: normalised wind speed over randomly selected anemometers from a wind farm

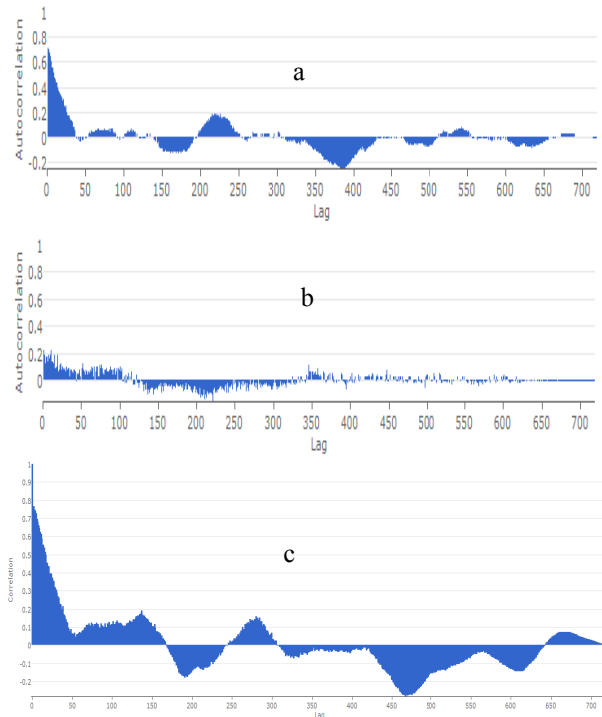


Figure 5: Correlation of observed joined dataset at 25, 30 and 45 meters (a,b,c).

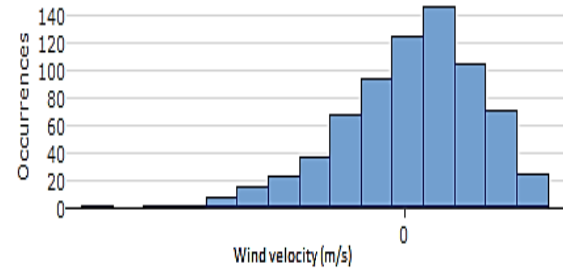


Figure 6: wind speed distribution based on figure 5

B. Model training and testing

The artificial neural network (ANN) architecture used for prediction is shown in figure 7.

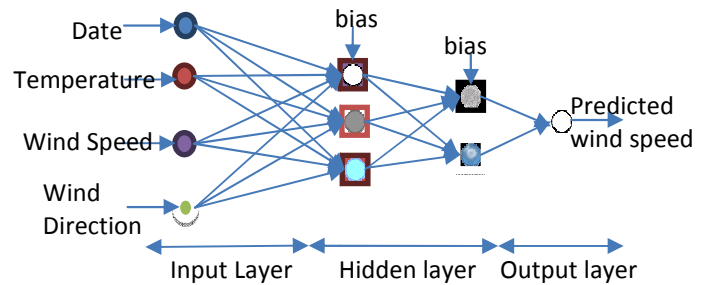


Figure 7: wind speed neural network architecture for variable prediction

Using a Azure ML based cloud system, a customized multilayer feed forward back-propagation artificial neural network was designed using C#, integrated to Azure. *Tan sigmoid* transfer function was used at the hidden layer.

Training was carried out after data pre-processing to the customized neural network for time series regression. Hence, big data and cloud data processing are applied – parallel computing at different microprocessor cores for speedy and efficient output. Three years of data at 10 minutes interval were fed in to the network as a block data and were further cascaded into four segments of 6 months each (inside the segmented first six months of data were merged and joined). In addition, avoiding over-fitting and generalization were also considered. Furthermore, early stopping – VC dimension was implemented in the training algorithm; and sweep parameters were implemented. Time stamp before 15/04/2002 – equivalent to 70% of the data were used for training. 16/04/2002 to 23/05/2002 – equivalent to 20% of the data were used for validation while the rest were used for testing. Since we are interested in minimising root mean square error (RMSE), sweep parameters are not optimised. Maximum number of iteration was set to 100 and 10^{-18} was chosen for accuracy tolerance. Evaluation of results was carried out based on coefficient of correlation (R^2) between actual and predicted. That is the correlation and RMSE of

$\sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{P_{j,m} - P_{j,f}}{P_N} \right)^2}$ (where $P_{j,m}$ and $P_{j,f}$ denotes measure and forecasted power output respectively, N represents number of predictions and P_N stands for rated power output) to ascertain best prediction capacity of the network. After training and testing, a forecast of one week was carried out on the merged (inner join) dataset used in the project as shown in Figure 8. Learning is shown in blue colour and the red colour shows predicted results. The figure also shows that when forecast horizon increases, the forecast accuracy decreases.

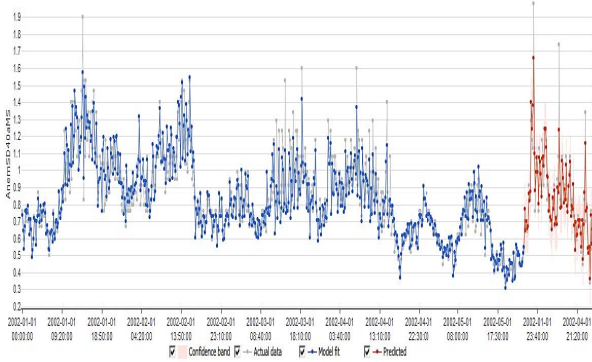


Figure 8: NN prediction based on merged dataset from the wind farm.

To further understand the distribution regularity in wind farm, a wind farm power output was estimated from the one week ahead prediction using $P = \frac{k}{1 + ae^{-bv}}$ as shown in Figure 9; where v represents the predicted wind speed, P stands for power output; a , b and k are S-shaped statistical curve parameters.

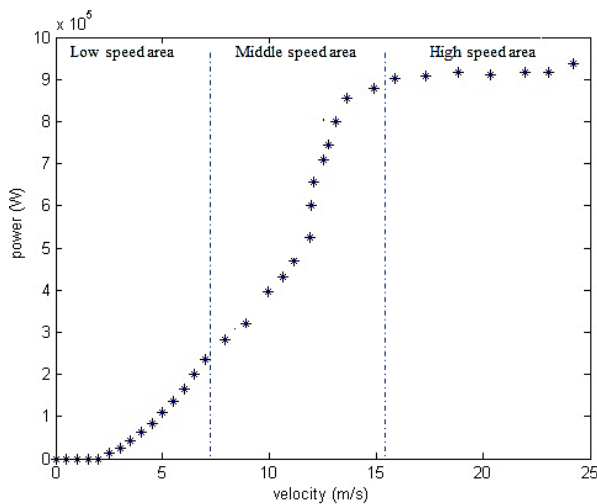


Figure 9: Estimated wind farm power output based on sample predicted wind speed.

V. CONCLUSION

This paper presents an introduction to research into Big Data for smart grids. Our initial work on wind farm output prediction is also reported. Further exploration is yet to be conducted. Nowadays, changes occur in an extremely rapid pace such that one technology has not yet reached its end of design life before a newer technology is introduced. Noticeably, “Second-Generation Big Data Systems” is proposed recently [12]. Furthermore, on the top of Hadoop and other popular Big Data systems, “BigDataBench” was launched in December 2014 [13] at Cmabridge, UK. All these will affect the progress towards Big Data applications for smart grids.

REFERENCES

- [1] J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” Communications of the ACM, vol. 51, no. 1 (2008), pp. 107-113.
- [2] http://en.wikipedia.org/wiki/Apache_Hadoop
- [3] <http://pentahobigdata.com/resources/analyst-resources>
- [4] A. Huang, M. Crow, G. Heydt, J. Zheng, and S. Dale, “The future renewable electric energy delivery and management (freedm) system: The energy internet,” Proc. IEEE, vol. 99, no. 1, pp. 133–148, 2011.
- [5] X. She, S. Lukic, A. Huang, S. Bhattacharya, and M. Baran, “Performance evaluation of solid state transformer based microgrid in freedm systems,” in Proc. 2011 26th Annu. IEEE Appl. Power Electron. Conf. Expo. (APEC).
- [6] G. Karady and X. Liu, “Fault management and protection of FREEDM systems,” in Proc. 2010 IEEE Power Energy Soc. Gen. Meeting.
- [7] M. Baran and M. Steurer, “A digital testbed for FREEDM system development,” in Proc. 2012 IEEE Power Energy Soc. Gen. Meeting.
- [8] X. Lu, W. Wang, A. Juneja, and A. Dean, “Talk to transformers: An empirical study of device communications for the freedm system,” in Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm), 2011.
- [9] A. Huang, “FREEDM system—A vision for the future grid,” in Proc. IEEE Power Energy Soc. Gen. Meet. (PES’10).
- [10] Y. Jiang, P. A. Tatcho, and H. Li, “The performance analysis of FREEDM system fault with SST,” in Proc. FREEDM Syst. Center Annu. Conf., 2010.
- [11] Vladislavleva, E., et al., “Predicting the energy output of wind farms based on weather data: Important variables and their correlation”. Renewable Energy, 2013. 50(0): pp. 236-243.
- [12] Fadi H. Gebara., et al., “Second-Generation Big Data Systems”, Computer, Vol. 48, No. 1, 2015, pp.36-41.
- [13] <http://prof.ict.ac.cn/BigDataBench/>

Cash Flow Prediction Using a Grey-Box Model

Yang Pang^a; Kwaku Opong^a; Luiz Moutinho^a; Yun Li^b

a. Adam Smith Business School; b. School of Engineering

University of Glasgow

Glasgow, United Kingdom

Corresponding Email: y.pang.1@research.gla.ac.uk

Abstract— This paper tackles the problem of financial forecasting by extending methods developed in automation, engineering and computing science. Current methods existing in the literature for firm-level cash flows are first analysed. Then a grey-box modelling method is developed to elevate the performance of cash-flow prediction. Linear panel data modelling is used as a benchmark model. Experiments with out-of-sample tests are used to validate the grey-box approach. Encouragingly, nonlinear grey-box modelling outperforms linear panel data modelling in both one-period-ahead and multi-period-ahead predictions.

Keywords—Prediction; cash flow; panel data; grey-box

I. INTRODUCTION

Making prediction is an inevitable and crucial part in economic and financial analysis, as well as in business practice. Predictions usually lead to decision making. For instance, sales prediction will affect a firm's inventory management; forecasts made by financial analysts are used to construct portfolios, monetary and fiscal policies of a country are also made with respect to the forecast of the country's future economic state etc. Various methods and techniques have been developed to enhance predictive power of models used to forecast economic variables. In both economics and finance, academic research is important in explaining the associations and interactions between particular variables. Seeking high accuracy in prediction is of less concern. As a result, studies in these areas tend to adopt simple and parsimonious models, usually in the linear form. The Nobel winning capital asset pricing model (CAPM) is one example (see e.g. [1]), which is used to model expected return of equity market. Similarly, a stock return predictive model that is widely cited is also a linear model [2].

Applications in science and engineering, however, place greater effort in reaching high accuracy in predictions. Thus methods in engineering are generally more complicated. The border between these different disciplines is becoming blurry in some aspect. Many models that are widely applied in economics and finance were originally developed and used in the area of engineering, mathematics or other disciplines of science. For instance, Kalman filter [3] were developed to deal with problems in signal processing and it is now commonly applied in econometrics for estimating state-space models; neural network in the area of artificial intelligence has also been widely been applied to work in financial practice [4]. Besides, it is a common phenomenon that finance borrows ideas from or even merges with other fields, forming interdisciplinary studies.

This paper investigates various modelling methods in cash flow prediction. This is an application with analysis of firm-level financial or accounting information. There have been previously developed prediction models to predict cash flows in accountancy and finance. This paper applies the grey-box model developed in [5] and investigates the potential dynamic and nonlinear features of the model to cash flows that have been overlooked in previous modelling procedures.

The structure of this paper will be as follows. Section 2 introduces the background relating to cash flow prediction. Section 3 discusses the models that will be applied in this piece of work and section 4 discusses the research designs and data sources. Section 5 will present the empirical results by analysing data with the proposed models. In the final, section 6 will be the conclusive part.

II. BACKGROUND

Cash, not merely as a number, is the true income of business transactions. Firms will pay their debts and expenses with cash so they will fail without cash incomes no matter how good their earnings number looks. Earnings are not and should not necessarily be treated as equivalent to cash flow. Before the introduction of statement of cash flow, income statement reflected a firm's revenue, expense and profit, but the numbers meant nothing if they could not be eventually translated into the same amount of cash. Due to trade credit, a transaction could occur when revenues and expenses are recognised in accounting even without immediate cash settlement. There is the risk that these pre-recognised amounts may not fully end up with the equivalent cash because of default for instance. As a result, it is uncertain whether accounting revenues and the cash income will be equal. Similarly, the accounting expenses are not necessarily equal to the amount of cash paid out in the current period. In addition, there are non-cash items being recorded as expenses such as depreciation, which affect earnings but not cash flow. Earnings that are the difference between revenues and expenses are therefore often unmatched with cash income. If cash is genuinely the measure of profit, accounting earnings do not accurately reflect it.

In the U.S, a statement of cash flow has been a compulsory part of financial reports since 1987 when statement of financial accounting standards (SFAS) No. 95 was published. Before the U.S, Canada was the first country which required cash flow disclosure in 1985. Thereafter, the time series property of cash flow started to be studied along with earnings. Before then cash flow was indirectly estimated by deducting accrual terms and

Pang Yang is grateful to University of Glasgow for a Kelvin-Smith scholarship.

non-cash items from earnings. There were early studies on the time series property of cash flow and the tool used in this procedure is mainly under the ARIMA framework (see e.g. [6]). The ARIMA model is univariate model, therefore it ignores potential influence of exogenous variables. [7] states importantly that unlike earnings, univariate time series model may not be sufficient to model cash flow because there are predictable components in cash flows brought on by accrual terms (denote their model as DKW for short). It would be better for cash flow prediction to include more variables in addition to lagged cash flow. [8] therefore adjusted DKW model by disaggregating earnings into cash flow and accrual term components as multiple predictors other than using aggregated number of earnings (denote their model as BCN model). Based on BCN model, there were further extensive studies (see e.g. [9]).

III. METHODOLOGIES AND MODELS

As the disclosure of firms' financial information mainly comes from their annual report, the frequency of data in practice is relatively low (even for firms disclosing their information quarterly, the frequency is still not high enough). Therefore, the length of sample data is usually short. This limitation of dataset has restricted the application of many advanced and complicated time-series methods that rely heavily on sufficiency of data. In economics, people often deal with panel data that is to put time series observations of many individuals together, where certain assumptions on the distribution of parameters apply. In this way, particular model parameters could still be estimated on a pool of individuals with short data samples. The DKW model is in the form of:

$$CF_{i,t+k} = \gamma_{i,0} + \gamma_{i,1}CF_{i,t} + \gamma_{i,2}EARN_{i,t} + \varepsilon_{i,t} \quad (1)$$

where CF denotes net operating cash flow and $EARN$ denotes earnings. BCN model suggests that these accrual terms could have different effect in the model and therefore it extends (1) as:

$$CF_{i,t+1} = \beta_0 + \beta_1CF_{i,t} + \beta_2\Delta INV_{i,t} + \beta_3\Delta AP_{i,t} + \beta_4\Delta AR_{i,t} + \beta_5DEP_{i,t} + \beta_6AMORT_{i,t} + \beta_7OTHER_{i,t} + \varepsilon_{i,t+1} \quad (2)$$

where ΔINV denotes changes in inventory, ΔAP denotes changes in account payable, ΔAR denotes changes in account receivable, DEP denotes depreciation, $AMORT$ denotes amortisation and $OTHER$ denotes other accruals. Clearly (2) has more details than (1) and thus requires more data points to estimate the many parameters. In the two papers, DKW and BCN models were indeed estimated in different ways. As DKW model has 3 parameters to be estimated for each firm, it is possible to undertake estimation procedure individually for sample firms. BCN model requires estimation of 8 parameters; therefore the model was estimated using pooled regression in the original paper, which restricted the parameters to be identical across different firms.

Despite the different ways in the estimation of parameters, the parameters in (1) and (2) are static. However, from DKW model's assumption, the cash flow process could be re-derived,

which would take a dynamic form instead of (1) or (2). The derivation is shown as in the following.

Net operating cash flow is the difference of cash received and cash paid out, which is represented as in DKW's paper:

$$\begin{aligned} CF_t &= (SALES_t - \Delta AR_t) - (PURCHASE_t - \Delta AP_t) \\ &= (SALES_t - \Delta AR_t) - (COST_t + \Delta INV_t - \Delta AP_t) \\ &= EARN_t - \Delta WC_t \end{aligned} \quad (3)$$

where the definitions of each term are as follow:

CF : Net operating cash flow

$SALES$: Sales

ΔAR : Changes in account receivable

$PURCHASE$: Purchase that is equal to the sum of cost and changes in inventory

ΔAP : Changes in account payable

$COST$: Cost

ΔINV : Changes in inventory

$EARN$: Earnings that is the difference of sales and cost

ΔWC : Changes in working capital that is equal to $\Delta AR_t + \Delta INV_t - \Delta AP_t$

Assume cost, account receivable, account payable and inventory to be constantly proportional to sales, therefore earnings and working capitals are also constant proportion of sales. Assign two constant α and β so that:

$$EARN_t = \alpha SALES_t \quad (4)$$

$$WC_t = \beta SALES_t \quad (5)$$

Define r to be the growth rate of sales and the following relation holds:

$$SALES_t = (1+r_t)SALES_{t-1} \quad (6)$$

With some manipulations, there will be recursive relationships for earnings and working capital as:

$$\begin{aligned} EARN_t &= \alpha SALES_t \\ &= \alpha(1+r_t)SALES_{t-1} \\ &= (1+r_t)EARN_{t-1} \end{aligned} \quad (7)$$

$$\begin{aligned} \Delta WC_t &= \beta \Delta SALES_t \\ &= \beta r_t SALES_{t-1} \\ &= \beta r_t (1+r_{t-1})SALES_{t-2} \\ &= \left(\frac{r_t}{r_{t-1}} + r_t\right) \Delta WC_{t-1} \end{aligned} \quad (8)$$

Rewrite (3) as:

$$\begin{aligned}
CF_t &= EARN_t - \Delta WC_t \\
&= (1+r_t)EARN_{t-1} - \left(\frac{r_t}{r_{t-1}} + r_t\right)\Delta WC_{t-1} \\
&= (1+r_t)EARN_{t-1} - (1+r_t)\Delta WC_{t-1} + \left(\frac{r_{t-1}-r_t}{r_{t-1}}\right)\Delta WC_{t-1} \\
&= (1+r_t)CF_{t-1} + \left(\frac{r_{t-1}-r_t}{r_{t-1}}\right)\Delta WC_{t-1}
\end{aligned} \tag{9}$$

Combining (8) and (9), cash flow and changes in working capital will be described as:

$$E_{t-1} \left(\begin{bmatrix} CF_t \\ \Delta WC_t \end{bmatrix} \right) = E_{t-1} \left(\begin{bmatrix} 1+r_t & \frac{r_{t-1}-r_t}{r_{t-1}} \\ 0 & \frac{r_t}{r_{t-1}}(1+r_{t-1}) \end{bmatrix} \right) \times \begin{bmatrix} CF_{t-1} \\ \Delta WC_{t-1} \end{bmatrix} \tag{10}$$

Therefore, there are potentially nonlinearity and dynamics in the unobservable parameters. To take account of such effects, this paper will propose the application of Grey-box model in enhancing the simple linear model form.

Grey-box was developed in [5], where ‘grey’ means the combination of black box and clear box. Take (2) for instance, the model is in a clear linear form, and differences in estimation method will only alter the value of parameters but do not change the model to be black box. Model (10) is also a clear box model because it is a theoretical model based on ideal assumptions. Empirically, the parameters of cash flow model (1) and (2) will not exactly comply with the form described in (10). Their evolution, instead, is not clearly explainable due to complex environment and inevitable omissions of model assumptions. Therefore, processes in social sciences cannot be precisely captured by perfect mathematical models because there are interactions between infinitely many variables and extremely high uncertainty. For such processes that cannot be captured by clear box, black box will be a powerful alternative option. Assume parameters in (2) are nonlinear and time varying:

$$\beta_{i,t} = F(\mathbf{z}_t) \tag{11}$$

where $F(\mathbf{z}_t)$ is a nonlinear function of some variables \mathbf{z} . The form of function F is unknown; therefore it would be numerically approximated with a black box model. There are several options for such functions. For instance, a neural network (NN) is considered as a universal approximation that is able to approximate any function [10]. Similarly, Taylor series and Fourier series are two more examples that can approximate functions with any degree of accuracy. This paper adopts Padé approximant [5] for the nonlinear function as the method is efficiently accurate with only a few coefficients to determine.

In this paper, two forms of grey-box model (GM) will be simultaneously examined.

GM1:

$$\begin{aligned}
CF_{i,t+1} &= \beta_{i,t,0} + \beta_{i,t,1}CF_{i,t} + \beta_{i,t,2}\Delta INV_{i,t} + \beta_{i,t,3}\Delta AP_{i,t} + \beta_{i,t,4}\Delta AR_{i,t} \\
&\quad + \beta_{i,t,5}DA_{i,t} + \beta_{i,t,6}OTHER_{i,t} + \varepsilon_{i,t+1} \\
\beta_{i,t} &\approx \frac{c_0 + c_1r_t + c_2r_t^2}{1 + d_1r_t + d_2r_t^2}
\end{aligned} \tag{12}$$

GM2:

$$\begin{aligned}
CF_{i,t+1} &= \beta_{i,t,0} + \beta_{i,t,1}CF_{i,t} + \beta_{i,t,2}\Delta INV_{i,t} + \beta_{i,t,3}\Delta AP_{i,t} + \beta_{i,t,4}\Delta AR_{i,t} \\
&\quad + \beta_{i,t,5}DA_{i,t} + \beta_{i,t,6}OTHER_{i,t} + \varepsilon_{i,t+1} \\
\beta_{i,t} &\approx \frac{c_0 + c_1AGE_{i,t} + c_2AGE_{i,t}^2}{1 + d_1AGE_{i,t} + d_2AGE_{i,t}^2}
\end{aligned} \tag{13}$$

where DA is depreciation and amortisation (which are aggregated for convenience in data processing); AGE denotes firm age, which is calculated as the difference of time t and the initial time t_0 when the first observation appears in the sample; c s and d s are the coefficients of Padé approximants to be determined. GM2 uses age as a proxy variable for firms’ growth path based on empirically observed pattern that firms’ growth is related to firms’ ages.

In comparison, model (2) with static parameters will be estimated as benchmark (BM). This paper use two estimation methods of model (2) to allow differences in individual intercept terms:

BM:

$$\begin{aligned}
CF_{i,t+1} &= \beta_{i,0} + \beta_{i,1}CF_{i,t} + \beta_{i,2}\Delta INV_{i,t} + \beta_{i,3}\Delta AP_{i,t} + \beta_{i,4}\Delta AR_{i,t} \\
&\quad + \beta_{i,5}DA_{i,t} + \beta_{i,6}OTHER_{i,t} + \varepsilon_{i,t+1}
\end{aligned} \tag{14}$$

The model is named fixed effect (or random effect if different assumption is made) model in micro-econometrics. Allowing individual effect as represented by the intercept terms would increase the models practical predictive power. The parameters of β_1 to β_6 are still assumed to be homogeneous among firms. They will be estimated by two methods: demean and first difference (for more details please see the textbooks [11]), and the intercept term is calculated as:

$$\begin{aligned}
\hat{\beta}_{i,0} &= \overline{CF}_{i,t} \\
&\quad - (\beta_1\overline{CF}_{i,t-1} + \beta_2\overline{\Delta INV}_{i,t-1} + \beta_3\overline{\Delta AP}_{i,t-1} \\
&\quad + \beta_4\overline{\Delta AR}_{i,t-1} + \beta_5\overline{DA}_{i,t-1} + \beta_6\overline{OTHER}_{i,t-1})
\end{aligned} \tag{15}$$

IV. RESEARCH DESIGN AND DATA

The models (GM1, GM2, BM demean and first difference) will be examined in the performance of practical prediction for cash flow. The comparison procedure will be in three stages: in-sample fitness, out-of-sample one period prediction and out-of-sample multiple period prediction. The out-of-sample test is adopted to take account of the fact that complicated models

tend to over-fit data. Over-fitting models would have very good performance in-sample but poorer performance out-of-sample, therefore models that are not tested by out-of-sample data might cause harm when put into practical use.

The dataset used for empirical analysis is the U.S.A annual data collected from the WRDS Compustat dataset. The period of the sample spans from 1988 to 2012. All available firms in the U.S. will be included in the sample as long as the required variables in the models are available except financial service firms (SIC codes from 6000 to 6999). Outliers' exclusion procedure follows that in Barth et al. (2001), leaving the sample with 99845 firm-year observations. All variables are deflated by average total assets. Data before year 2005(inclusive) is used to estimate model parameters and data thereafter are used for out-of-sample comparison.

The basic criterion for the model comparison is sum squared errors (SSE). With panel data that include different individuals, SSE that is a general measure aggregating over the whole sample may not be sufficient to clearly indicate different models' performance. Therefore, average rank is also used to compare the performance of models as a second test. For each observation, the model generating the minimum magnitude of prediction error will be deemed as the best and hence ranked 1st and the opposite for the worst model that produces the largest error. The ranks of each model will be averaged over all observations, indicating the average performance of the models in making individual prediction. The model with the smallest average rank would be considered to be better performing in general.

In out-of-sample prediction, all 4 models could be applied in one-period-ahead prediction. The multiple-period prediction will be implemented in the spirit of vector autoregressive (VAR) model (see textbook [12]) that is to make predictions for all predictive variables, and then use the predicted variables recursively to predict into further periods in the future. In multiple-period prediction, BM parameters are static and thus there is no difference to one-period prediction. However, the GM have dynamic parameters, which will also be recursively determined over time. GM1 relies significantly on prediction of future sales growth rates, where recursion hence cannot be naturally extended from one-period to multiple-period ahead. Therefore, in multiple-period prediction comparison, GM1 is excluded.

V. EXPERIMENTAL RESULTS

In the dataset, all observations that are before 2005 inclusive are pooled together for estimating the factor loadings in (14). Then the intercept term is calculated individually using (15). For the implementation of (15), there is naturally a

requirement that the individual firm need to have at least 2 observations in the period to calculate mean values for the variables. For firms with only one observation, the intercept term could be still calculated, which will be equivalent as the prediction error. In the in-sample comparison, firms with one observation are excluded, but which are included back for out-of-sample prediction. Therefore, in the in-sample stage, there are 39631 firm-year observations for comparison.

The parameters of β_1 to β_6 in (14) estimated by demean and first difference methods deviate significantly from each other as Table 1 shows, especially for the autoregressive (AR) parameter β_1 . The AR coefficient estimated using the first difference method is negative, which is substantially due to the estimation procedure, which incurs negative autocorrelation bias. As a result, the intercept terms calculated based on the two methods are not close either. It is expected that the demean methods would give more reliable results than the first difference method.

It is noteworthy that in comparison with the BM model (14), the intercept term involved in GM1 and GM2 is not assumed to be constant within an observed individual. GM models hence are more flexible than the BM model as they do not place requirement on the number of observations of individual firms to calculate the intercept terms.

The Padé approximants coefficients in GM1 and GM2 are estimated by minimizing the sum squared error of the sample observations. After the coefficients are determined, the relationship of sales growth rates and firm ages with the cash flow model parameters can be calculated. Fig. 1 and Fig. 2 plot the relationship of sales growth rates and firm ages with β_1 respectively for demonstration.

In both figures, the nonlinearity is apparent, which suggest that the two selected variables, i.e. sales growth rates and firm ages tend to have some explanatory power to the model parameters, otherwise the figures would show flat straight lines. The curve in Fig.1 is generally monotonic. It suggests that higher growth in sales indicates higher AR parameter. Fig. 2 in general indicates that as time goes by, the AR parameter tends to increase except for the first two years.

Table 1

Estimation Methods	Estimated Parameters in the Linear Model					
	β_1	β_2	β_3	β_4	β_5	β_6
Difference	-0.2013	0.1709	-0.3062	0.3081	0.4493	0.2337
Demean	0.5054	0.3194	-0.4872	0.4757	0.7514	0.3657

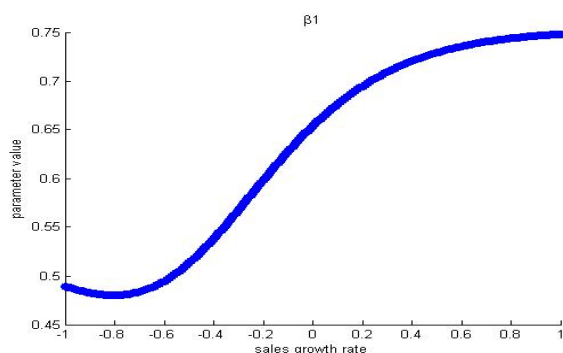


Figure 1: The relationship of AR parameter with sale growth rates.

The in-sample fitness of the four models are compared in Table 2. The first two rows denote the models in comparison. The third row contains the mean squared errors (MSE) generated by each model. BM model estimated by demean method fits the data better because of the consideration of individual effects. However BM model estimated by difference method has the poorest fit with the data, suggesting that the parameters estimated in this way contains greater bias for practical prediction and poorer fit. The GM1 and GM2 have similar data fitting ability, but poorer than the BM demean model because they do not calculate intercept terms for individual firms. The numbers in the last row in Table 2 are the average ranks of the four models. The conclusion drawn from this criterion is not different from the MSE: the BM demean method have the lowest average rank so it has the best in-sample data fitness of all these four models. BM difference method is again the worst in this measure and the two GM models are similar to each other, both lying between the two BM models.

The parameters determined in-sample are used in out-of-sample test without re-estimation. In the one-period-ahead sample, the firms that do not appear in the estimation sample are not predictable using BM models because there is no way to decide their intercept terms. However the GM models could adapt well to this situation. The one-period-ahead results are shown in Table 3. The measures are calculated based on 17965 firm-year observations. Both MSE and average rank measures show a significant outperformance of grey-box models over the benchmark models in one-period-ahead prediction. Despite the better performance of demean model over the grey-box models in the in-sample data fitting capacity, the latter have shown stronger out-of-sample performance by the two criteria. GM2

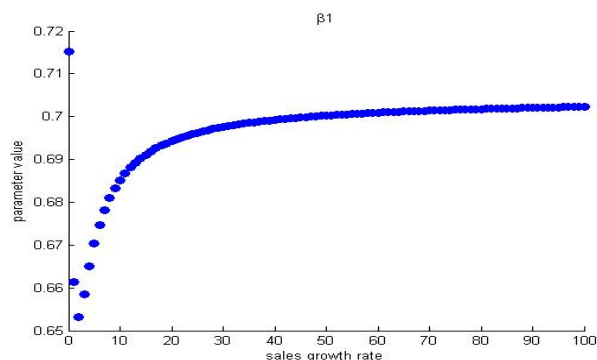


Figure 2: The relationship of AR parameter with firm ages.

Table 2

Measure of Performance	In-Sample Performance of the Models			
	GM1	GM2	BM	BM ^a
MSE	0.0075	0.0075	0.0053	0.0078
Average Rank	2.5247	2.5215	2.3760	2.5778

a. Difference method; the other BM is default to be by demean method.

that uses firm age as input to black box has the lowest MSE among the four models and GM1 has the second lowest MSE. The difference in the MSE by the two grey-box models is negligible. The benchmark models have higher MSE, especially that estimated by the first difference method. The comparison of average rank in the last row also favours the grey-box models. Both GM have lower average rank than the two BM, which suggest that the grey-box models do not only provide more accurate predictions for the sample in general, but also have better performance in predicting on individual level. It is noteworthy that GM1 and GM2 does not dominate each other as GM1 has the lowest average rank while GM2 has the lowest MSE.

In the multiple-period-ahead prediction test, all forecast are made based on the observations in year 2005. Result of year 2006 is excluded because it is not multi-period prediction. Therefore, this test could examine the predictive power of the GM2 and BM models for at most 7 years. The results are summarised in Table 4. The results are impressive and encouraging for the application of GM2, i.e. grey-box model with firm age. It has outperformed the two benchmark model again in both criteria. The MSE of GM2 is smaller than 0.02 whereas that of the benchmark models are much higher, especially the first difference model. GM2 also has the lowest average rank of the three models and thus it dominantly outperforms the benchmark models. The demean method has resulted in better practical performance than the first difference method, according to the two criteria.

VI. CONCLUSION

This paper has introduced the grey-box modelling technique to financial forecasting. This method has been shown to be powerful in engineering and its application to financial industry is therefore investigated, such as for firm-level cash flow prediction. Cash flow is not easy to predict. There are theories suggesting that it could be predicted with earnings components. In previous studies, simple linear models are developed for this application. However, it has been shown that the process of cash flow could be more complicated than linear process. Nonlinearity and dynamics that have been overlooked

Table 3

Measure of Performance	One-period Out-of-Sample Prediction Performance			
	GM1	GM2	BM	BM ^a
MSE	0.0111	0.0109	0.0140	0.0376
Average Rank	2.1283	2.2367	2.7716	2.8634

a. Difference method; the other BM is default to be by demean method.

Table 4

Measure of Performance	Multiple-period Out-of-Sample Prediction Performance		
	GM2	BM	BM ^a
MSE	0.0199	0.0244	0.0278
Average Rank	1.6530	1.9273	2.4197

a. Difference method; the other BM is default to be by demean

are the main issue addressed in this paper.

The grey-box model incorporates a black-box model to fit nonlinear data into a clear box model which explains the theoretical mechanism of the target variables. There are two grey-box models implemented in this paper in comparison with the benchmark linear models. With the assistance of Padé approximants, the grey-box model has captured some nonlinearity in the modelling system. The modelling performance is examined for out-of-sample prediction. The results have shown great improvement of grey-box model in making practical predictions of cash flows. The grey-box models outperform the benchmark models in making both one-period-ahead and multi-period-ahead predictions. The conclusion is consistent between the two criteria adopted in this paper. Therefore, the results encourage the application of grey-box modelling in the business world where nonlinear interactions are inevitable. Grey-box model could incorporate more variables that are considered to contain useful information without changing the original clear-box model structure. Therefore it could add extra power to simple linear models.

REFERENCES

- [1] Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk*. *The Journal of Finance*, 19(3), 425-442.
- [2] Campbell, J. Y., & Shiller, R. J. (1988). Stock prices, earnings, and expected dividends. *The Journal of Finance*, 43(3), 661-676.
- [3] Kalman, Rudolph Emil. "A new approach to linear filtering and prediction problems." *Journal of basic Engineering* 82, no. 1 (1960): 35-45.Sdf
- [4] Trippi, R. R., & Turban, E. (1992). *Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance*. McGraw-Hill, Inc..
- [5] Tan, K. C., & Li, Y. (2002). Grey-box model identification via evolutionary computing. *Control Engineering Practice*, 10(7), 673-684.Asdf
- [6] Hopwood, W. S., & McKeown, J. C. (1992). Empirical evidence on the time-series properties of operating cash flows. *Managerial Finance*, 18(5), 62-78.
- [7] Dechow, P. M., Kothari, S. P., & L Watts, R. (1998). The relation between earnings and cash flows. *Journal of Accounting and Economics*, 25(2), 133-168.
- [8] Barth, M. E., Cram, D. P., & Nelson, K. K. (2001). Accruals and the prediction of future cash flows. *The Accounting Review*, 76(1), 27-58.
- [9] Cheng, C. A., & Hollie, D. (2008). Do core and non-core cash flows from operations persist differentially in predicting future cash flows?. *Review of Quantitative Finance and Accounting*, 31(1), 29-53.
- [10] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4), 303-314.
- [11] Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Organization Based Intelligent Process Scheduling Algorithm (OIPSA)

¹Munam Ali Shah, ²Muhammad Bilal Shahid, ³Sijing Zhang, ⁴Safi Mustafa, ⁵Mushahid Hussain

^{1,2,4,5}Department of Computer Science, ³Department of Computer Science and Technology
^{1,2,4,5}COMSATS Institute of Information Technology, Islamabad, Pakistan

³University of Bedfordshire, Luton, UK

¹mshah@comsats.edu.com, ¹hallian20@gmail.com, ³sijing.zhnag@beds.ac.uk, ⁴safi07@live.com, ⁵mushahidh@yahoo.com

Abstract— In a multi-tasking environment, the purpose of a scheduling algorithm is to give CPU time to each process in such a way that maximum throughput could be achieved. One way is to manually set the priority of the processes but conventional operating systems put equivalent scheduling policies on each set of process and do not observe organizational preferences. It is believed that every organization performs same set of tasks most of the time. Tasks performed by an organization must be given priority according to their level of activeness rather than some hard rules defined at a design level. In this paper, we propose a novel algorithm that schedule processes according to the organization's need. Our proposed Organization Based Intelligent Process Scheduling Algorithm (OIPSA) intelligently learns the processes that are frequently used within an organization's operating system and give priority to the users' most wanted processes. The results show that OIPSA decreases response time, waiting time and turnaround time for the organization preferred processes and enhance the overall efficiency of the system when compared with conventional scheduling algorithms.

Keywords—OS process, scheduling, Organization preference, scheduling algorithm

I. INTRODUCTION

We are in an era in which everyone is moving towards making intelligent and efficient systems. Intelligence is also required at organizations' operating systems that can adapt organizational preferences and behave accordingly. In recent years, there have been quite a few improvements and efforts made towards the development of adaptive systems that make use of reflection model [1]. Reflection modelling is extensively recognized as a very effective and manageable mechanism for runtime adaption and redesign of a system [1].

In a multi-tasking or multi-programming environment, processes when fetched in the main memory are always competing for the CPU time. When one process is in execution cycle other processes are waiting for their turn or waiting for some other event to occur such as I/O operation. This

is the main reason, scheduling is important for an operating system. Scheduling decides which process will get the CPU at a given time and which process will be put in the wait condition. Some of the goals that Scheduling algorithm should accomplish to determine performance and efficiency of Scheduling algorithms are throughput, turnaround time, response time, fairness and waiting time [5][6]. Scheduling is also considered as an organized procedure through which process, thread and flow of data is controlled and is given access to different system resources [2]. Scheduling is mainly performed for load balancing and to offer quality of services [3]. All the scheduling algorithms perform some common tasks such as *multi-tasking* (to keep CPU as busy as possible) and *multiplexing* (to transmit multiple data using single physical channel) [2], [4].

Process scheduling has been always a point-of concern for operating system designers. At present, there are lots of scheduling algorithms designed, developed and implemented. Some examples are: First Come First Serve (FCFS) [12], Shortest Job First (SJB), Round Robin [RR], Multi Queue Scheduling and many more [5]. However, there will be always a need for a better and more efficient scheduling algorithm which can use CPU resources efficiently and fairly [6]. In some scenarios, an operating system is unable to select an algorithm at design level which can target organizational preferences and help in improving the throughput of the system [7][10][11]. Hard rules defined at design level cannot produce desirable results for majority of the organizations and users get same priority level issues. A similar scenario is personalized operating systems in handheld devices but there exist no personalized operating system at organizational level. In these circumstances, an organization's preference based operating system can perform better.

In this paper, a novel scheduling algorithm called Organization Based Intelligent Process Scheduling Algorithm (OIPSA) is proposed. OIPSA considers the organizations preferences and schedules the processes accordingly. The rest of the paper is organized as follows: Section II reviews the existing literature. Section III provides detailed description of the proposed OIPSA. Section IV provides performance evaluation and comparison of OIPSA with the help of

experiments and results. The paper is concluded in Section V.

II. RELATED WORK

Generally, multi-processors operating system focuses on using the maximum time of CPU and to use CPU resources efficiently for all types of processes. Oldest, easiest, fairest and most commonly used scheduling algorithms is Round Robin (RR) [8]. The aim of its design is specifically for clock/time sharing systems. In RR scheduling algorithm, multiple processes can be present in the ready queue at same time. Scheduler's job is to allocate a short amount of

account during scheduling. This will result in the more user friendly behavior of the operating system.

Another technique which uses AI focuses on the priority of relationship between activities instead of Absolute Time Quantum. It was proved that using this technique scheduling will be efficient even if set of activities are not well defined and also when the time of task to complete is large [10]. The static scheduling algorithm in [11] claims that operating system should tune its priorities accordingly [11].

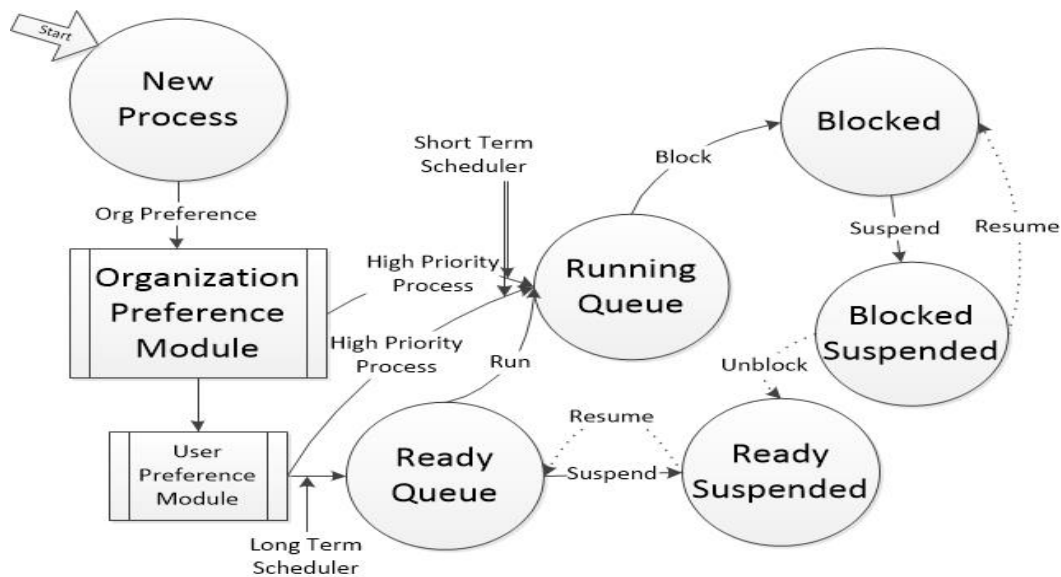


Figure 1. Process State Diagram with Organization Preference Module

time known as quantum to each process. In this way, all the processes can have CPU time fairly and can progress simultaneously [11]. Recently, a lot of work is being done in the field of CPU computational power and bandwidth of network, few studies show support of multimedia job using CPU. Nieh and Monica [9] proposed a scheduler for multimedia applications, called SMART, which is categorized the CPU time. This scheme supports conventional time sharing applications and real time system as well.

There is also a significant amount of research done in the field of process scheduling using artificial intelligence techniques such as fuzzy logic. One of the techniques for scheduling using fuzzy logic was to decide time quantum for those jobs which are neither too big nor too small. In this way, every job gets a reasonable CPU time and also, the throughput of the system does not reduce due to needless context switching [5]. Sungsoo Lim and Sung-Bae Cho [7], proposed that scheduling must be done according to the type of processes rather than using uniform scheduling policy for each set of process. They also suggested that user preference must be taken into

III. ORGANIZATION BASED INTELLIGENT PROCESS SCHEDULING ALGORITHM (OIPSA)

This section highlights salient features of our proposed OIPSA algorithm. We start our discussion by first presenting the hypothesis for the proposed algorithm and then present the methodology and design for OIPSA.

A. Hypothesis

We believe that each organization perform certain tasks repeatedly and regularly. The users within that organization perform tasks according to their need in general and organizational needs in particular. Using this assumption, we assign priorities to the certain organizational processes. This results in a better working environment and more optimal results than the one which are given by conventional scheduling algorithms. We believe that it is difficult to handle individual processes. To validate our assumption, we perform simulations on different scheduling algorithms and our proposed OPISA. We assume that organization preferred processes are already assigned top priority and are stored in the system. Our

assumption is based on the Windows operating system which stores application preferred data in the following registry:

HKEY_LOCAL_USER\MACHINE\Software\Microsoft\Windows.

B. Methodology

Simulation experiments and results are used to prove that proposed algorithm performs better in most of the cases. The input values are provided in Table 1. For our simulation, we use open source simulator. A file is taken as an input which is used as a parameter and is run for the proposed OIPSA and for other conventional scheduling algorithms. After this, we compare results and evaluate the performance of each of the algorithm with OIPSA.

C. Design

When a new process is started, it becomes an input for the A new process when started, will become input of Organization Preference Module as shown in Figure. 1. Organization Preference Module considers some predefined rules for new processes. On the basis of these rules, the organizational preference module put the process in one of three queues, i.e., *high priority process queue*; *medium priority process queue* and *low priority process queue*. Long-term Scheduler decides which process should be in ready queue based on multiple priority queue. As the new process is placed in the ready queue, short-term scheduler checks if the process submitted by the long-term scheduler has more priority than the process which is currently running in the running queue. If the recently submitted process doesn't have priority over the running process, then it waits. Otherwise, an interrupt is sent to the CPU. The process that was in the running state is sent to the blocked queue and the newly submitted process is given to the CPU. In Figure 1, it can be observed that the Organization Preference Module can send process to running queue if it is up to certain level, and sends to User Preference Module, which can send process directly to running queue if a priority of new process is up to certain level of user importance. This ensures that whenever Organization Preferred Process is created, it gets CPU in no time reducing response time, turnaround time and waiting time.

As discussed above, conventional operating systems usually schedule each and every process uniformly without taking notice of Organization's Preferences, or User's Preferences, because they are hard coded at design level. The designers of the operating system cannot decide target audience of operating system as the user of operating systems are very versatile in nature. Every organization uses operating system for different uses. Some use it for handling complex tasks, while some use it for information processing. Some users use the operating system for software development purposes and few

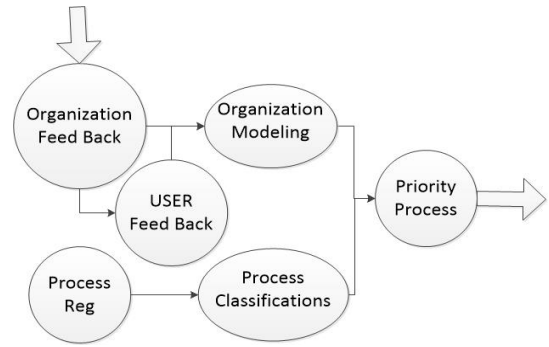


Figure 2. Core Modules of the OIPSA

use it for routine tasks. Scheduling decided at design level can be better in certain cases but at the same time it can be devastating for some other organizations. While considering user preferences, the single system will work according to user needs. This will overcome the drawback of prioritizing the preferences in a static fashion. Our proposed OIPSA adjusts the user needs and organizational preferences in a dynamic way.

OIPSA is comprised of three major modules as shown in Figure. 2. The *Process Classification* module, which classifies a process into high, medium and low priority processes. Second module is, *Organization Modelling* module, which models the organization's most used processes or organization preferences. This module also checks the user feedback and take organization feedback into consideration. Third module is *Priority Process* module which classifies the priority of each process at run time. The input for the priority module is based on the output from Priority Classification and Organization Modelling modules.

D. Process Classification

Figure. 3 depicts the processes within Priority Classification module. For the Priority Classification, we are keeping a track of processes and number of times they have been a resident in the job queue. The more identical the processes are in job queue, the more they will get priority. Also, according to some specified rules they will be placed in either High Priority Pool, Medium Priority Pool or Low Priority Pool.

E. Process Priority Decesion

In routine, the operating system schedules processes at system level without taking into account

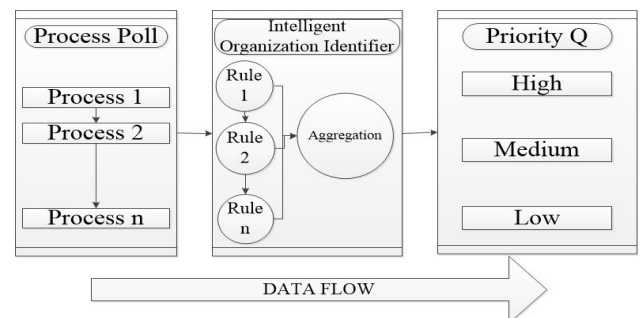


Figure 3. Priority classification module

the organization's liking and disliking. Therefore, if an organization wishes to alter the scheduling policy, the organization must change priority by themselves or they must reset the operating system. Both these tasks are time taking and hard for novice users. Using

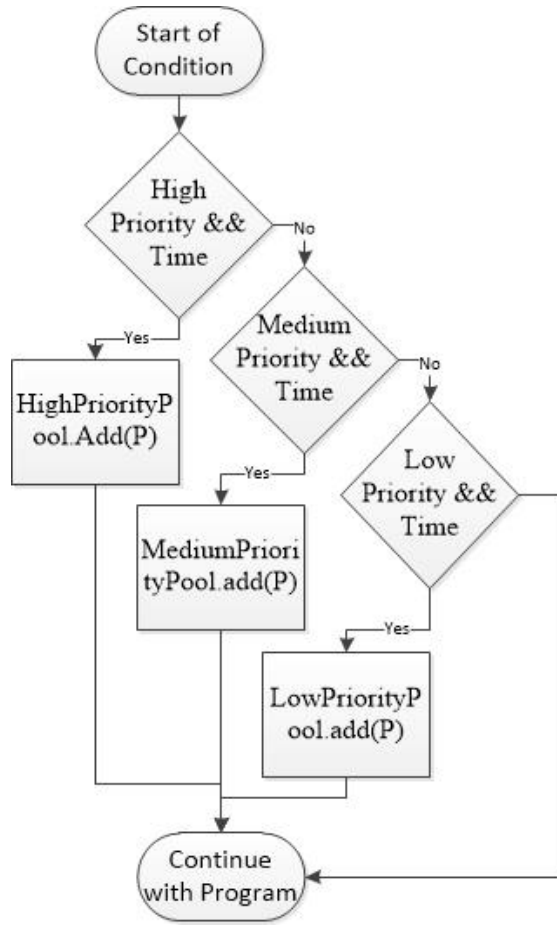


Figure 4. Flowchart of Priority Scheduling in OIPSA

OIPSA, an operating system can keep the users of an organization satisfied without the need of changing priorities or resetting operating system by themselves. The OIPSA only needs to model preferences of organization. If organization is satisfied, the users comes under its umbrella automatically.

F. Implementation

We combine our technique of assigning priorities with one of the implemented scheduling algorithms, i.e., Priority Scheduling Algorithm [11]. The flow chart of the proposed OIPSA is provided in Figure. 4.

Table 1: Input for OIPSA

Process	Burst Time	Priority
1	10	5
2	20	2
3	30	3

IV. EXPERIMENTAL RESULTS

In this section, we perform experiments on our proposed OIPSA. We take a set of three processes with different priority values and different burst times. Table. 1 shows the processes and the associated values. For the first experiment, we choose a data source that is suitable for FCFS algorithms. Same is repeated with opposite values and priorities. It could be observed that jobs with the short burst time are scheduled first. In this case, FCFS performs better when jobs with the shorter burst time are first and jobs with the longer burst time are at the end of the queue. This reduces turnaround time, waiting time and response time.

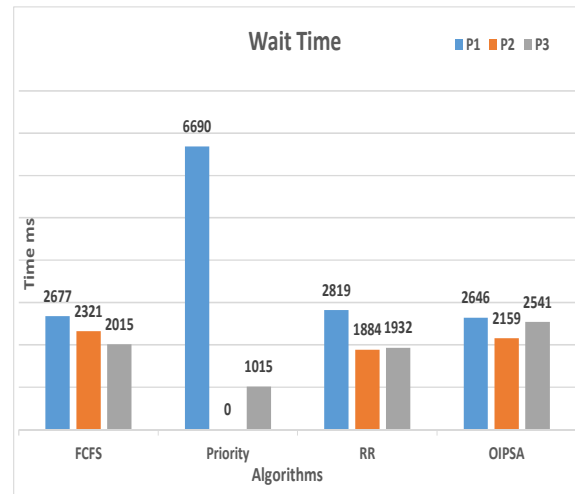


Figure 5. Waiting time of Individual Processes (P=Process)

A. Waiting time of Individual Processes

From above chart (Figure. 5), we can see that FCFS has performed pretty well for all the processes, as they have very low waiting time. Round Robin (RR) performed satisfactory for all the processes. If we observe the performance of our proposed OIPSA algorithm, we see the results obtained closely match with that of FCFS algorithm. This is because OPISA doesn't give same quantum time to all the processes. Higher priority process gets more time than lower priority process and that is why its waiting time is very close to the FCFS for this set of input.

B. Response Time of Individual Processes

In Figure 6, we can observe that response time of Round Robin (RR) for all the processes is comparatively lowest because of its fair quantum time to all the processes. For Priority algorithm, the processes with more priority get CPU very frequently then the processes with low priority. The response time in FCFS is better as shorter processes are first and longer burst time processes are latter.

If we observe the performance of the proposed OIPSA, it could be noticed that the OIPSA is not even close to RR algorithm. The reason is that OPISA does not give fair quantum time for all the processes. The higher priority will get more quantum time that is why

response time for each process is higher than RR Algorithm.

C. Turnaround Time of Individual Processes

From Figure 7, we can see that the turnaround time for the FCFS is satisfactory as short CPU burst

job are at start of queue. High CPU burst jobs are at the end of queue. P2 for priority algorithm has the least turnaround time as of lowest priority process. RR algorithm does not perform better when it comes to turnaround time as it believes in fairness.

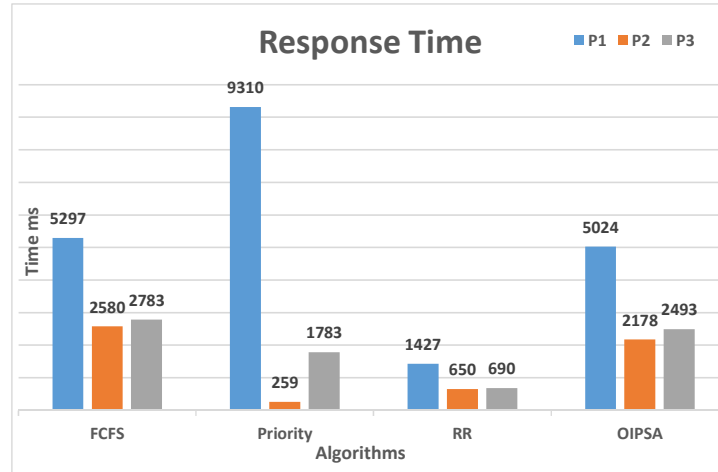


Figure 6. Responce time of Individual Processes (P=Process)

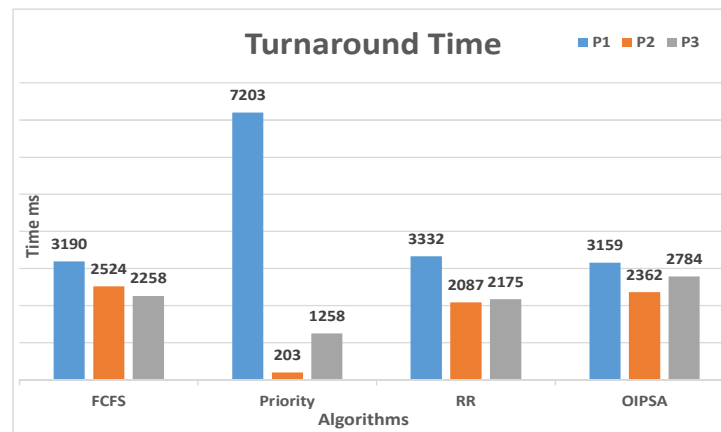


Figure 7. Turnaround Time of Individual Processes (P=Process)

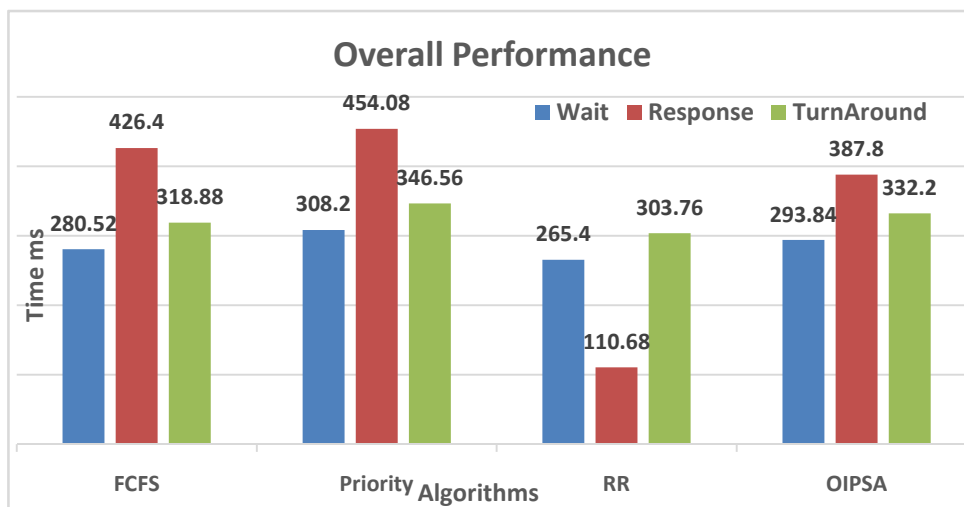


Figure 8. Overall Performance of all Algorithms

Notice the performance of the proposed OIPSA in Figure 7, we see that it performed well for P1 when compared with the Priority Algorithm as its priority increased as the time passes. P2 turnaround time is higher than P2 of Priority, as of low priority. In this case too, OIPSA is still very close to the FCFS.

D. Overall Performance of all Algorithms for the same data

The average values of all the experiments for the discussed algorithms have been computed and plotted in Figure 8. It could be observed that on average, FCFS gives optimal values and the proposed OIPSA is pretty close to FCFS.

V. CONCLUSION

In this paper we proposed a novel algorithm which uses organization's preferences as a basis for scheduling. Our proposed algorithm classifies each process into *high*, *medium*, and *low* priority pool and then depicts the organization preferences. Comparing with existing algorithms and user preference algorithms, OIPSA gives better results. It schedules according to organization's needs. User personal preferences are ignored as user comes under the umbrella of an organization. An important point to note here is that the performance of OIPSA at start will be just satisfactory but as OIPSA learns organization's preferences, it will perform much better and will enhance overall efficiency of the system according to organization's needs. In future, we will be adding more preferences. User preference will be encapsulated in organization preferences. We aim to enhance the existing algorithms to meet the organization needs. A hybrid approach comprising of organization's need and user preferences will also form part of our future work.

ACKNOWLEDGMENT

The principal author would like to acknowledge the grant of funding by the Higher Education Commission (HEC), Pakistan to present this paper.

REFERENCES

- [1] G. Coulson, G. Blair, and P. Grace, "On the performance of reflective systems software," IEEE Int. Conf. Performance, Comput. Commun. 2004, pp. 763–770.
- [2] A. Silberschatz, P. B. Galvin, and G. Gagne, "Operating system concepts", 1998.
- [3] B.A. Shirazi "Introduction to Scheduling and Load Balancing," 1990.
- [4] A. A. Aburas and V. Miho, "Fuzzy Logic based algorithm for uniprocessor scheduling," in 2008 International Conference on Computer and Communication Engineering, 2008, pp. 499–504.
- [5] P. K. Varshney, N. Akhtar, and M. F. H. Siddiqui, "Efficient CPU Scheduling Algorithm Using Fuzzy Logic," vol. 47, no. Iccts, pp. 13–18, 2012.
- [6] A. M. Wang, "Fuzzy-Based Scheduling Algorithm," pp. 1–12, 2010.
- [7] S. Lim and S. Cho, "Intelligent OS Process Scheduling Using Fuzzy Inference with User Models," pp. 725–734, 2007.
- [8] J. Nieh, "2001 USENIX Annual Technical Conference," 2001.
- [9] "The Design, Implementation and Evaluation of SMART: A Scheduler for Multimedia Applications." [Online]. Available: <http://suif.stanford.edu/papers/nieh97b/paper.html>. [Accessed: 06-Apr-2014].
- [10] P. Laborie, "Algorithms for propagating resource constraints in AI planning and scheduling: Existing approaches and new results," Artif. Intell., vol. 143, no. 2, pp. 151–188, Feb. 2003.
- [11] C. L. Liu and J. W. Layland, "Scheduling algorithms for multiprogramming in a hard-real-time environment," in *Journal of the ACM*, vol. no. 20, no. 1, pp. 46–61, 1973.
- [12] U. Schwiegelshohn and R. Yahyapour, "Analysis of first-come-first-serve parallel job scheduling," *SODA*, 1998.

An Improved Search Space Resizing Method for Model Identification by Standard Genetic Algorithm

*Kumaran Rajarathinam, J. Barry Gomm, DingLi Yu and Ahmed Saad Abdelhadi
Mechanical Engineering and Materials Research Centre (MEMARC), Control Systems Group,
School of Engineering, Liverpool John Moores University,
Byrom Street, Liverpool, L3 3AF, UK
*K.Rajarathinam@2011.ljmu.ac.uk

Abstract—In this paper, a new improved search space boundary resizing method for an optimal model's parameter identification by Standard Genetic Algorithms (SGAs) is proposed and demonstrated. The premature convergence to local minima, as a result of search space boundary constraints, is a key consideration in the application of SGAs. The new method improves the convergence to global optima by resizing or extending the upper and lower search boundaries. The resizing of search space boundaries involves two processes, first, an identification of initial value by approximating the dynamic response period and desired settling time. Second, a boundary resizing method derived from the initial search space value. These processes brought the elite groups within feasible boundary regions by consecutive execution and enhanced the SGAs in locating the optimal model's parameters for the identified transfer function. This new method is applied and examined on two processes, a third order transfer function model with and without random disturbance and raw data of excess oxygen. The simulation results assured the new improved search space resizing method's efficiency and flexibility in assisting SGAs to locate optimal transfer function model parameters in their explorations.

Keywords—search space boundary resizing; predetermined time constant approximation; genetic algorithms; convergence constraints; premature convergence; transfer function model identification.

I. INTRODUCTION

One of the most common problems that may be encountered during model's or control's parameters optimisation by optimisation algorithms is premature convergence due to search space boundary constraints. An optimisation process has prematurely converged to a local optimum if it is no longer able to explore other parts of the search space region than the area currently being explored and there exists another region that may contains a superior solution [1]. Particularly, a set of transfer function parameter's to be optimised for a continuous higher order model distinguishes the dynamic characteristics of the system. At present, numerous well known algorithms and

techniques are in application for improving the search space boundary constraints.

Standard Genetic Algorithms (SGAs) are unsophisticated and a very promising approach of an evolutionary computation method in identification of model's parameters. Though, premature convergence is still attributable to the searching space constraints and is a common phenomenon in SGAs [2]. Without prior knowledge of a model's parameters or time constant values, it is highly infeasible to predict the search Upper Space Boundary (SB_{Upper}) and Lower Space Boundary (SB_{Lower}). Especially, when the optimum values are located near to the boundary region or outside the boundary region. Insignificant numbers of researches are involved in improvement of searching space to an optimal solution. Based on the complex Box technique, a boundary search method for optimisation problems in the case of the optimal solution at the boundary was proposed [3]. It has been demonstrated and verified, if there is an optimal solution at the boundary constraint set.

Recently, a modified GAs is applied in solving the n-Queens difficulty in chessboard [4]. The holism and random choices cause solving difficulties for SGAs in searching a large space. To improve the solving difficulty, the minimal conflicts algorithm is collaborated with SGAs. The minimal conflicts algorithm gives a partial view for SGAs by a locally searching space. But, the collaboration of algorithms consumed time for searching. A new approach called the self-adaptive boundary search strategy for penalty factor selection within SGAs was proposed [5]. This approach guides the SGA to preserve around constraint boundaries and improves the efficiency of attaining the optimal or near optimal solution. A technique for resolving the structural optimisation difficulties in quantising the subjective uncertainties of active constraints is proposed by fuzzy logic formulation [6]. Another method to improve the prematurity and to sustain the diversity population was proposed by Niche Genetic Algorithm (NGM) associated with isolation mechanism [7]. A comparison study was done on NGM and Annealing Genetic Algorithm where the Annealing Genetic

Algorithm has better premature convergence [8]. However, the Annealing Genetic Algorithm is time consuming by extra procedures. Another method, named Accelerating Genetic Algorithm (AGM) was proposed to resize the feasible region into the elite individual's adjacent region for better local searching and convergence [9]. Search space boundary reduction for the candidate diameter for each link by pipe index vector and critical path method, along with modified genetic operator's derivatives, was proposed [10] [11]. Further, an improved AGM based on the saddle distribution by which adding random individuals into the initial population to increase the searching ability of optimal solution was proposed [12].

A literature review discloses that most researched techniques are considered based on limited or confined search space boundaries and has an initial knowledge of search space parameters. Also, the discussed research information involves complex mathematical approaches and inevitably can be time consuming for convergence. This paper proposes and investigates a new improved search space method, named the predetermined time constant approximation (T_{sp}) to enhance the SGAs exploration and exploitation towards the global optima. This method employs a novel search space boundary extension technique by T_{sp} , which guides the search to concentrate on optimal values within the boundaries of the feasible region of the solution space. The structure of this paper is as follows; first, the SGAs convergence states for an optimal value by search space boundary constraints are discussed. Second, the approximation process of predetermined time constant methods is discussed. Further, search space boundary extensions for better exploration and for optimal exploitation are discussed here. Finally, the effectiveness of the T_{sp} method is assessed with two processes. Also, the improved AGM based on the saddle distribution method is compared with excess oxygen model to measure the effectiveness of proposed method. The proposed methods are developed and tested in simulations based on Matlab/Simulink models.

II. POLYNOMIAL COEFFICIENTS

Consider a system can be modelled by the general order differential equation,

$$a_n \frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \dots + y = K_p f(t - \theta) \quad (1)$$

where $f(t - \theta)$ is the input signal or forcing function with time delay, $y(t)$ is the output signal and K_p is process gain. Assuming zero initial condition, $y(0)=0$, $y'(0)=0$, and taking the laplace transform of equ. 1 gives the general order transfer function is of the form,

$$G(s) = \frac{Y(s)}{F(s)} = \frac{K_p}{a_n s^n + a_{n-1} s^{n-1} + \dots + a_1 s + 1} e^{-\theta s} \quad (2)$$

where $a_n \dots a_1$ are coefficients of the denominator polynomial. The denominator polynomial coefficients provide a foundation for determining a system's dynamic response characteristics. In particular the system's poles directly define the components in

the homogeneous response. Thus, optimal poles identification is primarily considered here.

III. CONVERGENCE CONSTRAINTS OF SGAs BY SEARCH SPACE BOUNDARY

In most situations, selecting the search space boundary regions is delicate if there is no prior knowledge of optimum value location. Thus, a randomly selected search space boundary is a significant factor which leads the SGAs to often converge and get trapped in local optima, resulting in suboptimal solutions. Particularly, it locates near the boundary or outside of the boundary.

As illustrated in Fig. 1, the SGAs convergences by search space boundary constraints can be classified by three states;

- State 1 – If the optimal value (X_i) is located within uniformly distributed elite group around boundary region $[X_i - \Delta_{GO}, X_i + \Delta_{GO}]$, the genetic operators have higher probability of converging to global optimum. Thus, the randomly generated initial population within well distributed elite group search boundary has higher probability exploring and exploiting a better parent chromosome. Further, the selected parent chromosome will be evaluated by genetic precision process (selection, crossover and mutation) to produce fitter offspring without any convergence constraint.
- State 2 – If the X_i is located near $([SB_{Lower}, X_i - \Delta_{GO}], [X_i + \Delta_{GO}, SB_{Upper}])$, the SGAs possibly will converge to local minima. The elite group which is distributed near the boundary may have located a part of the elite group at the outer boundary. If the elite group at the outer part have the genetic information of an optimal value, the genetic operators will suffer to exploit the optimal value and the exploration process will retard. As a result, the search space boundary constraints will lead the SGAs to converge to local minima.
- State 3 – If the X_i is located outside the boundary region $[SB_{Lower} > X_i > SB_{Upper}]$, the SGAs will fail to explore and exploit the optimal value. The simulation may be retarded and stopped.

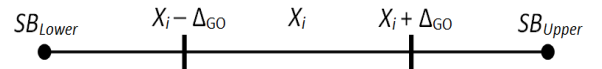


Fig. 1. Schematic diagram of feasible search space boundary region

where SB_{Lower} is lower search boundary, SB_{Upper} is upper search boundary and Δ_{GO} is the genetic operator for convergence precision.

IV. PREDETERMINED TIME CONSTANT APPROXIMATION

By approximating the distribution of the elite group in a boundary region at the initial stage, gives the genetic operators opportunity to locate the optimal value rapidly without any constraint. To improve searching space boundaries for optimal model identification, a straightforward trial and error technique without a mathematical constraint is introduced here, named

predetermined time constant approximation (T_{Sp}). The approximation process can be simplified as follows;

- Selecting δT_s , where δ is the settling band in %. ($\delta = 3, 4$ and 5). The selection of desired δ is according to the raggedness of dynamic response.
- Estimating process's dynamic response period ($DR_{P(\tau_2-\tau)}$). At $C(t) = 0_{(T=\tau)}$ to $C(t) = 1 \pm \delta(\%)/(T=\tau_2)$. ($C(t)$ is desired settling point.)
- Approximating an initial $\tau_I = DR_{P(\tau_2-\tau)} / \delta$.
- Calculating initial T_{Sp} by identified τ_I according to the respective transfer function coefficients ($a_n s^n + a_{n-1} s^{n-1} + \dots + a_1 s + 1$).

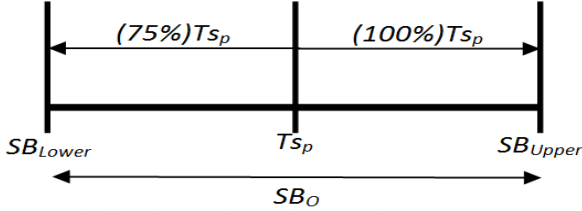


Fig. 2. Optimising the search space boundary by T_{Sp}

$$SB_O = \{SB_O; SB_{Lower} \leq T_{Sp} \leq SB_{Upper}\} \quad (3)$$

As illustrated in Fig. 2 and equation 3, the SB_O is optimum search space boundary, SB_{Lower} is lower search boundary and SB_{Upper} is upper search boundary. For an SB_O , the SB_{Upper} and SB_{Lower} are extended by 100% and 75% from initial T_{Sp} , respectively. Especially, 100% of extension for SB_{Upper} is required as the optimal solution can be mostly located near to the upper boundary region. Such an extensive search space extension is required for SGAs to explore the elite groups within boundaries and to exploit the X_i without any constraint while maintaining the population diversity. Also, such an extension required for characterizing the dynamic homogenous of higher order model parameters.

Generally, the all process of search space boundary adjustment and an optimal X_i identification can be stated as follows;

1. Initial attempt – Identified T_{Sp} according to the respective transfer function coefficients are applied with 100% extension on SB_{Upper} . The SB_{Lower} is extended to approximately 95% from initial T_{Sp} instead of 75% for better exploration at the beginning stage. Execute the SGAs.
2. Second attempt – Genetically identified T_{Sp} of respective transfer function coefficients by initial attempts are extended accordingly (SB_{Upper} to 100% and SB_{Lower} to 75%) to optimise SB_O . Execute the SGAs.
3. Subsequent attempt – Continuing the SGAs execution with unchanged boundary search approximation by second attempt, until optimal X_i and minimum sum of square error (SSE) attained.
4. *Subsequent attempt – If the extended boundary in second attempt is not a SB_O , consecutive boundary adjustment is

essential until SB_O is achieved. Then, continuing the SGAs execution until optimal X_i and SSE attained.

V. EXPERIMENTAL RESULTS

To illustrate the non-complexity and effectiveness, the proposed time constant approximation method is applied on two example processes; a 3rd order transfer function and real data from a process step response.

A. Process 1 – 3rd Order Transfer Function

For simulation study, the transfer function of a 3rd order process is selected with process gain ($K_p = 10$),

$$G(s) = \frac{10}{15s^3 + 78s^2 + 6s + 1} \quad (4)$$

The particular motive of selecting this 3rd order transfer function is that it has a real pole at -5.1245 and a pair of complex poles at $-0.0378 \pm 0.1076i$ which are exhibiting a significant oscillatory response. Also, to assess the T_{Sp} method's flexibilities and effectiveness, the 3rd order transfer function coefficients are moderately small parameters. So, an appropriate search space boundary extension is required.

According to the 3rd order process step response (Fig. 3), the $DR_{P(\tau_2-\tau)} = 123s - 0s = 123s$. Selecting $\delta T_s = 5T_s$, as the desired T_s is 1% settling band, gives the initial τ_I is 24.6s. Therefore, the T_{Sp} for the 3rd order polynomial coefficients can be approximated by,

$$\begin{aligned} \tau_1 s &= 24.6; - \rightarrow (T_1 s)^3 + 3(T_1 s)^2 + 3T_1 s + 1 \\ &= 14887 s^3 + 1815.5 s^2 + 73.8 s + 1 \end{aligned} \quad (5)$$

According to table 1, the SGAs explored well the entire search space boundaries and exploited the elite group within the chosen boundary region $[X_i - \Delta_{GO}, X_i + \Delta_{GO}]$ for T_{Sp} values of S^2 and S^I at the initial attempt. This can be seen by the consistency of the T_{Sp} values of S^2 and S^I in further execution with readjusted boundaries at the 2nd attempt. This has enhanced the exploitation of an optimal X_i at each subsequent attempt by the SGAs.

On other hand, the simulation results reveal that the elite group of T_{Sp} values of S^3 are distributed near to SB_{Lower} region. This is clearly noticeable at the 1st, 2nd and 3rd execution results that the T_{Sp} value of S^3 is remaining around SB_{Lower} . This caused the SGAs to fail to exploit an optimal X_i and converge to local minima as a part of the elite group is located outside of SB_{Lower} (state 2). As a result, 3 adjustments on boundaries, especially on SB_{Lower} are required to optimise the SB_O and to bring the elite groups within a feasible boundary region. As expected, the boundaries are optimised and the elite groups are explored well at the 4th execution. Further SGAs execution enhanced an optimal X_i exploitation.

The flexibilities and effectiveness of T_{Sp} methods is further assessed on 3rd order transfer function model with 5% disturbance. Initially, identified transfer function coefficients without the disturbance are applied on the 3rd order model with

disturbance. The simulation result in Fig. 4 and table 2 reveals that the exploration of elite groups and exploitation of an optimal X_i for the 3rd order model with disturbance is a very similar process to without disturbance. Thus, the effectiveness of T_{Sp} method is well demonstrated in optimizing the SB_O and exploiting the X_i with or without disturbance. Based on

minimum SSE , the selected 3rd order model transfer function without disturbance is;

$$G(s) = \frac{9.997}{21.98s^3 + 77.68s^2 + 6.197s + 1} \quad (6)$$

TABLE I. SIMULATION RESULTS OF 3RD ORDER TRANSFER FUNCTION EXECUTIONS

Execution	S^3		S^2		S^1		$T_{Sp} (S^3)$	$T_{Sp} (S^2)$	$T_{Sp} (S^1)$	SSE	Gen
	SB_U	SB_L	SB_U	SB_L	SB_U	SB_L					
1	29774	10	3630	10	148	2	141.3	76.75	7.439	60.092	70
2	280	35	150	20	15	2	42.55	77.73	6.281	8.4924	50
3	85	12	150	20	15	2	23.25	77.67	6.182	7.7894	30
4	50	5	150	20	15	2	22.98	77.69	6.179	7.7899	20
5	50	5	150	20	15	2	21.23	77.67	6.157	7.8149	20
6	50	5	150	20	15	2	22.18	77.67	6.189	7.7915	30
7	50	5	150	20	15	2	21.98	77.68	6.197	7.6025	25
8	50	5	150	20	15	2	21.41	77.69	6.171	7.6171	35
9	50	5	150	20	15	2	23.53	77.67	6.186	7.7898	25
10	50	5	150	20	15	2	22.62	77.68	6.175	7.7914	15
11	50	5	150	20	15	2	23.49	77.69	6.183	7.7895	20

TABLE II. SIMULATION RESULTS OF 3RD ORDER TRANSFER FUNCTION WITH 5% DISTURBANCE EXECUTIONS

Execution	S^3		S^2		S^1		$T_{Sp} (S^3)$	$T_{Sp} (S^2)$	$T_{Sp} (S^1)$	SSE	Gen
	SB_U	SB_L	SB_U	SB_L	SB_U	SB_L					
1	29774	10	3630	10	148	2	380.4	82.03	11.27	150.832	90
2	760	95	165	20	22	3	95.15	77.78	6.296	60.1486	78
3	190	24	155	20	13	2	25.29	77.57	6.211	33.4558	43
4	50	6	155	20	13	2	24.02	77.57	6.196	33.4456	37
5	50	6	155	20	13	2	24.67	77.58	6.049	33.4481	32
6	50	6	155	20	13	2	24.05	76.33	6.398	33.4452	28
7	50	6	155	20	13	2	26.14	77.91	6.215	33.4627	22
8	50	6	155	20	13	2	24.25	77.51	6.198	33.4459	30
9	50	6	155	20	13	2	22.99	77.58	6.186	33.4503	21
10	50	6	155	20	13	2	22.89	77.58	6.183	33.4511	42
11	50	6	155	20	13	2	22.76	77.84	6.114	33.4596	34

TABLE III. SIMULATION RESULTS OF EO₂ EXECUTIONS

Execution	S^3		S^2		S^1		$T_{Sp} (S^3)$	$T_{Sp} (S^2)$	$T_{Sp} (S^1)$	SSE	Gen
	SB_U	SB_L	SB_U	SB_L	SB_U	SB_L					
1	3.5e6	10	8.6e4	10	7.2e2	10	8088.2	10085	178.73	0.86796	70
2	1.6e4	2e3	2e4	2e3	3.5e2	40	4039.7	14074	180.02	0.49128	20
3	1.6e4	2e3	2e4	2e3	3.5e2	40	2699.7	13304	180.38	0.51873	40
4	1.6e4	2e3	2e4	2e3	3.5e2	40	4875.7	14995	183.64	0.49413	40
5	1.6e4	2e3	2e4	2e3	3.5e2	40	8187.7	14524	181.41	0.48654	20
6	1.6e4	2e3	2e4	2e3	3.5e2	40	8079.1	16513	184.16	0.53421	35
7	1.6e4	2e3	2e4	2e3	3.5e2	40	4330.5	14555	177.2	0.5109	90
8	1.6e4	2e3	2e4	2e3	3.5e2	40	4137.2	15028	181.88	0.48758	22
9	1.6e4	2e3	2e4	2e3	3.5e2	40	9903.9	16043	182.3	0.51771	80

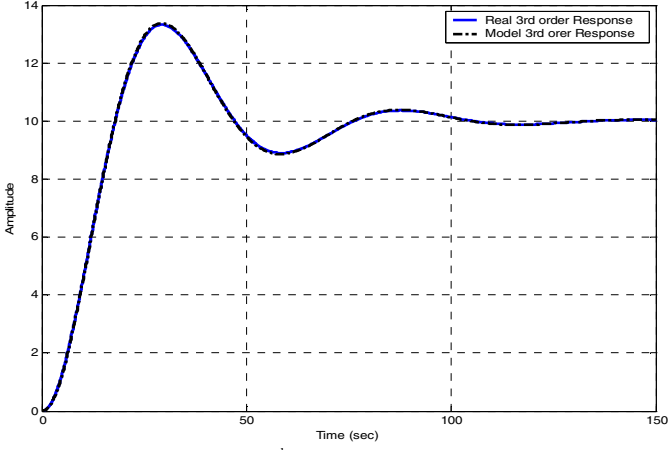


Fig. 3. Transient response of 3rd order transfer function real and model process

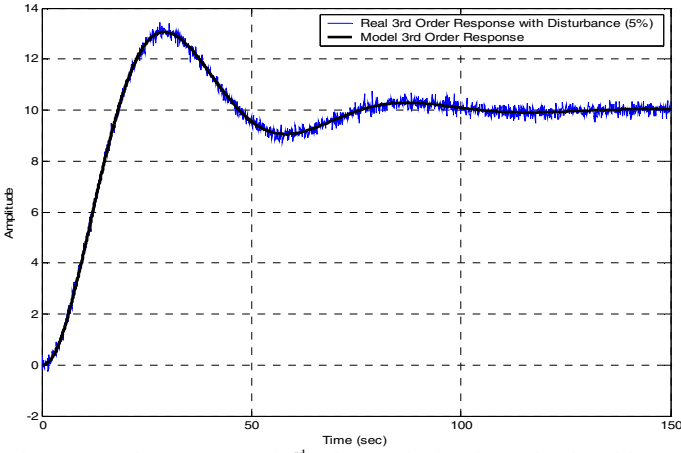


Fig. 4. Transient responses of 3rd order transfer function real and model with 5% disturbance

and with 5% disturbance is;

$$G(s) = \frac{9.976}{24.05s^3 + 76.33s^2 + 6.398s + 1} \quad (7)$$

By comparing the identified T_{Sp} coefficients with 3rd order transfer function model's parameters, the S^2 and S^I values have 98% similarity. But, the S^3 value only has 54% of similarity. According to table 4 and Fig. 5, the complex poles of all 3rd order models illustrate that the imaginary parts are considerably constant. But, the real part is slightly moved along the real axis causing a small change in the damping ratio for these roots. These small changes in the complex poles are consolidated with the differing position of the other real root.

TABLE IV. ROOTS OF 3RD ORDER MODEL'S

Model's	S^3	S^2 & S^I	Damping Ratio
Real	-5.1245	-0.0378±0.1076i	0.331
Without Disturbance	-3.4564	-0.0389±0.1079i	0.339
With Disturbance	-3.0921	-0.0408±0.1085i	0.352

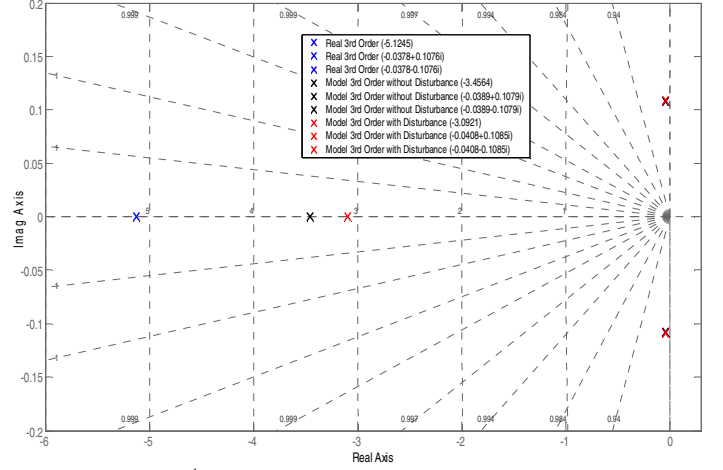


Fig. 5. Roots of 3rd order transfer function real and model with 5% disturbance

Notice that the peak time is the same for all waveforms because the imaginary part remains the same. Nevertheless, the identified model responses, with and without noise, closely match the response of the actual system as shown in Figs. 3 and 4.

B. Process 2 – Excess Oxygen (EO_2)

A raw numerical data of excess oxygen (EO_2) is collected from a real industrial furnace by empirical technique for 1000 seconds with 5 seconds interval. As illustrated in Fig. 6, the process response of EO_2 is exhibiting an approximate first-order plus dead-time (FOPDT) dynamic system. The data was gathered by the step input of increasing air ratio from 9.5 to 10.5 in volumetric.

As discussed earlier, the time constant (τ_s) of transfer function are primarily considered here for optimal model identification by T_{Sp} method. Whereas, the process gain (K_p) and transport delay (θ) can be approximated by close observation of the transient response. As illustrated on the transient response of EO_2 , $K_p \approx 1.54$ and $\theta \approx 160s$. As a result, an extension on the search space boundaries are approximated for $K_p \in [1 : 2]$ and $\theta \in [50 : 200]$.

According to the EO_2 response, the $DR_{P(\tau_2-\tau_1)} = 700s - 100s = 600s$. Selecting $\delta T_s = 5T_s$, as the desired T_s is 1% settling band, gives the initial τ_1 as 120s. For EO_2 , the selection of an optimal

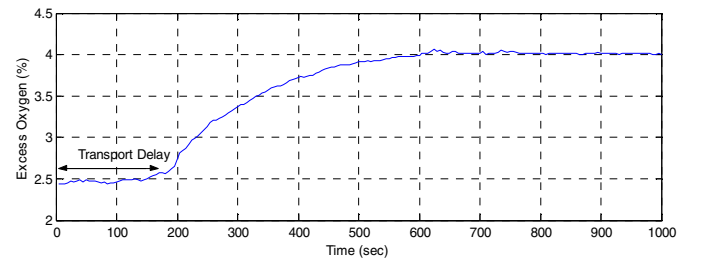


Fig. 6. Step response of EO_2

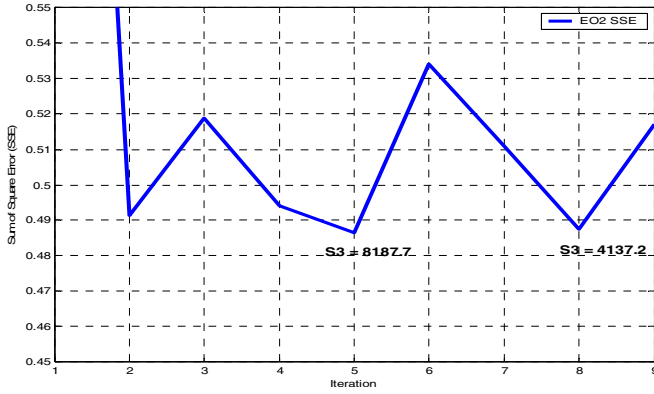


Fig. 7. Two optimal values of S^3 for EO_2

model is a 3rd order transfer function. Therefore, the T_{Sp} for the 3rd order polynomial coefficients can be approximated,

$$\begin{aligned} \tau_1 s &= 120 ; - \rightarrow (T_1 s)^3 + 3(T_1 s)^2 + 3T_1 s + 1 \\ &= 1.728 e^6 s^3 + 4.32 e^4 s^2 + 3.6 e^2 s + 1 \end{aligned} \quad (8)$$

As illustrated in table 3, the distribution of elite groups within boundary region $[X_i - \Delta_{GO}, X_i + \Delta_{GO}]$, the exploitation of optimal X_i and the consistency of the T_{Sp} values of S^2 and S^l in further execution by SGAs are exhibiting similar process characteristics as 3rd transfer function model.

Based on the initial attempt, the elite groups of T_{Sp} value of S^3 are uniformly distributed around $X_i - \Delta_{GO}$ region. As illustrated in table 3, the T_{Sp} value of S^3 is still continuously evolving within the boundary SB_O region at each execution. Therefore, further readjustment of SB_O boundaries is not required as the elite groups are still within the boundary range (state 1) as discussed in section 3. For this 3rd order model of EO_2 , the T_{Sp} values by the 5th execution are selected as the SSE and Gen (generation) is minimum and optimal. The identified transfer function is,

$$G(s)_{EO_2} = \frac{1.555}{8187.7s^3 + 14524s^2 + 181.41s + 1} e^{-109.36s} \quad (9)$$

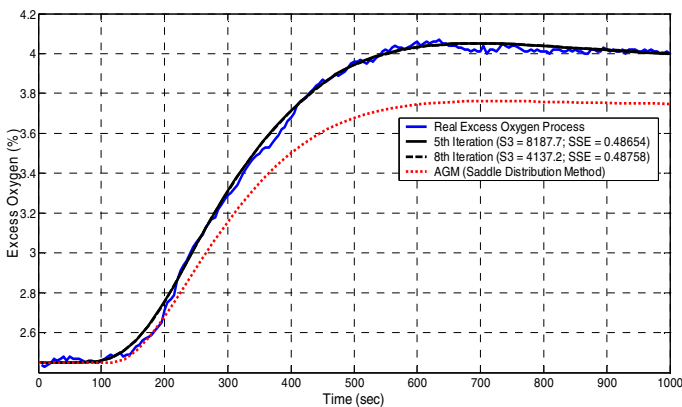


Fig. 8. Transient responses of 2 global optimal values with real process of EO_2

However, the inconsistency of S^3 shows that there are two optimal values of X_i ($X_i = 8187.7; 4137.2$), which frequently appear within the SB_O region at 1st, 2nd, 4th, 5th, 6th, 7th and 8th execution. This has been verified by simulation results in Fig. 7 and 8 of both optimal X_i values of S^3 and minimum SSE . Further, the improved AGM based on saddle distribution method is suffered to optimise the search space region and to characterise the homogeneous response of higher order polynomial coefficients as the AGM method is generally applicable for confined search space region in control parameters optimisation.

VI. CONCLUSION

The proposed predetermined time constant (T_{Sp}) method enhanced the optimization of search space boundaries for global optima convergence. The response's dynamic period and settling time provide better presumption of an initial T_{Sp} for search space optimisation. The extended SB_{Upper} and SB_{Lower} for an optimal search boundary (SB_O) derived from an initial T_{Sp} brought the elite group within a feasible bounded search region. Further, SGAs execution improved the exploration of elite groups to locate and exploit the optimal values for the identified model parameters. As expected, the polynomial coefficients (for S^l , S^2 and S^3) of two processes are optimised well by SGAs.

References

- [1] R.K. Ursem, "Models for Evolutionary Algorithms and Their Applications in System Identification and Control Optimisation", PhD Thesis, University of Aarhus, Denmark, 2003.
- [2] J.H. Holland, "Adaptation in Nature and Artificial System", Ann Arbor: The University of Michigan Press, 1975.
- [3] B.F. Zhu, Z.M. Li and B.C. Zhang, "Structural Optimal Design: Theory and Applications", Hydro-Electrical Press, Beijing, China, 1984.
- [4] J.E.A. Heris and M.A. Oskoei, "Modified Genetic Algorithm for solving N-Queens Problem", Iranian Conf. Intel. Sys. (ICIS), pp. 1-5, 2014.
- [5] Z.Y. Wu and A.R. Simpson, "A Self-Adaptive Boundary Search Genetic Algorithm and its Application to Water Distribution Systems", Journal of Hydraulic Research, Vol. 40, Issue (2), pp. 191-203, 2002.
- [6] Z.Y. Wu and Y.T. Wang, "Arch Dam Optimisation Design Under Strength Fuzziness and Fuzzy Safety Measure", Proc. Of Int. Conf. on Arch Dam, Hehai University, Nanjing, China, pp. 129-131, 1992.
- [7] Y. Lin, J.M. Hao, Z.S. Ji and Y.S. Dai, "A Study of Genetic Algorithm based on Isolation Niche Technique", Journal of Systems Eng., Vol. 15, pp. 86-91, 2000.
- [8] Q.Y. Tu and Y.D. Mei, "Comparison of Genetic Simulated Annealing Algorithm and Niche Genetic Algorithm for reservoir Optimal Operation", Hydropower Automation and Dam Monitoring, Vol. 32, pp. 1-4, 2008.
- [9] J.L. Jin, X.H. Yang and J. Ding, "An Improved Simple Genetic Algorithm-Accelerating Genetic Algorithm", Systems Eng. Theory & Practice, pp. 8-13, 2001.
- [10] K.S. Mahendra, R. Gupta and P.R. Bhawe, "Optimal Design of Water Networks using Genetic Algorithm with Reduction in Search Space", Journal Water Resour. Plann. Mngmnt., ASCE, Vol. 134, Issue (2), pp. 147-160, 2008.
- [11] K. Vairavamoorthy and M. Ali, "Pipe Index Vector: A Method to Improve Genetic Algorithm-Based Pipe Optimisation", Journal Hydraulic Engg, ASCE, Vol. 131, Issue (12), pp. 1117-1125, 2005.
- [12] B. Xu, P. Zhong and L. Tang, "Improvement on Boundary Searching of Accelerating Genetic Algorithm", Inter. Conf. on Intelligent Design and Engineering Application", pp.301-305, 2012.

Reconfigurable software architecture for a hybrid micro machine tool

Wenbin Zhong, Wenlong Chang, Luis Rubio, Xichun Luo*
Department of Design, Manufacture and Engineering Management
University of Strathclyde
Glasgow, G1 1XJ, UK
*xichun.luo@strath.ac.uk

Abstract—Hybrid micro machine tools are increasingly in demand for manufacturing microproducts made of hard-to-machine materials, such as ceramic air bearing, bio-implants and power electronics substrates etc. These machines can realize hybrid machining processes which combine one or two non-conventional machining techniques such as EDM, ECM, laser machining, etc. and conventional machining techniques such as turning, grinding, milling on one machine bed. Hybrid machine tool developers tend to mix and match components from multiple vendors for the best value and performance. The system integrity is usually at the second priority at the initial design phase, which generally leads to very complex and inflexible system. This paper proposes a reconfigurable control software architecture for a hybrid micro machine tool, which combines laser-assisted machining and 5-axis micro-milling as well as incorporating a material handling system and advanced on-machine sensors. The architecture uses finite state machine (FSM) for hardware control and data flow. FSM simplifies the system integration and allows a flexible architecture that can be easily ported to similar applications. Furthermore, component-based technology is employed to encapsulate changes for different modules to realize “plug-and-play”. The benefits of using the software architecture include reduced lead time and lower cost of development.

Keywords—hybrid micro machine tool; reconfigurable software architecture; finite state machine; component-based technology

I. INTRODUCTION

Conventional machining techniques have been increasingly challenged by some highly engineered mechanical products such as gas turbines, advanced automotive systems and heavy off-road equipment, which rely on high strength materials [1]. Although the recent improvements on tool materials have improved the machinability on advanced materials such as aeroengine alloys, structural ceramics and hardened steel to some extent [2,3], the productivity and manufacturing cost remain problems to be solved. Nevertheless, hybrid machining processes have the unique advantage of being able to machine the difficult-to-cut materials with high material remove rate, while achieving fine surface finish and reduced tool wear. Figure 1 shows the machined surface topography of hardened steel (53HRC) by ultrasonic-assisted diamond machining and the diamond tool tip after the machining. It can be seen that good surface quality (R_a of 4.38 nm) and negligible tool wear

were achieved. There are a number of hybrid machining processes having been developed, which can be classified as assisted and combined hybrid machining processes [5].

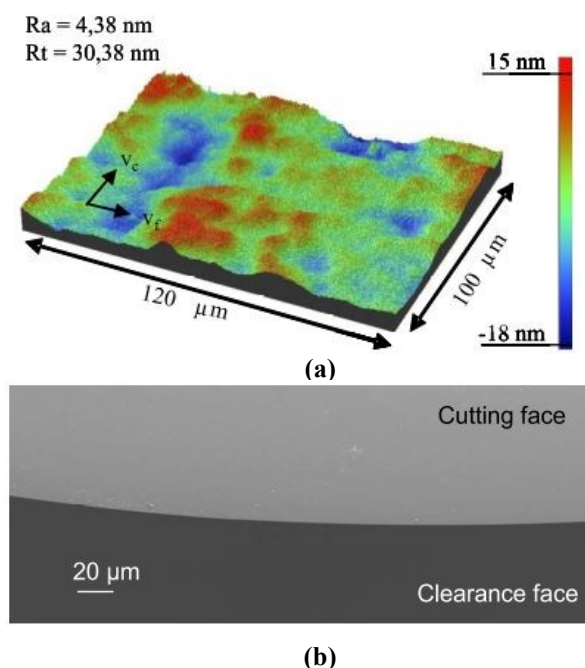


Figure 1. Ultrasonic-assisted diamond machining on hardened steel: (a) Machined surface topography; (b) SEM image of the monocrystalline diamond tool after machining [4]

However, there are a number of issues in the development of hybrid machines that are fundamentally different from the development of conventional machines. The reason lies in that hybrid machines need to incorporate more components from different vendors to achieve the desired functionalities. A typical configuration of a hybrid machine is that a multi-axis machining control system works with a non-conventional machining controller as well as many advanced on-machine sensors. Therefore, a reconfigurable software architecture that enables hybrid machines to react to changes rapidly is highly desirable. Unfortunately, there are no generic approaches for the design of software for such highly customized hybrid machines presently, which is largely due to the lack of standards in sensor interfaces, signal processing algorithms and control systems. Great efforts are required to maintain and upgrade the machines because of the low reconfigurability of the software architecture. In recent years, open architecture controllers,

which are PC-based solutions with a homogenous and standardized environment, are widely witnessed among industry. They provide the possibility of continuously integrating new advanced functionality into the control system from hardware side. Figure 2 summarizes the trend of development of open controller architecture and its components.

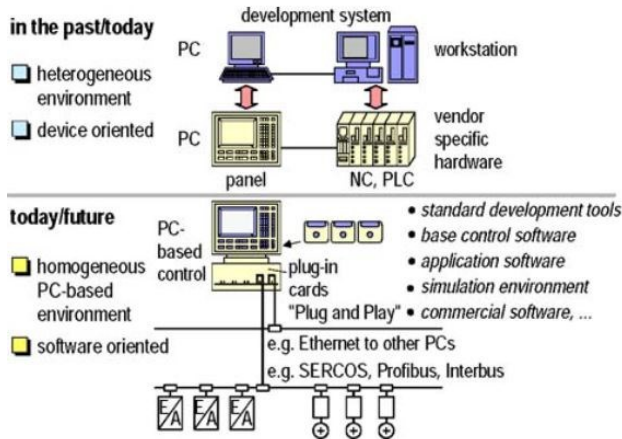


Figure 2. PC-based, software-oriented Control Systems [6]

Nor and Cheng successfully developed a PC-based control system for a five-axis ultra-precision micro-milling machine — ‘Ultra-Mill’ [7]. Some third party sensors were connected to the system based on open controller architecture. However, the software architecture was based on the CNC controller used in the application. The software integrated development environment (IDE) was also provided by the CNC vendor. Such device-oriented architecture with proprietary hardware and software components decreases system integrity as well as reconfigurability.

In this paper, the requirement for the reconfigurability of control software architecture for hybrid micro machine tools is investigated. The focus will be on the design of software architecture for a hybrid micro machine tool with laser-assisted 5-axis micro-milling capability, integrated with a material handling system, on-machine metrology and on-line force measurement. Two key technologies used in the implementation of the architecture are discussed. Finally three reconfigure scenarios are provided to illustrate characteristics of the software architecture and its advantages over device-oriented architectures.

II. SYSTEM DESCRIPTION

Ultra-precision micro-milling machine tools are widely used to manufacture 3D complex micro-components at sub-micron accuracy and nanometer surface finish. This project aims to develop a hybrid micro machine tool to push the limits of ultra-precision micro-milling, viz. enhancing the machining performance on hard materials and increasing the productivity. The machine consists of a 5-axis micro-milling system, a laser-assisted machining system, a material handling system, and two sensor systems, including an interferometer for on-machine metrology and a dynamometer for on-line cutting force measurement.

Each sub-system is developed or provided by a different vendor with its own hardware and software and can function independently. However, as a whole, they should be coordinated to collaborate seamlessly to achieve the goal — hybrid machining. In this application, the open controller architecture is adopted. A PC acts as the coordinator, while all the systems can exchange data with the PC through communication interface or peripherals, as shown in Figure 3. A software system running on Windows operating system is required to manage and process the data from each sub-system as well as providing the human-machine interface (HMI).

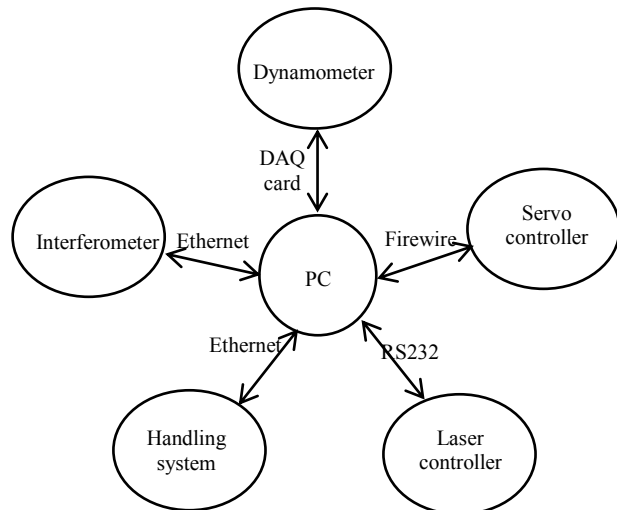


Figure 3. System connection diagram

A reconfigurable software architecture allows the modification of functionality in a very dynamic way. There are five characteristics of reconfigurability, which can be summarized below [8]:

- **Modular.** Decentralized structures are supported. All components are encapsulated as vendor-neutral modules.
- **Interoperable.** Components can cooperate in a consistent manner and can interchange data in a defined way.
- **Portable.** Components can be integrated in different environments without any changes, while maintaining their capabilities.
- **Scalable.** Topology of the architecture can be modified depending on the user requirement.
- **Extendable.** Functionality can be enhanced. A variety of components can run on the platform without any conflicts.

Modular and interoperable are the critical characteristics of reconfigurability. They ensure the rest three characteristics. To fulfill the criteria of being reconfigurable, the following principles should be taken into consideration:

- **Decoupling.** Too much coupling results in an inflexible architecture.

- Vendor-neutral. This feature ensures that a variety of components can be incorporated with minimum effort.
- Standard-based. The data exchange and application interface (API) should be standardized, so that a component can be distributed rapidly.

III. SOFTWARE ARCHITECTURE DEVELOPMENT

From the control perspective, a typical machine tool has three control layers, the hierarchy comprises servo control, process control and supervisory control from bottom to top. These control layers interact with machine, process and product respectively, and lead to three control loops. Each control loop is running upon a group of hardware, and controls some parameters within this layer, while a bottom layer will get instructions from its top layer. The control scheme can help look into the design of the software architecture with the reconfigurable design principles described in section 2.

Firstly, the scheme covers all the machine functions with very clear data flow. For easy cooperation of different components, the data, which contains control messages and sensor feedback, can be packaged as events. All the possible events are defined in a protocol that all components on the platform should follow. As a result, the software architecture becomes event-driven; Secondly, all hardware that services for the machine can be placed in one of the layer. After encapsulation, the function of the hardware can be a vendor-neutral component in the architecture; Finally, the coupling is remarkably reduced with the event-driven mechanism and vendor-neutral components. Therefore, in this software architecture, each sub-system, for example servo controller, can be regarded as one unified module with specific events in/out interfaces, and can be encapsulated into a component. During the process, FSM is used in implementing the event-driven mechanism, while component-based technology is used for modularization.

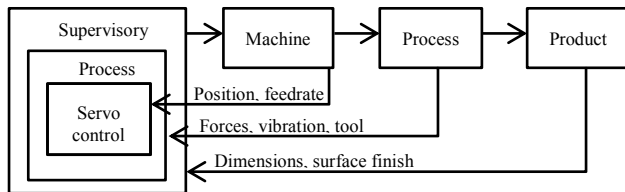


Figure 4. Machine, process and product interactions in the control loops

A. FSM Model

FSM can be observed widely in software industry but is rare in the machine tool software design. It can change from one state to another when specific event is triggered. In this application, FSM is implemented in both component level and architecture level. Figure 5 illustrates the work principle of a FSM. When an event is received, FSM will call the API defined in the knowledge base to process it, with the state change following. The FSM itself may trigger events, which will be dispatched to other modules. The type of events to be received and the type of events can be posted are agreed using the protocol. An

unknown event will cause an exception. The knowledge base along with the FSM is of the greatest importance. Typically, there are two types of APIs within the base. The first type is hardware abstract APIs, which are directly related with hardware and ensure the proper operation of the hardware; Another type is advanced algorithms for data processing, for example, the workpiece surface data from the metrology sensor is processed to analysis the surface quality, the results can be used to optimize the machining parameters.

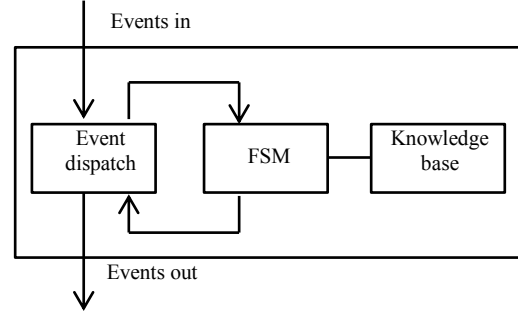


Figure 5. FSM work principle

Each FSM model for a sub-system comes with standardized data exchange, which is defined as events. The events implementation varies with the programming language to be used, a structure can be used when programming with C/C++, the structure may have the following definition:

```
typedef struct {
    DWORD dest;
    UINT message;
    WPARAM wParam;
    LPARAM lParam;
    DWORD time;
} EVENT;
```

In this example, `dest` identifies the destination of the event, `message` specifies the ID of the event, `wParam` and `lParam` specify some additional information of the event, `time` indicates the time when this event was posted. The available event ID and exact meaning of additional information are explained by the protocol. A new task is created to let one FSM work, which leads to parallel software tasks in the architecture, and makes the model more independent.

B. Modular

From the software perspective, a component is a reusable piece of software that serves as a building block within an application. It is also language-neutral that can be used in a variety of languages, which makes it easier to integrate. Essentially, the software is built with a set of components, with events flowing from one component to another. Besides the sub-systems described in Section 2, HMI is essential for the architecture. HMI, as its name suggests, is responsible for the interaction with operators. It collects command inputs from operators and dispatches them to the corresponding sub-systems, as well as receiving update events to update the display. All the sub-systems are based on a FSM model presented before, and

they are encapsulated into several components or modules as shown in Figure 6. There are several component based

technologies, of which Windows COM technology is used in this application.

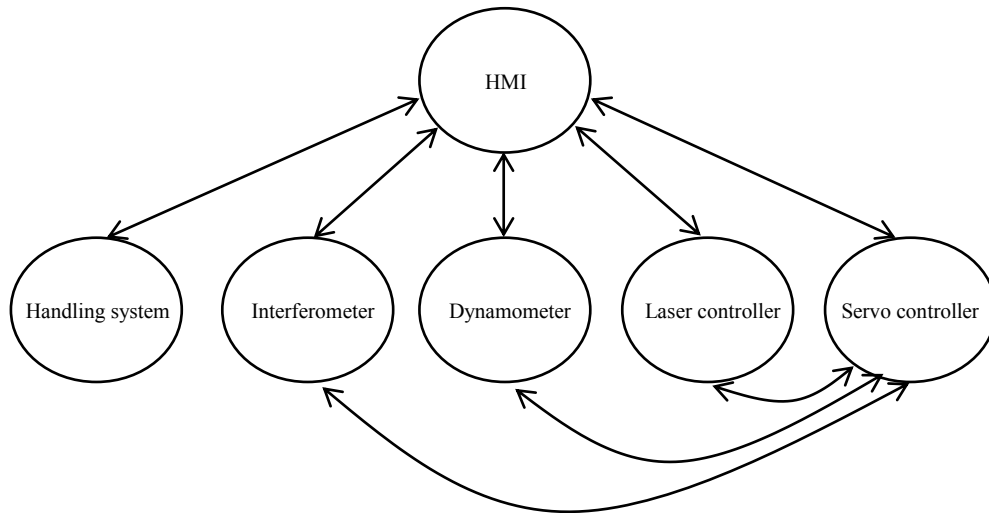


Figure 6. Modules and events flow in the architecture

IV. CASE STUDY

The case study includes the software reconfiguration in three scenarios: change a component, add a component and add a feature.

A. Change a Component

This case considers when a new hardware should replace the old one for better performance, for example, a new type of laser is to be installed. In this case, a new component interface is created by implementing the same component interface, states and state transitions, then it can replace the old one without causing any problem, the only change needed to be made is that adapting the hardware abstraction API to the new controller. The good scalability is achieved.

B. Add a Component

This case looks into the condition that if a new module should be added. For example an acoustic emission sensor is needed to detect tool breakage. In this case, the protocol defining the events should be extended, and the related FSM model should be expanded to accommodate the new events. The extendibility is improved due to the architecture.

C. Add a Feature

Another common case is that a new algorithm is developed to improve the machine performance when the machine has already been delivered. For this scenario, the quick solution is that adding the algorithm in the related component and adding an event at which the algorithm will be called.

V. CONCLUSION

A software architecture has been developed for a hybrid micro machine tool. It uses FSM technology for hardware control and data flow. The coupling between different modules is minimized. The data exchange between different components is standardized to events.

The architecture encapsulates modules using component-based technology, which improves the reconfigurability of the software system. Results from this paper provide a general methodology for designing the software architecture for hybrid machine tools, especially these incorporating several subsystems from different suppliers.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support of EPSRC (EP/K018345/1).

REFERENCES

- [1] B. Lauwers, F. Klocke, A. Klink, a. E. Tekkaya, R. Neugebauer, and D. McIntosh, "Hybrid processes in manufacturing," *CIRP Ann. - Manuf. Technol.*, vol. 63, no. 2, pp. 561–583, 2014.
- [2] E. O. Ezugwu, "Key improvements in the machining of difficult-to-cut aerospace superalloys," *Int. J. Mach. Tools Manuf.*, vol. 45, no. 12–13, pp. 1353–1367, 2005.
- [3] E. O. Ezugwu, J. Bonney, and Y. Yamane, "An overview of the machinability of aeroengine alloys," *J. Mater. Process. Technol.*, vol. 134, no. 2, pp. 233–253, 2003.
- [4] B. Bulla, F. Klocke, O. Dambon, and M. Hüntel, "Ultrasonic Assisted Diamond Turning of Hardened Steel for Mould Manufacturing," *Key Eng. Mater.*, vol. 516, pp. 437–442, 2012.
- [5] S. Z. Chavoshi and X. Luo, "Hybrid Micro-machining Processes: A Review," *Precis. Eng.*, vol. 41, pp. 1–23, 2015.
- [6] G. Pritschow, Y. Altintas, F. Jovane, Y. Koren, M. Mitsuishi, S. Takata, H. van Brussel, M. Weck, and K. Yamazaki, "Open controller architecture—past, present and future," *CIRP Ann. - Manuf. Technol.*, vol. 50, no. 2, pp. 463–470, 2001.
- [7] M. K. M. Nor and K. Cheng, "Development of a PC-based control system for a five-axis ultraprecision micromilling machine 'Ultra-Mill' and its performance assessment," *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.*, vol. 224, pp. 1631–1644, 2010.
- [8] Y. Koren, U. Heisel, F. Jovane, T. Moriwaki, G. Pritschow, G. Ulsoy, and H. Van Brussel, "Reconfigurable Manufacturing Systems," *CIRP Ann. - Manuf. Technol.*, vol. 48, no. 2, pp. 527–540, 1999.

Measurement Station Planning of Single Laser Tracker based on PSO

Zhenying Xu¹, Baozhong Wu¹, Meng Zhang¹
College of Mechatronic and Automation
National University of Defense Technology
Changsha, Hunan, P. R. China
zhenying_xu@163.com

China National South Aviation Industry Co., Ltd,
Changsha, Hunan, P. R. China

Yuehui Yan
Beijing Institute of Space Long March Vehicle
Beijing, P. R. China

Feng Zou

Abstract—The laser tracker owns large measuring range, high precision and is easy to carry, which makes it widely used in large scale measurement. For a specific measurement task, however, it is hard to locate a laser tracker measurement station. In view of this situation, this paper advances a new planning method, which combines optimization engine with computer virtual laser tracker measurement (VLTM). A design case constructed by particle swarm optimization (PSO), SpatialAnalyzer(SA) virtual measurement and a high-precision marble plane is studied to illustrate and demonstrate the feasibility of the proposed method. The result shows that this method is able to find out the feasible solution which meets the uncertainty requirements in the given area automatically and exactly.

Keywords—laser tracker; Planning of Measurement field; Flatness Measurement; SpatialAnalyzer virtual measurement; PSO

I. INTRODUCTION

Laser tracker system is high-precision and its measuring range is large. It is widely used in industrial measurement and fits to large piece assembly measuring, heavy machinery manufacturing, large parts detection and reverse engineering. Because of the uncertainty, the measuring result cannot determine the product is qualified or not directly. Therefore, a good measurement result contains measurement values, confidence level and confidence interval in modern error theory [1].

The planning of laser tracker measurement field is about the layout of measurement stations, common points and enhanced reference points. The confidence level and confidence interval are related to measurement field layout. Up to the present, the study on planning of measurement field get some guiding principles only; in large piece manufacturing, making a coordinate measurement plan is always depends on experience and experiments, which is inefficient. Unfortunately, sometimes, for example, with sophisticated measurement environment or demanding accuracy requirements, it is impossible to get a feasible solution. The traditional methods and computer technology are not close, which is detrimental to optimization and improvement.

This paper focuses on the study of automatic planning techniques of locating a single laser tracker measurement station. For planning the layout of laser trackers automatically, the optimization combines with virtual laser tracker measurement (VLTM), and sets the preselected area of space as constraints and measurement uncertainty requirements as target. In this paper, the measurement of high-precision marble surface flatness is as an example for verification. The result shows that the planning method is high efficiency and accurate-planning, and can plan the measurement field layout automatically.

II. PLANNING METHOD OVERVIEW

The planning of locating a single laser tracker measurement station is contained in laser tracker measurement field planning. While doing coordinate measurement with laser trackers, if all the targets are under the vision of a single laser tracker without any obscured points, one measurement station will suffice. In this case, to meet the measurement accuracy is the only constraint, and at the same time, the difficult to study reduced and significance clear.

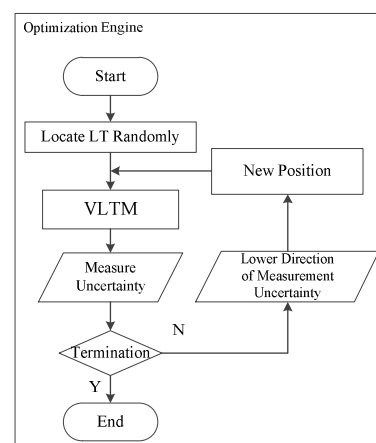


Figure 1. Design of optimization engine

The implementation approach of locating a single laser tracker measurement station is: taking the optimization engine as main part, and the uncertainty of VLTM as the input of

optimization engine; and then, calculating the lower direction of measurement uncertainty for re-planning the new VTLM station. As a result, the best position and its uncertainty are output. Fig.1 shows the implementation approach of the optimization engine.

In this paper, the optimization engine is PSO, and SA provides VTLM, and the goal is to get the best station or that meets the uncertainty requirements.

III. COMPUTER VIRTUAL MEASUREMENT

A. SA Simulation Introduction

Without linking to a real instrument, SA gains the measurement data by simulating the behavior of laser tracker with certified algorithms. The measurement errors follow a normal distribution, and setting up the instrument precision parameters and environmental parameters can make it roughly the same as the real one's, which makes the results of simulation believable.

Document [1] achieved the target of doing the large-scale measurements automatically with SA, VCMM method and MonteCarlo method, which is used to simulate the sources of uncertainty in the measurement process. Because of the complicating factors in measurement process and the difficulty to establish precise measurement model, the MonteCarlo method is the most effective way to evaluate the laser tracking measurement uncertainty in object oriented measuring.

In this paper, MonteCarlo method combined with SA simulation is used to evaluate the uncertainty.

B. Simulation Design

The simulation input is from a SA file, which contains the geometric model of measured object. By clicking the surface, a series of points are generated for output.

The simulation process is as follows:

- 1) After the program starts, select a point data file via dialog box, and read it with prescribed format and save it into memory.
- 2) Through the measure plan (MP) interface, the program links to SA, and it can manipulate SA automatically. Close auto event creation.
- 3) Add an API Tracker III into an empty SA file and set up the instrument precision parameters and environmental parameters to fit in with the real one.
- 4) Locate the instrument with the data from PSO, then loop simulation.
- 5) Read points data from memory and construct points in SA through functions. Make fabricate observations on these points with instrument error.
- 6) Fit plane to the measured points by using SA functions, and get the flatness at the same time, then save it into memory.

- 7) Delete all data except the instrument to improve running efficiency.
- 8) Loop 5)-7) until uncertainty converges
- 9) Read simulation results from memory and compute uncertainty for PSO

The process is as Fig.2:

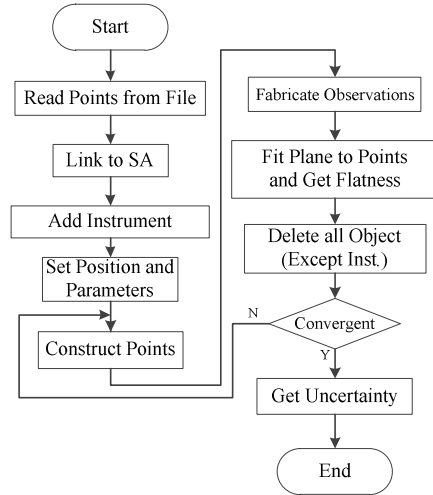


Figure 2. SA simulation process

IV. PSO AND ITS IMPROVEMENT

A. Basic PSO

Particle Swarm Optimization (PSO) is a kind of global optimization method based on swarm intelligence theory. Its swarm intelligence is generated by the cooperation and competition among the particles, which guides the optimal search [2]. The particles update speeds and positions according to their own best positions $pbest$ and the global best position $gbest$.

$$v_{ij}^{k+1} = \omega v_{ij}^k + c_1 r_1 (pbest_{ij}^k - x_{ij}^k) + c_2 r_2 (gbest_j^k - x_{ij}^k) \quad (1)$$

$$x_{ij}^{k+1} = x_{ij}^k + \alpha v_{ij}^{k+1} \quad (2)$$

In the formulas, subscript i is particle number, subscript j is dimension number, superscript k is iterative algebra; the particles' position and speed are in a given range; c_1 and c_2 are non-negative learning factors, r_1 and r_2 are random number in $[0, 1]$, ω is inertia factor and α is constraint factor; $pbest_{ij}^k$ is the j^{th} dimension of particle P_i 's best position; $gbest_j^k$ is the j^{th} position dimension of the global best one. Equation (1) and (2) make up the basic PSO.

B. Improve PSO with SA Simulation

PSO combined with SA simulation is different from the basic one. For the former one, its range of activities is discrete, and the fitness is from SA simulation.

It always takes a long time to simulate at a position. And the influence of the error of locating the laser tracker measurement station to measurement accuracy is not obvious in a small range. Considering of the previously mentioned, introduce discrete layout can narrow the range of activities and avoid doing simulation at the same place. It can improve simulation efficiency. For discretization, upper and lower bounds(X_{up} , X_{down}), the degree of dispersion $Step$ and the next position X' provided by PSO are needed. Then calculate the spacing between one point and another in all directions by (3).

$$Xstep = (X_{up} - X_{down}) / Step \quad (3)$$

After getting the spacing, calculate new position X by (4) or (5).

$$X = \lfloor X' / Xstep \rfloor \times Xstep \quad (4)$$

$$X = \lceil X' / Xstep \rceil \times Xstep \quad (5)$$

The discretization process is as Fig.3:

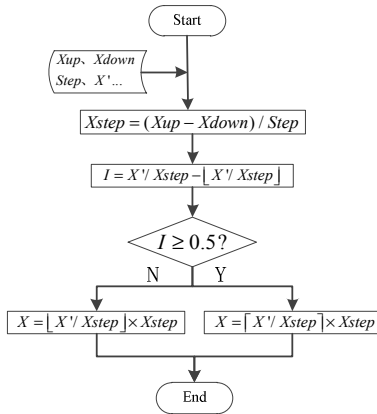


Figure 3. Discretization process

Do simulation after locating the new position and record it.

The fitness is SA simulation measurement uncertainty. The uncertainty is smaller, that is closer to the optimal solution. The objective function of simulation is

$$\begin{cases} \min U = Simu(InstID, P(x_i, y_i), U(r, \theta, \phi)), \\ (x_i, y_i) \in [X, Y], \\ -25000 \leq x_i \leq 25000, \\ -25000 \leq y_i \leq 25000. \end{cases} \quad (6)$$

$\min U$ represents the target is the smallest uncertainty, $Simu()$ is the simulation process, $InstID$ is the ID of the instrument, $P(x_i, y_i)$ is the instrument's position which is included in the discrete plane $[X, Y]$, its unit is mm, $U(r, \theta, \phi)$ is the uncertainty of instrument.

The algorithm implementation process is as Fig.4, proceed as follows:

- 1) Particle swarm initialization: Randomly initialize each particle's position and velocity, set of particles' size n
- 2) Calculate the fitness: first of all, search record to tell whether the current position has been reached before, if it has, read the fitness from the record; otherwise, do simulation there and record the current position and its fitness.
- 3) For the first generation of particles, $pbest$ is current position. And find out the $gbest$ from them. For subsequent generations of particles, if the current position is better than $pbest$, then update the $pbest$ as the current one. For each particle in the entire population, compare its $pbest$ with $gbest$, if it is better than $gbest$, then update the $gbest$.
- 4) Update speed and position according to (1) and (2).
- 5) Check the termination condition: determine whether the algorithm achieves fitness requirements or the best fitness value. If it does, terminate the program, otherwise go to 3).

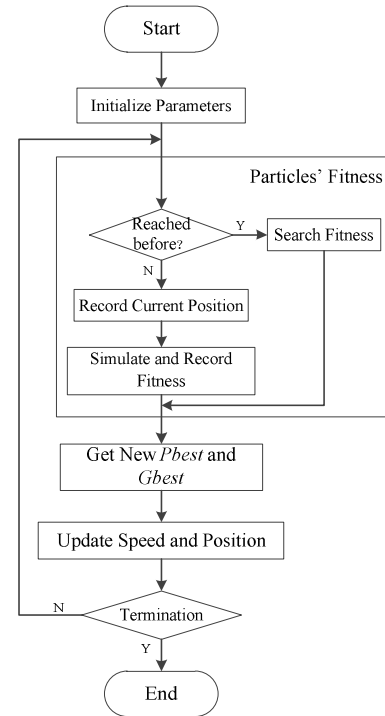


Figure 4. Algorithm process

V. FLATNESS MEASUREMENT SIMULATION AND EXPERIMENTAL RESULT DISCUSSION

A. Definition of flatness

The mathematical definition of flatness [3] is: two parallel planes with spacing equal to the tolerance t of the defined area. It can be expressed as a set of points satisfying (7):

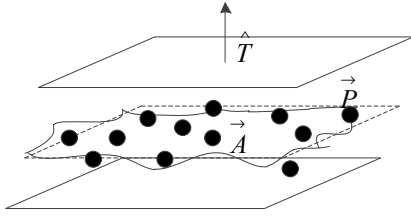


Figure 5. Mathematical definition of flatness

$$|\hat{T} \cdot (\vec{P} - \vec{A})| \leq \frac{t}{2}. \quad (7)$$

In (7), \hat{T} is the normal vector of two parallel planes which defines the tolerance zone, \vec{A} is the position vector of the median plane between two parallel planes.

To evaluate the flatness, there are several methods such as minimum zone method least squares method. In this paper, SA provides the flatness, and its algorithm is certified least squares methods.

B. Simulation design

The MonteCarlo method is used to simulate the sources of uncertainty in the measurement process. It requires multiple measurement process through computer simulation. However, it's hard to know how many times after the simulation results can be convergent. For using the SA to get the uncertainty, document [4] verified the credibility of 10,000 times simulation results, and drew figures to analyze the convergence of uncertainty. Fig.6 shows the uncertainty change trend of D_x , D_y , D_z , R_x , R_y , R_z , in a specific measurement plan. It shows that there is a convergent uncertainty after 3000 times simulation.

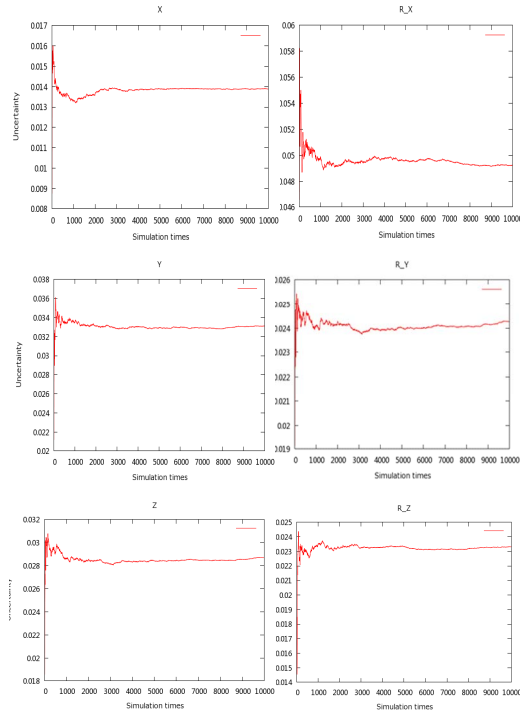


Figure 6. Trend of uncertainty of 10000 times simulation

While planning measurement field by PSO, it is unavoidable to measure the same object repeatedly. So, it requires greater simulation efficiency. Based on the literature's conclusion, the simulation time is set to 3000. It can not only guarantee the accuracy of the simulation, but also needs the minimum resource.

The basic parameters of the algorithm are set as follows:

- a) Particle size n : Generally set to 20~40, and $n=20$ in this paper;
- b) Particle dimension D : determined by the number of the objective optimization function's arguments, and $D=2$ in this paper;
- c) Range of activities: determined by the distribution of measuring points; in this paper, the measured object is set to the origin, and the range is $\pm 25m$, and there are 1000 points in each direction.
- d) Maximum speed V_{max} : due to the large measuring range, set $V_{max}=0.1\Delta$, $\Delta=X_{up}-X_{down}=50m$.
- e) Termination condition: g_{best} meets the fitness requirements or g_{best} is unchanged in the latest 200 times.

After analyze the data from experiment, the uncertainty of the instrument [5] is

$$\begin{aligned} u_{\theta} &= 0.483 \\ u_{\phi} &= 0.246 \\ u_r &= -0.0013 \\ u_{ppm} &= 1.18. \end{aligned} \quad (8)$$

In a simulation, the positions experienced by the laser tracker are shown in Fig.7:

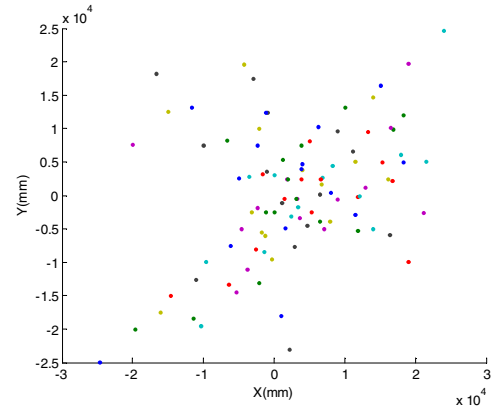


Figure 7. Experienced positions

As is shown in Fig.7, there is high efficiency by using PSO to find the right position, and it experienced less than 1% among the whole points.

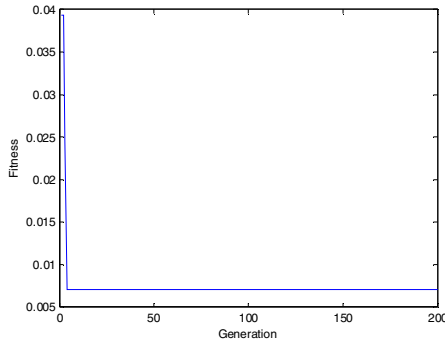


Figure 8. Particle fitness

As is shown in Fig.8, the particles evolve fast, and get the optimal solution in several generations. The optimal position is (1400, 500).

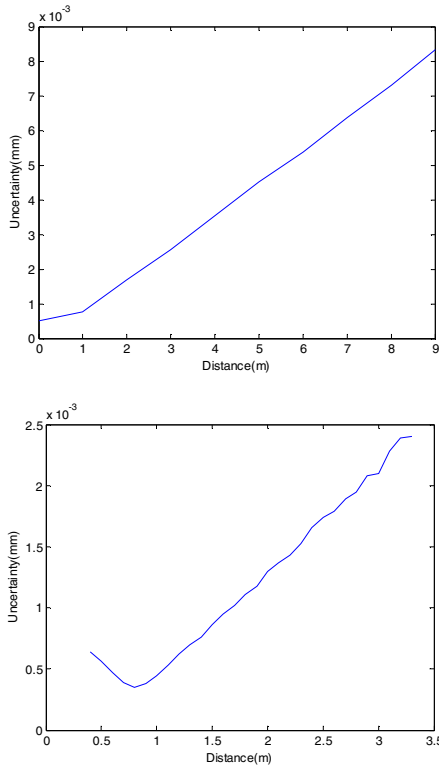


Figure 9. Uncertainty along X-axis

Fig.9 up is in the range of 0-9m, and the step equals 1m, Fig.9 down is in the range of 0-3m, the step is 0.1m. As is shown in Fig.9 that the optimal position is distributed at 1.4m in this interval, which is basically the same as the PSO's conclusion. The result of the algorithm is credible.

C. Experimental design and processing

A high-precision marble flat is used in this experiment, and its flatness equals $5\mu\text{m}$. Fundamentally, laser tracker station layout planning is to get the relative position of the plane and the laser tracker. Because moving the plane is more convenient, in this experiment, the laser tracker is stationary and the plane moves, which equals to moving the laser tracker.



Figure 10. Experimental environment

The relative position of the plane and laser tracker in the experiment is shown as Fig.11, and each position is repeated measurements 10 times.



Figure 11. Relative positions of the plane and laser tracker

The relative position of the plane and laser tracker in the experiment is shown as Fig.11, and each position is repeated measurements 10 times.

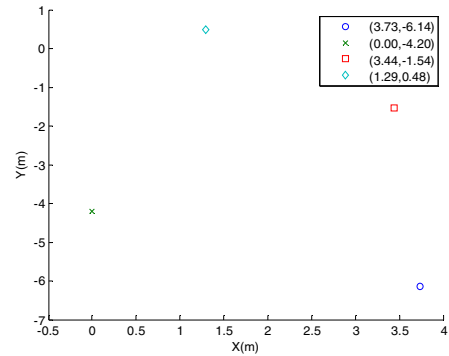


Figure 12. Positions after converted

Fit the 10 sets of data measured at (1.29, 0.48) to plane by SA, and display the flatness in table 1:

TABLE I. FLATNESS AT (1.29, 0.48)

No.	1	2	3	4	5
Flatness	0.0101	0.0058	0.0072	0.0064	0.0061
No.	6	7	8	9	10
Flatness	0.0061	0.0053	0.0104	0.0057	0.0048

Because the 1st, 8th data are larger than others apparently, consider the possibility of gross error. Use the Grubbs discriminate for testing two outliers [6] to determine.

Arrange the above data in order by size to order statistic.

$$\begin{aligned}
x_{(1)}, x_{(2)} &\leq x_{(3)} \leq \dots \leq x_{(n)} \\
s_0^2 &= \sum_{i=1}^n (x_{(i)} - \bar{x})^2 = 3.3488\text{e-}04 \\
\bar{x}_{9,10} &= \frac{1}{n-2} \sum_{i=1}^{n-2} x_{(i)} = 0.0060 \\
s_{9,10}^2 &= \sum_{i=1}^{n-2} (x_{(i)} - \bar{x}_{9,10})^2 = 2.8759\text{e-}05 \\
\bar{x}_{1,2} &= \frac{1}{n-2} \sum_{i=3}^n x_{(i)} = 0.0073 \\
s_{1,2}^2 &= \sum_{i=3}^n (x_{(i)} - \bar{x}_{1,2})^2 = 2.0694\text{e-}04
\end{aligned}$$

Computing high and low Grubbs test statistics

$$\begin{aligned}
g_b &= \frac{s_{n-1,n}^2}{s_0^2} = \frac{s_{9,10}^2}{s_0^2} = 0.0859 \\
g_b' &= \frac{s_{1,2}^2}{s_0^2} = 0.6180.
\end{aligned}$$

Look-up table, find out the value corresponding to $n=10$, $\alpha=0.05$, the critical value $G_b(0.05,10)=0.1864$.

Because of

$$g_b < g_b'$$

and

$$g_b < G_b(0.05,10),$$

the maximum $x_{(10)}=0.0104$ and $x_{(9)}=0.0102$ may have gross error.

Redetermination the data after excluding the two maxima

$$\begin{aligned}
g_b &= \frac{s_{n-1,n}^2}{s_0^2} = \frac{s_{7,8}^2}{s_0^2} = 0.2621 \\
g_b' &= \frac{s_{1,2}^2}{s_0^2} = 0.3088.
\end{aligned}$$

Because of

$$g_b < g_b'$$

and

$$g_b > G_b(0.05,8) = 0.1101,$$

there is no gross error in the set. The variance of new data is 0.007mm. Other data are also processed by the way for a single outlier or more, the results are shown in table 2.

TABLE II. FLATNESS UNCERTAINTY AT VARIOUS LOCATIONS

Position (m)	Distance (m)	Uncertainty (measured)	Uncertainty (simulated)
(1.29, 0.48)	1.4	0.0007	0.0008
(3.44, -1.54)	3.7	0.0037	0.0035
(0.00, -4.20)	4.2	0.0012	0.0035
(3.73, -6.14)	7.1	0.0069	0.0064
(3.73, -6.14)	7.1	0.0069	0.0064

As is shown in the table, by comparing the measured data and the simulation one, the uncertainty of measurement and simulation are substantially similar. And at (1400, -500) which provided by PSO, the uncertainty is smaller than other locations significantly. Thus, the conclusion is credible.

VI. CONCLUSION

In this paper, the stations layout planning for laser tracker measurement field is conducted according to the measurement uncertainty requirements of flatness. SA simulation and location query are introduced to PSO, and gets good results. The improved PSO is computationally efficient, and the error distribution of SA simulation is close to that of real laser tracker. In fact, the method proposed in this paper is not only limited to the PSO, SA, and flatness measurement. It applies to different optimization and virtual laser tracker measurement methods. For other measured objects exposed to a single laser tracker station, it applies, too. It greatly increases the versatility of the method. This paper have not considered some other problems such as the convergence of uncertainty, the blocked light path and the tasks which needs two stations at least. These will be the future research directions of this subject.

- [1] Yang Jingzhao. Research on key technologies in evaluation and planning of laser-tracker measurement for large scale manufacturing[D]. Changsha: Graduate School of National University of Defense Technology, 2014.
- [2] Eberhart R C, Kennedy J. A new optimizer using particle swarm theory[C] //Proc. Of the 6th Int'l Symp. on Micro Machine and Human Science. Nagoya, Japan: 1995: 39-43.
- [3] ASME Y14.5.1M-1994: Mathematical Definition of Dimensioning and Tolerancing Principles[S].
- [4] He Yucheng. Research on integrated platform of laser track measuring based on MBD[D]. Changsha: Graduate School of National University of Defense Technology, 2014
- [5] SA User Manual.
- [6] Wang Zhongyu, Liu Zhimin, Xia Xintao, Zhu Lianqing. Evaluation of measurement error and uncertainty[M]. Beijing: Science Press,2008:98-120.

Grey-box Identification for Photovoltaic Power Systems via Particle-Swarm Algorithm

Naji Al-Messabi, Cindy Goh, Yun Li

School of Engineering, University of Glasgow, Glasgow G12 8 QQ, U.K.,
(n.al-messabi.1@research.gla.ac.uk, Cindy.Goh@glasgow.ac.uk, Yun.Li@glasgow.ac.uk)

Abstract — Amongst renewable generators, photovoltaics (PV) are becoming more popular as the appropriate low cost solution to meet increasing energy demands. However, the integration of renewable energy sources to the electricity grid possesses many challenges. The intermittency of these non-conventional sources often requires accurate forecast, planning and optimal management. Many attempts have been made to tackle these challenges; nonetheless, existing methods fail to accurately capture the underlying characteristics of the system. There exists scope to improve present PV yield forecasting models and methods. This paper explores the use of apriori knowledge of PV systems to build clear box models and identify uncertain parameters via heuristic algorithms. The model is further enhanced by incorporating black box models to account for unmodeled uncertainties in a novel grey-box forecasting and modeling of PV systems.

I. INTRODUCTION

Photovoltaic (PV) energy is now positioned amongst the top three new power generation means installed in Europe and is expected to remain so [1]. Power from PV sources provides a number of benefits over other renewable energy sources (RES). It can be supplied locally to loads, reducing the cost of transmission lines and associated power losses. Furthermore, advances in technology and large scale manufacturing have led to the decline in PV cost at a steady rate [2]. Despite a high capital setup cost, the operation and maintenance costs of PV are almost zero [3].

Nonetheless, like other RES, PV sources pose a number of integration challenges such as the impact on voltage profile [4],[7], impact on operational costs of the grid [5], regulation and load-following requirements [6]. Advance knowledge of the expected yield from PV sources will help tackle these challenges, to allow for proper planning of available generation sources and provide insights into the impact of PVs on the power network. However, the forecasting task requires non-primitive techniques, as power yield from PVs is intermittent in nature. The intermittent and non-linear characteristics of PV data is due to an interplay of various factors such as the variability in sunrise and the amount of sunshine, sudden changes in atmospheric conditions, cloud movements and dust [8]. The PV power data can thus be viewed as consisting of two parts: the deterministic and the stochastic parts. The former represents the mathematical equations of irradiance that depend on location, sun's position, and equations of PV cells, whilst the latter represents the

sudden atmospheric changes such as dust, clouds, and wind blow.

Various mathematical models that capture physics of PVs or clear-box models are possible but are inaccurate or impractical for large systems [9]. However, clear-box models possess various strengths such as that their structures are of physical meaning and usually have fewer parameters to estimate [10].

On the other hand, data-driven or black-box models based on statistics or artificial-intelligence are popular methods as they are simple and easy to use. Dynamic Neural Networks (DNNs) such as the 'Focused Time-Delay Neural Networks' (FTDNN) and the 'Distributed Time-Delay Neural Networks' (DTDNN) [11] have been studied for PV forecasting. These methods can handle nonlinear time-series data that are dynamic in nature. However, black-box models require good data for proper modeling – both quality and quantity. It is also difficult to design due to large number of parameters and lack of a systematic way to arrive at an optimal structure.

This paper will therefore focus on the identification of proper grey box photovoltaic models. Section II provides an overview of related works in clear-box models for solar PV. In Section III, uncertain parameters in the clear-box model are identified and optimised using Particle Swarm Optimisation (PSO). The model is then extended into a grey-box model to account for unknown effects during forecast. Two grey-box models are developed and the results and observations are discussed in Section IV. Section V concludes the paper.

II. SOLAR PV POWER MODELS

A. General overview

There exist various forecasting models proposed for PV systems [9]-[30]. The simplest are naïve or persistence models where the next power value is assumed to be same as the previous step. Such models are usually taken as reference models in forecasting studies [13].

The other approach to forecast PV power is to model solar data using statistical methods. Regression models can be used where power value is expressed as a regression of previous power values, irradiance, and temperature [14]. Statistical approaches adopt classical time-series forecasting methods that assume the data to be stationary. Auto-regressive (AR), AR with exogenous input (ARX), and AR with integrated moving average (ARIMA) [13] are some of the famous

statistical models used in solar PV forecasting. As the parameters in these models usually do not represent a physical phenomenon or quantity, such models are often referred to as 'black-box' models or functional approximates. The artificial neural network (NN) is another example of these models and is gaining popularity in PV forecasting owing to their modularity in handling non-linear models. There are various structures of NN, but they can be categorized into two: static [15] and dynamic [16]. However, artificial intelligence models can suffer from generalization problems. Also, there is no systematic way in arriving at the structure of the model. An alternative to this is the 'clear-box' model based on physical principles. The benefits of these models were outlined in the introduction. These equations are based on the physics of PV modules and are detailed in the preceding section.

B. PV Module Equations

There are different physical models proposed for PV modules - double diode models [17], simplified single diode model (SSDM), and further SSDM are some of them in a descending order of complexity. The higher complexity can provide better accuracy on the expense of increased computational burden which not suitable for real-time online applications.

The best model that gives a good compromise between simplicity and accuracy [18] is the simplified single diode model shown in Fig. 1.

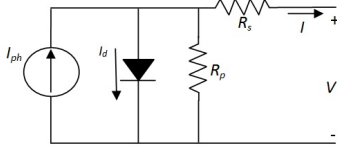


Figure 1. PV cell/array Single Diode Simplified model

The following equations describe the relation between the current and voltage output of the PV cell/array:

$$I = I_{ph} - I_d - \frac{V + R_s I}{R_p} \quad (1)$$

$$I_d = I_o \left(e^{\left(\frac{V + R_s I}{a V_t} \right)} - 1 \right) \quad (2)$$

Where I is the output current of the cell in Amperes, V is the solar cell voltage in Volts, I_{ph} is the photocurrent in Amperes, I_d is the Shockley diode equation, I_o is the reverse saturation or leakage current of the diode, $V_t = kT/q$ is the thermal voltage of the array, q is the electron charge ($1.60217646 \times 10^{-19}$ C), k is the Boltzman constant ($1.3806503 \times 10^{-23}$ J/K), T is the temperature of the cell in Kelvin, and a is the ideality factor constant. More details of these equations can be found in [18]. To calculate power yield, values for I and V are usually computed using numerical methods [17],[18]. The mathematical approach is usually tedious especially when applied to large or widely spread PV systems [9].

C. Simplified PV Equations

Furthermore, the aforementioned equations of PV modules require numerical solution and thus are sometimes replaced with simplified equations that relate the power output with the efficiency of the system and variation in radiation and temperature [19],[20]. These equations are basically a translation of performance measurement from standard test measurements (STC; Air Mass 1.5 spectrum with global irradiance ($G=1000\text{W/m}^2$ and module temperature = 25°C). One famous simple method is that of Osterwarld [19] which can be described as follows:

$$P_m = P_{mo} \cdot \frac{G}{G_o} [1 - \gamma \cdot (T - 25)] \quad (3)$$

Where P_m is the cell/module maximum power (W), P_{mo} is the cell/module maximum power in STC (W), γ is the cell maximum power coefficient ($^\circ\text{C}^{-1}$) which ranges from -0.005 to -0.003 $^\circ\text{C}^{-1}$ in crystalline silicon and can be assumed to be -0.0035 $^\circ\text{C}^{-1}$ with good accuracy.

Another version of equation (7) is given below [20]:

$$P_{MPP} = G_t \cdot A_a \cdot \eta \cdot [1 + K_T(T_c - T_{ao})] \quad (4)$$

$$\eta = \eta_m \eta_{dust} \eta_{mis} \eta_{DCloss} \eta_{MPPT}$$

G_t is the global irradiance on the titled surface in W/m^2 , K_T is thermal derating coefficient of the PV module in $\%/^\circ\text{C}$, A_a area of the PV array in m^2 , η_m is the module efficiency, η_{dust} is 1-the fractional power loss due to dust on the PV array, η_{mis} is 1-the fractional power loss due module mismatch, η_{DCloss} is 1-the fractional power loss in the dc side, η_{MPPT} is 1-fractional power loss due to the MPPT algorithm, T_c is the cell temperature in $^\circ\text{C}$, T_{ao} is the ambient temperature at STC conditions in $^\circ\text{C}$. The ac power of the PV system is then estimated by using manufacturer's efficiency curve of three phase inverter.

The simplified PV equation adopted for this work is given below [21]:

$$P_{pv} = G_t \cdot A \cdot \eta_{PV} \cdot \eta_{loss} \cdot \eta_{inv} [1 - \gamma \cdot (T_m - 25)] \quad (5)$$

In this equation, miscellaneous losses including dust were lumped together in η_{loss} ; PV cell efficiency η_{PV} and MPPT or inverter efficiency η_{inv} are kept separate. T_m is the module temperature.

The aforementioned equations require detailed modeling of the global irradiance falling on a tilted surface G_t as outlined in the next section.

D. Irradiance Falling on a Tilted Surface: Hottel's equations

There exist various models for calculating irradiance on a tilted panel. However, some of these models rely on other meteorological data such as total irradiance on horizontal surface, diffuse irradiance on horizontal surface, beam normal irradiance. Models of this type include those of Perez [22],[23] and Klucher [22],[24]. Others are not accurate in cloudy conditions, Temps and Coulson [25], or in clear skies, Liu and Jordan [26]. Simple models that require no additional solar measurements were proposed by Hottel [9],[27],[28] and are

adopted in this work. Description of this model is outlined below:

To explain irradiance equations, it is important first to present equations of solar angles as they are a pre-requisite to calculate solar equations.

The derivation of irradiance on tilted surfaces requires the calculation of different solar angles. These equations are mainly based on [29] and [30]. Solar angles that define the position of the sun with respect to a PV plane are illustrated in Fig. 2.

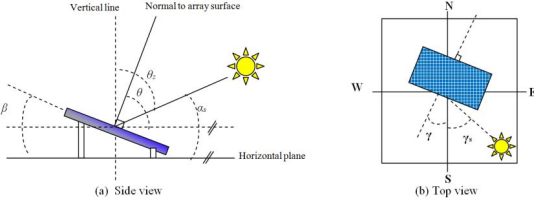


Figure 2. Solar angles of a PV plane

β = Tilt angle of array.

A_s = Solar elevation (altitude): the angle between the horizontal and line to the sun.

θ = Angle of incidence: the angle between normal to array surface and direct irradiance on a tilted surface (or line to the sun).

θ_z = Zenith angle: the angle between vertical line to earth and line to the sun.

γ_s = Solar azimuth angle: the angular displacement from south of the projection of beam radiation on the horizontal plane. Displacements east of south are negative and west of south are positive.

γ = Surface azimuth angle: the deviation of the projection on a horizontal plane of the normal to the surface from the local meridian, with zero due to south, east negative, and west positive; $-180^\circ \leq \gamma \leq 180^\circ$.

The zenith angle θ_z can be written as follows:

$$\cos \theta_z = \cos \varphi \cdot \cos \delta \cdot \cos \omega + \sin \varphi \cdot \sin \delta \quad (6)$$

Where

δ is the declination angle given by

$$\delta = 23.45 \cdot \sin \left(360 \cdot \frac{284 + n}{365} \right) \quad (7)$$

φ is the latitude in degrees is the angular location north or south of the equator, north positive; $-90^\circ \leq \varphi \leq 90^\circ$.

ω is the hour angle which is the angular displacement of the sun east or west of the local meridian due to rotation of the earth on its axis at 15° per hour; morning negative, afternoon positive. The hour angle can be calculated by first calculating the solar time given by:

$$\text{Solar time} = \text{standard time} + 4 \cdot (L_{st} - L_{loc}) + E \quad (8)$$

Where L_{st} is the standard meridian for the local time zone, L_{loc} is the longitude of the location in question, and longitudes are in degrees west. The parameter E is the equation of time in minutes and is given by:

$$E = 22.92(0.000075 + 0.001868 \cos B - 0.032077 \sin B - 0.014615 \cos 2B - 0.04089 \sin 2B) \quad (9)$$

Where B is calculated as follows:

$$B = (n - 1) \frac{360}{365} \quad (10)$$

The hour angle ω can then be written as:

$$\omega = (\text{Solar time} - 12) \cdot 15 \quad (11)$$

Furthermore, the incidence angle θ can be calculated using the following formula:

$$\begin{aligned} \cos \theta = & \sin \delta \sin \varphi \cos \beta - \sin \delta \cos \varphi \sin \beta \cos \gamma \\ & + \cos \delta \cos \varphi \cos \beta \cos \omega \\ & + \cos \delta \sin \varphi \sin \beta \cos \gamma \cos \omega \\ & + \cos \delta \sin \beta \sin \gamma \sin \omega \end{aligned} \quad (12)$$

The solar irradiance falling on a tilted surface, G_t (W/m^2) is composed of three parts: the direct irradiance G_{tb} (W/m^2), the diffuse irradiance G_{td} (W/m^2) and reflected irradiance G_{tr} (W/m^2), i.e.

$$G_t = G_{tb} + G_{td} + G_{tr} \quad (13)$$

The three components of irradiance can be calculated as follows:

$$G_{tb} = G_{on} \tau_b \cos \theta \quad (14)$$

$$G_{td} = G_{on} \cos \theta_z \tau_d \cdot \frac{(1 + \cos \beta)}{2} \quad (15)$$

$$G_{tr} = \rho \cdot G_{on} \cos \theta_z \tau_r \cdot \frac{(1 + \cos \beta)}{2} \quad (16)$$

Where G_{on} is the extraterrestrial radiation (W/m^2), τ_b is the beam atmospheric transmittance, τ_d is the diffuse atmospheric transmittance, and τ_r is the reflected atmospheric transmittance. G_{on} can be calculated as follows:

$$G_{on} = G_{sc} \cdot \left(1 + 0.033 \cdot \cos \left(\frac{360 \cdot d}{365} \right) \right) \quad (17)$$

Where G_{sc} is $1367 \pm 5 \text{ W/m}^2$ and d is the day of the year.

The beam atmospheric transmittance τ_b can be calculated as follows:

$$\tau_b = a_0 + a_1 \cdot e^{-k / \cos \theta} \quad (18)$$

Where a_0 , a_1 , and k are constants that can be calculated as follows:

$$a_0 = r_0 \cdot [0.4237 - 0.00821(6 - A)^2] \quad (19)$$

$$a_1 = r_1 \cdot [0.5055 + 0.00595(6.5 - A)^2] \quad (20)$$

$$k = r_k \cdot [0.2711 + 0.01858(2.5 - A)^2] \quad (21)$$

Where A is the altitude of the location in km, r_0 , r_1 , and r_k are correction factors for different types of climates and are given in Table I.

TABLE I. COFFIECIENTS VALUES DEPENDENT ON CLIMATE

Climate type	r_0	r_1	r_k
Tropical	0.95	0.98	1.02
Midlatitude Summer	0.97	0.99	1.02
Subarctic Summer	0.99	0.99	1.01
Midlatitude Winter	1.03	1.01	1.00

III. PV SYSTEM DESCRIPTION AND IDENTIFICATION VIA PSO

The test-bed system, located in Masdar city close to Abu Dhabi airport, is a $220,000 \text{ m}^2$, 10MW PV plant [31],[32]. The plant consists of around 87,777 panels: 17,777 are

polycrystalline and 70,000 are thin-film from Suntech and First Solar respectively. The parameters for the model were taken from data sheets of panels [33],[34] and from Engineers in Masdar and are given in Table II. These are the parameters with best engineering values taken from data sheet and engineers of the PV system. The model with these values will be referred to as the *clear box* model.

TABLE II. PARAMETERS OF CLEAR BOX PV MODEL

Longitude	54.45°
Latitude	24.43°
Altitude	1 m
Area of SunTech panels, A_{Sun}	30,911 m ²
Area of First Solar Panels, A_{First}	72,500 m ²
Miscellaneous losses (Suntech group), η_{loss_Sun}	5% (i.e. $\eta_{loss_Sun} = 95\%$)
Miscellaneous losses (FirstSolar group), η_{loss_First}	6% (i.e. $\eta_{loss_First} = 94\%$)
Efficiency of PV panel (Suntech), η_{PV_Sun}	11%
Efficiency of PV panel (FirstSolar), η_{PV_First}	10%
Efficiency of Inverter (Suntech panels), η_{inv_Sun}	95%
Efficiency of Inverter (FirstSolar), η_{inv_First}	94%
Temperature coefficient (%/C°), γ	5%
r_o	0.27
r_1	0.29
r_2	0.32
Albedo, ρ	0.35

A. CAutoD for Uncertainties in Clear-box Model:

The clear box model assumes different parameters with the values outlined in data sheet or are of constant value throughout the system. However and in reality values *change* with variation in atmospheric conditions and with aging of the materials. For the PV system studied in this work, some parameters were of uncertain value and thus were candidates for exploration of *better* practical values. These are the different PV efficiencies, albedo of ground, temperature coefficient, and miscellaneous losses; eight parameters in total. The practical ranges of these values are given in Table III. The identification of the best or optimized parameter value was conducted through PSO as explained in the preceding paragraphs.

TABLE III. RANGE OF UNCERTAIN PARAMETERS FOR CAutoD
CLEAR BOX PV MODEL

η_{loss_Sun}	92 - 96%
η_{loss_First}	92 - 96%
η_{PV_Sun}	10 - 12 %
η_{PV_First}	10 - 12 %
η_{inv_Sun}	92 - 99 %
η_{inv_First}	92 - 99%
γ	2% - 6%
r_o	0.2 - 1.0
r_1	0.2 - 1.0
r_2	0.2 - 1.0
ρ	0.3 - 0.5

B. Grey-box PV Model:

Enhancement of the model was explored by introducing black-box models to account for unknown effects. The parameters of efficiency of inverter are known to be adaptive and are function of their loading [35]. Therefore, the following two grey-box models are proposed to model the change in efficiency with loading:

$\eta_{inv_Sun} = 0.45 + a_1 \cdot P_o + a_2 \cdot P_o^2 + a_3 \cdot P_o^3 + a_4 \cdot P_o^4 + a_5 \cdot P_o^5$ $\eta_{inv_First} = 0.45 + a_6 \cdot P_o + a_7 \cdot P_o^2 + a_8 \cdot P_o^3 + a_9 \cdot P_o^4 + a_{10} \cdot P_o^5$	Grey-box 1
$\eta_{inv_Sun} = \frac{0.45 + a_1 \cdot P_o + a_2 \cdot P_o^2 + a_3 \cdot P_o^3}{1 + a_4 \cdot P_o + a_5 \cdot P_o^2 + a_6 \cdot P_o^3}$ $\eta_{inv_First} = \frac{0.45 + a_7 \cdot P_o + a_8 \cdot P_o^2 + a_9 \cdot P_o^3}{1 + a_{10} \cdot P_o + a_{11} \cdot P_o^2 + a_{12} \cdot P_o^3}$ $P_o = \frac{P_{DC}}{P_{DCo}}$	Grey-box 2

where P_o is the fractional loading calculated by dividing the P_{dc} output of PV by nominal DC rating of the inverter.

The first model, Grey-box 1, is of a polynomial form while the second, Grey-box 2, is based on Padé approximation. Both will be explored to find the most suitable model for this application. The coefficients of these models were identified using PSO.

PSO was chosen as an identification algorithm to optimize the parameters in the CAutoD clear model and to find best coefficients in the grey box models respectively. The objective function for the identification is to minimize the Root Mean Square Error (RMSE) between actual and identified model in terms of PV power output. The RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (P_a^i - P_p^i)^2}{n}} \quad (22)$$

Where P_a^i is the i^{th} actual output power, P_p^i is the i^{th} predicted power by model, and n is number of data points.

Particle Swarm Optimization [36] is inspired by social behaviour of bird flocking or fish schooling. It can be applied as follows:

Step 1: Initialize a population (array) of particles with random positions and velocities v on d dimension in the problem space. The particles are generated by randomly selecting a value with uniform probability over the d^{th} optimized search space $[x_d^{min}, x_d^{max}]$.

Step 2: For each particle x , evaluate the desired optimization fitness function, J , in d variables.

Step 3: Compare particles fitness evaluation with x_{pbest} , which is the particle with best local fitness value. If the current value is better than that of x_{pbest} , then set x_{pbest} equal to the current value and x_{pbest} locations equal to the current locations in d -dimensional space.

Step 4: Compare fitness evaluation with population overall previous best. If current value is better than x_{gbest} , the global best fitness value then reset x_{gbest} to the current particle's array index and value.

Step 5: Update the velocity v as follows:

$$v_{id}(k) = w(k) v_{id}(k-1) + \varphi_1 \cdot rand_1(x_{idpbest}(k-1) - x_{id}(k-1)) + \varphi_2 \cdot rand_2(x_{idgbest}(k-1) - x_{id}(k-1)) \quad (23)$$

where, k is the number of iteration, i is the number of the particles that goes from 1 to n , d is the dimension of the variables, and $rand_{1,2}$ is a uniformly distributed random

number in (0, 1), $\phi_{1,2}$ are acceleration constants and are set, as recommended by investigators [36], equal to 2. The weight w is often decreased linearly from about 0.9 to 0.4 during the search process.

Step 6: Update position of the particles,

$$x_{id}(t) = v_{id}(t) + x_{id}(t-1) \quad (24)$$

Step 7: Loop to Step 2, until a criterion is met, usually a good fitness value or a maximum number of iterations (generations) m is reached.

PSO identification will search for 8 parameters in the CAutoD clear box model, 10 parameters in Grey-box model 1, and 12 parameters in Grey-box model 2.

IV. SIMULATION RESULTS AND DISCUSSION

The models described in the preceding sections are simulated and compared as outlined below. The clear irradiance model based on Hottel's equations (6)-(21) is simulated and illustrated in Fig. 3. Furthermore, the clear PV model is simulated with values given in Table II. For the CAutoD clear box model, PSO is used to search for the optimum values of the uncertain parameters in the clear-box model. The data available from the system was used in two phases: first to tune the models and second to test or validate the model. To train the model, data (hourly PV power and module temperature) of days 5-20 in July 2010 (summer) and days 5-20 in January 2011 (winter) were used. The fitness function of the identification is chosen as the average error of July and January as shown below:

$$J = RMSE_{train} = (RMSE_{July} + RMSE_{January}) / 2 \quad (25)$$

Once models are identified, the five consecutive days in both July 2010 and January 2011 are used to test the models and to compute the average forecasting $RMSE$ ($RMSE_{test}$). For PSO, the number of particles is set to $n = 50$ with a maximum number of search iterations of 300.

The results of the four models: **clear-box**, **CAutoD clear-box**, **Grey-box 1**, and **Grey-box 2** are summarized below. The coefficients identified by PSO for the respective models are given in Table IV. In the same table, RMSEs for training and for testing are given for the four models. The progress for identification is given in Fig. 4 for three optimized models. PV power predicted (test data) by the four models is given in Figures 5 and 6. The difference between accuracy of the models is seen better by analyzing the errors of modeling shown in Figures 7 and 8.

The following observations can be deduced from the results:

1. In general, tuning the parameters of the clear-box model proved to enhance the forecasting capabilities of the model. This was clear in the CAutoD clear-box, Grey-box 1, and Grey-box 2. Parameters from manufacturers require further adaptation to the unique atmospheric and location conditions of a given PV system. The $RMSE$ s of tuned models are generally better than that of the clear-box model.

2. On the whole, introducing grey-box model enhanced the modeling accuracy compared with clear-box and optimized CAutoD clear-box as evident from both the training and testing the models. This can also be seen error plots in Figures 7 and 8.
3. PSO exhibited stagnation in identifying clear-box and grey-box models but at different times of progress. CAutoD clear-box identification was first to go through stagnation followed by Grey-box 2, though grey-box 2 reached a better $RMSE$. Grey-box 1 was better at stagnation indicating a better suited model for tuning via PSO.
4. Although, Grey-box 2 model produced the best $RMSE_{train}$, it was slightly surpassed by Grey-box model 1 in the testing phase. It can be said that Grey-box 2 exhibited *generalization* issue in comparison with the less complex Grey-box model 1.
5. Increasing or decreasing the order of Grey-box 1 and 2 was found to deteriorate the models accuracy and therefore kept at the given values.
6. General observation: the simplified model, equation 5, was found more sensitive to module temperature in July than in January i.e. excluding the temperature coefficient part (in brackets) had higher impact (worsen accuracy) in July than in January. This is expected as higher temperatures in July will impact the performance of the PV panels.

The following areas of further improvements were identified:

1. Different identification methods can be explored to further enhance the modeling capabilities. For example, PSO can be further enhanced to overcome the stagnation issue.
2. Different structure for grey-box models can be explored. Also, clear-box model can be further explored to identify candidate parameters to be replaced by black-box models to account for uncertainties.

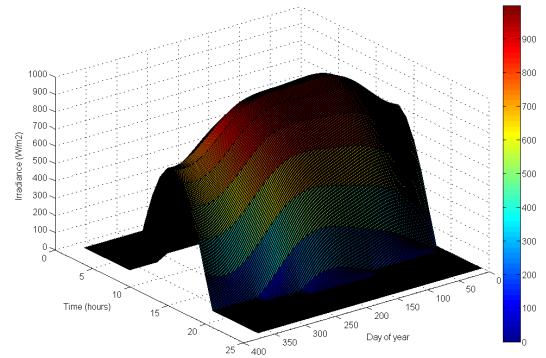


Figure 3. Clear-day irradiance model for one year

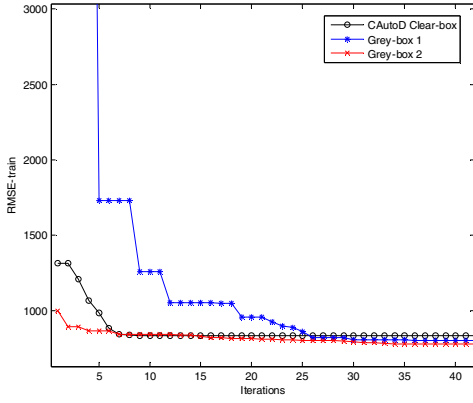


Figure 4. Progress of PSO identification for different models

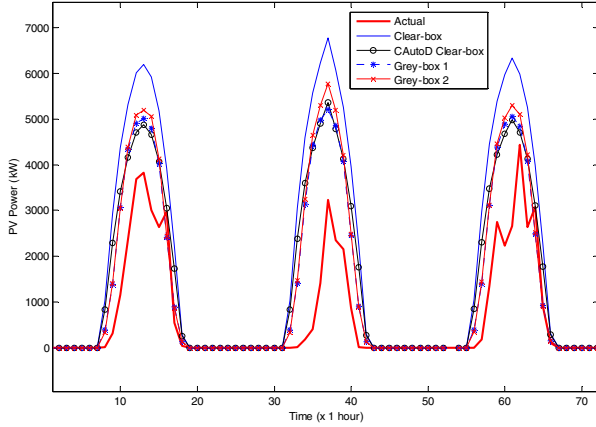


Figure 5. PV power predicted for different models: January 2011 (20-22 January). These were cloudy days.

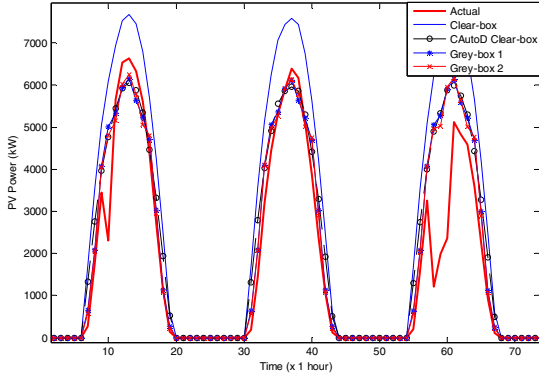


Figure 6. PV power predicted by different models: July 2010 (20-22 July).

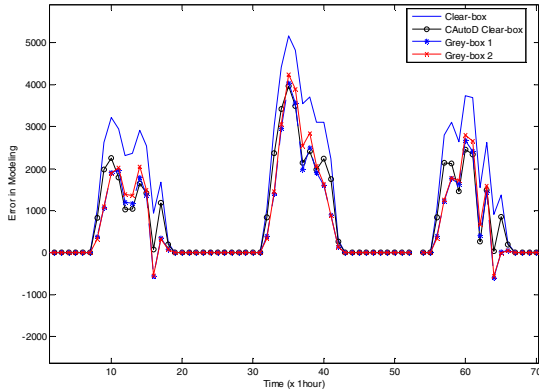


Figure 7. Testing error in modeling for different models: January 2011 (20-22 January).

TABLE IV. IDENTIFIED MODELS AND RMSE

Parameters	Clear-box	CAutoD clear-box	Grey-box 1	Grey-box 2
η_{loss_Sun}	95%	92%	96%	96%
η_{loss_First}	94%	96%	96%	96%
η_{PV_Sun}	11%	10%	10%	10%
η_{PV_First}	10%	10%	10%	12%
η_{inv_Sun}	95%	92%	Replaced by Grey box	
η_{inv_First}	94%	92%		
γ (%/Co)	5%	5.3%	2.3%	6%
ro	0.27	0.2	1	0.2
$r1$	0.29	0.2	0.2	0.2
$r2$	0.32	1	1	0.2
ρ	0.35	0.3	0.3	0.5
a_1	Not applicable		-7.11	-2.42
a_2			0.55	5.21
a_3			6.36	-4.58
a_4			-1.46	-2.23
a_5			-5.24	-2.97
a_6			0.054	7.34
a_7			2.55	-1.32
a_8			3.73	4.16
a_9			-8.20	6.47
a_{10}			3.37	3.70
a_{11}			Not applicable	
a_{12}				
$RMSE_{train}$, kW	1322	835	775	770
$RMSE_{test}$, kW	1636	1067	993	1010

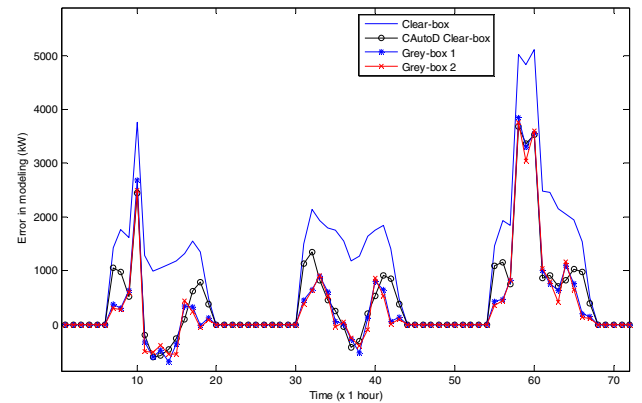


Figure 8. Testing error in modelling for different models: July 2010 (20-22 July).

V. CONCLUSION

The enhancement of modeling and forecasting of clear-box PV power models through introduction of black-box models was discussed in this paper. It was found that practical values of parameters can be tuned to improve the accuracy of the models. Further enhancement can be achieved through the introduction of grey-box models to account for uncertainties in the PV models. A free-derivative particle swarm engine was utilized in the identification process and was found particularly beneficial with simple grey-box models. The work presented is a novel step towards exploring the benefits grey-box models can add in the PV forecasting and hence supporting a better integration of these renewable energy sources in power networks.

REFERENCES

- [1] Global Market Outlook for photovoltaics 2013-2017, European Photovoltaic Industry Association (EPIA), May 2013.
- [2] R. M. Swanson, "Photovoltaic Power Up", Science, Vol. 324, pp. 891-892, 15 May 2009.
- [3] IRENA reference: "Renewable Energy Technologies: Cost Analysis Series, Photovoltaics", IRENA working paper, Volume 1: Power sector, Issue 4/5, June 2012.
- [4] A. Woyte, V. V. Thong, R. Belmans, J. Nijs, "Voltage Fluctuations on Distribution Level Introduced by Photovoltaic Systems," IEEE Trans. On Energy Conversion, Vol. 21, No. 1, pp. 202-209, March 2006.
- [5] M. Sandoval, S. Grijalva, "An assessment on the impacts of photovoltaic systems on operational costs of the grid? The case of the state of Georgia," 2013 IEEE PES Conference On Innovative Smart Grid Technologies Latin America (ISGT LA), pp.1-6, 15-17 April 2013.
- [6] J. Ma, S. Lu, R. P. Hafen, P. V. Etingov, Y. V. Makarov, V. Chadliev, "The impact of solar photovoltaic generation on Balancing Requirements in the Southern Nevada system," Transmission and Distribution Conference and Exposition (T&D), 2012 IEEE PES, pp.1-9, 7-10 May 2012.
- [7] T. Stetz, F. Marten, M. Braun, "Improved Low Voltage Grid-Integration of Photovoltaic Systems in Germany," IEEE Transactions on Sustainable Energy, vol.4, no.2, pp.534-542, April 2013.
- [8] L. Cibulka, M. Brown, L. Miller, A. V. Meier, "User Requirements and Research Needs for Renewable Generation Forecasting Tools that will Meet the Need of the CAISO and Utilities for 2020," A White Paper Report Prepared by CIEE, September 2012.
- [9] C. Tao, D. Shanxu, C. Changsong, "Forecasting power output for grid-connected photovoltaic power system without using solar radiation measurement," 2010 2nd IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG), pp.773-777, 16-18 June 2010.
- [10] U. Forsell, P. Lindskog, "Combining semi-physical and neural network modelling: An example of its usefulness" Proceedings of the 11th IFAC symposium on system identification, Vol. 4, pp. 795-798, Kitakyushu, Japan, 1997.
- [11] N. Al-Messabi, Yun Li, I. El-Amin, C. Goh, "Forecasting of photovoltaic power yield using dynamic neural networks," The 2012 International Joint Conference on Neural Networks (IJCNN), pp.1-5, 10-15 June 2012.
- [12] Y. d. Valle, G. K. Venayagamoorthy, S. Mohagheghi, J.-C. Hernandez, R. G. Harley, "Particle Swarm Optimization: Basic Concepts, Variants and Applications in Power Systems" IEEE Transactions on Evolutionary Computation, Vol. 12, No. 2, pp. 171-195, April 2008.
- [13] P. Bacher, H. Madsen, H. A. Nielsen, "Online short-term solar power forecasting," Solar Energy, Vol. 83, pp. 1772-1783, 2009.
- [14] B. Kroposki, K. Emery, D. Myers, and L. Mrig, "A Comparison of Photovoltaic Module Performance Evaluation Methodologies For Energy Ratings," 24th IEEE Photovoltaic Conference, pp. 858-862, 1994.
- [15] A. Mellit, A. M. Pavan, "A 24-h Forecast of Solar Irradiance using Artificial Neural Network: Application for Performance Prediction of a Grid-connected PV plant at Trieste, Italy," Solar Energy, Vol. 84, pp. 807-821, 2010.
- [16] L. A. Fernandez-Jimenez, A. Munoz-Jimenez, A. Places, M. Mendoz-Villena, E. Garcia-Garrido, P. Lara-Santillan, E. Zorzano-Alba, J. Zorzano-Santamaria, "Short-term Power Forecasting System for Photovoltaic plants," Renewable Energy, Vol. 44, pp. 311-317, 2012.
- [17] J. A. Gow, C. D. Manning, "Development of a Photovoltaic Array Model for use in Power-Electronics Simulation Studies," IEE proceedings on Electric Power Applications, Vol. 146, no. 2, pp. 193-200, March 1999.
- [18] M. G. Villalva, J. R. Gazoli, E. R. Filho, "Comprehensive Approach to Modeling and Simulation of Photovoltaic Arrays," IEEE Transactions on Power Electronics, Vol. 24, No. 5, May 2009.
- [19] C. R. Osterwald, "Translation of Device Performance Measurements to Reference Conditions," Solar Cells, Vol. 18, pp. 269-279, 1986.
- [20] W. Omran, M. Kazerani, M. M. A. Salama, "A Clustering-Based Method for Quantifying the Effects of Large On-Grid PV Systems," IEEE Transactions on Power Delivery, Vol. 25, No. 4, October 2010.
- [21] Md. Habibur Rahman and Susumu Yamashiro "Novel Distributed Power Generating System of PV-ECaSS Using Solar Energy Estimation," IEEE TRANSACTIONS ON ENERGY CONVERSION, VOL. 22, NO. 2, JUNE 2007
- [22] R. Perez, "An anisotropic model for the diffuse radiation incident on slopes of different orientations," Proceedings of ASSES, Minneapolis, pp. 883-888, 1983.
- [23] R. Perez, "A new simplified version of the Perez diffuse irradiation model for tilted surfaces," Solar Energy, Vol. 39, No. 3, pp. 221-231, 1987.
- [24] T. M. Klucher, "Evaluation of models to predict insolation on tilted surfaces," Solar Energy, Vol. 23, pp. 111-114, 1979.
- [25] R. C. Temps, K. L. Coulson, "Solar radiation incident upon slopes of different orientations," Solar Energy, Vol. 19, pp. 179-814, 1977.
- [26] B. Y. H. Liu and R. C. Jordan, "The long term average performance of flat-plate solar-energy collectors," Solar Energy, Vol. 7, No.2, pp. 53-74 1963.
- [27] H. C. Hottel, "A simple model for estimating the transmittance of direct solar radiation through clear atmospheres," Solar Energy, vol. 18, no. 2, pp. 129 - 134, 1976.
- [28] Charnon Chupong and Boonyang Plangklang, "Forecasting power output of PV grid connected system in Thailand without using solar radiation measurement," Energy Procedia, Vol. 9, pp. 230-237, 2011
- [29] J. Duffie, W. Beckman, Solar Engineering of Thermal Process, Third Edition, John Wiley and Sons, 2006.
- [30] M. Iqbal, "An introduction to solar radiation," Academic Press Canada, 1983.
- [31] Website: Date 18 April 2015: <http://www.masdar.ae/en/energy/detail/masdar-city-solar-pv-plant>
- [32] 10MW Solar Power Plant, Enviromena Power Systems Report, 2014 Website: http://enviromena.com/2015/wp-content/uploads/2014/01/10MW-Masdar_PD_2013-WEB.pdf
- [33] SunTech: SunTech panel data sheet: STP270S - 20/Wd+, STP265S - 20/Wd+ : www.suntech-power.com IEC-STP-WdS+-NO1.01-Rev 2014
- [34] Thin film: Models: First Solar FS Series 2 PV Module data sheet: www.firstsolar.com, FS Series 2 PV Module PD-5-401-02 NA MAY 2011
- [35] Aleksandar Pregelj, "Impact of Distributed Generation on Power Network Operation," PhD thesis, Georgia Institute of Technology, December 2003.
- [36] R. Eberhat and Y. Shi, Particle Swarm Optimization: Developments, applications, and resources, Proceedings of the 2001 Congress on Evolutionary Computation, vol. 1, 2001, pp. 81-86.

KEY CHALLENGES AND OPPORTUNITIES IN HULL FORM DESIGN OPTIMISATION FOR MARINE AND OFFSHORE APPLICATIONS

Joo Hock ANG^{1,2}, Cindy GOH², Yun LI²

1) Sembcorp Marine Ltd., Singapore 759956

2) School of Engineering, University of Glasgow, Glasgow, G12 8LT, UK

Correspondence email: joothock.ang@sembmarine.com

Abstract— New environmental regulations and volatile fuel prices have resulted in an ever-increasing need for reduction in carbon emission and fuel consumption. Designs of marine and offshore vessels are more demanding with complex operating requirements and oil and gas exploration venturing into deeper waters and harsher environments. Combinations of these factors have led to the need to optimise the design of the hull for the marine and offshore industry. The contribution of this paper is threefold. Firstly, the paper provides a comprehensive review of the state-of-the-art techniques in hull form design. Specifically, it analyses geometry modelling, shape transformation, optimisation and performance evaluation. Strengths and weaknesses of existing solutions are also discussed. Secondly, key challenges of hull form optimisation specific to the design of marine and offshore vessels are identified and analysed. Thirdly, future trends in performing hull form design optimisation are investigated and possible solutions proposed. A case study on the design optimisation of bulbous bow for passenger ferry vessel to reduce wave-making resistance is presented using NAPA software. Lastly, main issues and challenges are discussed to stimulate further ideas on future developments in this area, including the use of parallel computing and machine intelligence.

Keywords— *Simulation-based hull form design optimisation; geometry modeling; shape transformation; performance evaluation; computational fluid dynamic (CFD)*

I. INTRODUCTION

Increasing environmental regulations and fuel price volatility are two top concerns faced by the marine and offshore industry today. As a consequence, eco-friendly shipping and fuel efficiency are now key design criteria for marine and offshore vessels. Furthermore, the need for bigger vessels and new operating conditions have pushed the frontiers of ship and offshore vessel design beyond conventional design boundaries. For example, as oil and gas exploration ventures into deeper waters and harsher environment, the technical requirements for offshore structures and vessels operating in these environments also increases tremendously. The shape of the hull is tightly coupled to the efficiency of the vessel and has direct impact on its hydrodynamic performance. It is, therefore, a crucial aspect to optimise in order to achieve an overall improvement in vessel performance. Simulation based design (SBD) is widely used in engineering

applications and is well-known to improve product performance and design efficiency [1].

Computer-aided design (CAD) tools are commonly used in ship design firms and shipyards for modelling and hydrodynamic evaluation purposes. Simulation-based hull form optimisation, in this case, offers a potential solution to overcome the challenges faced in the design of marine and offshore vessels. This paper will hence focus on the state-of-the-art SBD for hull form optimisation. Section II provides an overview of related works in hull form design optimisation, which includes the key processes and techniques applied. Section III will identify the key challenges for each process and highlight the associated opportunities and key trends. Lastly, a case study of optimisation of passenger vessel bulbous bow for reduced resistance will be presented in section IV, followed by conclusions.

II. RELATED WORK IN SIMULATION-BASED HULL FORM DESIGN OPTIMISATION

In the marine and offshore vessel design space, SBD can be used to analyse and improve the hydrodynamic performance of vessels including reduction of resistance and better sea keeping capabilities. Traditional approach to vessel design and testing is a manual and laborious process with a long lead time that could take up to several months to complete. In addition, conventional model test is an expensive process with little tolerance for design error or modifications. Testing can only be done on a single design with minimal variations. A significant advantage of SBD over traditional methods is that it can help shorten the entire design cycle by an appreciable amount. Prior to any physical model testing, SBD can be used to develop ‘sufficiently optimum’ initial designs virtually which can then be further refined afterwards. The hull form being the largest single component of the entire ship or floating structure exerts the most influence not only on hydrodynamic performance, but also production and subsequently operation of the vessel. It is therefore a crucial aspect to optimise in order to achieve maximum gain on the overall design of the vessel. SBD for performing hull form design and optimisation mainly comprises of 3 key processes namely geometry modelling and manipulation algorithms, optimiser and performance evaluations. These processes can be integrated into a

common framework for automated hull form optimisation as illustrated in Figure 1.

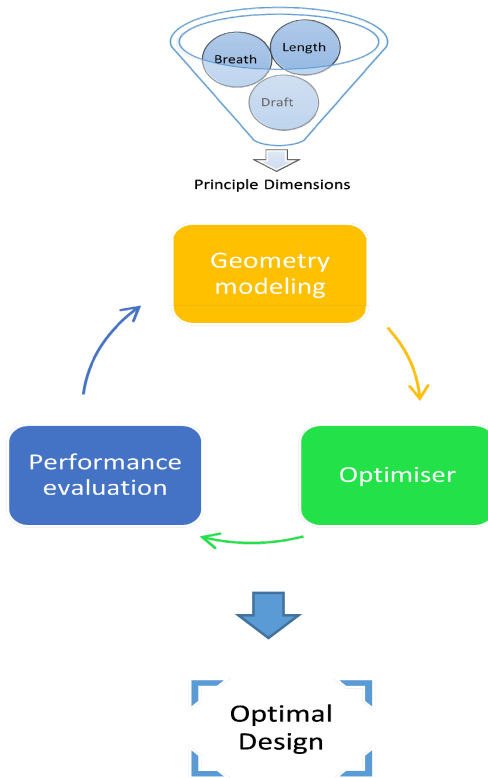


Figure 1. Simulation-based hull form design optimisation framework

A. Geometry Modelling and Shape Transformation

Geometry modelling of the hull form involves the generation of a geometric shape by first defining in points, followed by representing it using curves before transforming into surfaces. It is an important and integral part of hull form design optimisation and has been studied extensively. Key geometry modelling methods for hull form design optimisation includes Beizer curves, B-splines and non-uniform rational B-splines (NURBS) [2]. Beizer curves are defined by a set of control points and are in many aspects similar to interpolation. Nonetheless, unlike the latter, it ‘stretches’ towards, rather than passing through the central control point [3]. B-splines method are developed based on hull form parameters and includes variations such as cubic B splines, fairness optimized B spline or F-splines [4], which helps to overcome the limitation to main algorithm. Due to its ability to accurately represent complex shapes with very small amount of definition data and the ease to manipulate control points, NURBS has become the ‘de-facto’ algorithm for surface representation in hull form geometry design.

After the initial geometry of the hull form is generated, geometry modification or shape transformation methods can then be applied to edit the initial curves and surfaces. In the area of shape transformation, much work has been reported in the literature. Lackenby [5] and 1-Cp are one of the earliest geometry modification methods in hull transformation. It translate the hull section longitudinally using simple

calculation and uses simple scaling functions to deform the forward and aft section [6]. Another common method for shape transformation is parametric modelling [7]. By capturing the essence of the intended shapes and the possible variations, it offers better control over the shape to be optimised and can find improved solutions within a shorter timeframe [8]. A popular method for shape transformation of the hull form is Free-form deformation (FFD). Based on the principles of enclosing complex geometric shapes within simpler ones, it provides mapping to change the coordinate position of more complex shape [9]. In hull form design optimisation, it has been applied to transforming the hull design of catamaran [9] and navy vessel [10].

Geometry modelling and shape transformation are complex and delicate processes in hull form optimisation. This is particularly so when an optimal design has multiple objectives to satisfy. Variations to one parameter may improve the performance of one objective at the expense of another, and as a consequence compromise the overall performance of the design. For instance, reducing the input parameters or “control point” while consistently maintaining a smooth surface or ‘feasible design’ of the hull form are two seemingly conflicting objectives to achieve. It is therefore important to consider the optimisation process as part of the entire SBD framework.

B. Optimisation

Over the years, SBD has benefited greatly from the developments in optimisation algorithms as well as rapid advancements in computational resources. Suffice to say, there is no one-size-fits-all solution and depending on the objectives and application, the most appropriate optimisation algorithm will be used. In the area of hull form design, the main types of optimisation algorithms are gradient based, deterministic methods, heuristic methods and evolutionary methods. Gradient based methods such as Adjoint equations, Gauss Newton method [11], steepest descent, conjugate gradient [12] and Sequential Quadratic Programming (SQP) [13] have been applied to optimise the hull form of tankers, catamarans and container ship for improved efficiency [14-17].

While these algorithms could be used to determine the characteristics of the search space, such as the extrema, they require information on the search gradient which may not always be available. As a result, the solutions found tend to be sub-optimal. This is particularly so for non-convex problems [1]. They are thus more suitable for local fine-tuning after the proximity of the global maxima or minima has been identified. Deterministic methods are another class of optimisation algorithm that have been applied to improve the design and performance of the ship hull. They include a Hill-climbing technique and Nelder and Mead or downhill simplex method [7]. In [18] and [19], heuristic methods such as fish shoal algorithm (FSA) and Simulated Annealing were applied to optimise the hull of catamaran for reduced resistance and global optimisation model for ship design respectively. Simulated annealing is a stochastic optimisation method

which is capable of controlling the deterioration of object function so as to escape local minima [19].

In more recent studies, evolutionary algorithms (EA) have become increasingly popular in hull form design optimisation. EA is a class of generic population-based metaheuristic global optimisation techniques [11]. Inspired and modelled closely after biological evolution, it uses mechanisms of selection, recombination and mutation to traverse the search space for fitter solutions. Unlike conventional optimisation algorithms, EA does not make any assumption of the underlying fitness landscape and scales well to higher dimensional problems, making them suitable for NP-complete problems [19]. In hull form design optimisation, commonly used EAs including Genetic Algorithm (GA) [11] and Particle swarm optimization (PSO) [11] have been applied to improve the hydrodynamic performance of ships [20-23].

C. Performance Evaluation

Performance evaluation is an integral part of the SBD framework and key to ensure that solutions found are true global optima. It also provides a measure to weed out weaker solutions as the optimisation progresses. In the development of ship hull form, resistance is one of the most important performance parameters that need to be considered. Using linear potential code and Computational Fluid Dynamic (CFD), the evaluation of resistance can be carried out during the early stages of ship design. CFD is a popular tool for evaluating new hull form design. By solving field equations that describes the dynamics of fluid flow, CFD is able to provide accurate simulation of fluid flow [24]. CFD have also been used to assess the performance of vessels in terms of propulsion, seakeeping, maneuvering and propeller designs [25]. Prevalent performance evaluation methods include potential flow and Reynolds Averaged Navier-Stokes Equation (RANSE) [12, 26]. Potential flow is governed by Laplace equation and discretised using body surface and free surface panels [27]. It is a very useful algorithm, especially in the analysis of free surface flows [24]. RANSE are used to solve viscous fluid flows and able to represent complex free surfaces, which enables it to accurately evaluate total resistance, propulsion, appendages and added resistance. A key advantage of the RANSE methods is its ability to capture global and localised wave patterns as well as effects of viscosity at full scale [28]. Nonetheless, a common criticism of this method is the high computational resources and time needed to perform an evaluation. In addition, the quality of results obtained differs significantly depending on the settings, user's experience and software provider [24].

Despite the array of methods available, they are not widely adopted in the marine and offshore community. Crafting of new hull forms remains largely a tedious trial-and-error process where most new designs are created through modifications of existing ones based on the designer's experience. This, we believe, is primarily due to the lack of an integrated framework which is capable of (i) capturing designers' experience and knowledge, (ii) translating them into effective problem representations, (iii) efficiently exploring the design search space for new,

innovative and optimal design solutions. Key challenges and bottlenecks in SBD are discussed in detail in the next section.

III. KEY CHALLENGES AND OPPORTUNITIES IN SIMULATION-BASED HULL FORM DESIGN OPTIMISATION

Despite the benefits that SBD offers in hull form design optimisation, its take-up rate in the marine and offshore vessel design space remains limited. Recent developments in optimisation and performance evaluation techniques, coupled with advancements in high performance computing (HPC) have enabled simulation-based design optimisation to be applied successfully in the design of marine and offshore vessels. The key challenges are discussed with potential gaps for improvements and their possible solutions highlighted as follow:

A. Geometry Modelling and Shape Transformation

Traditional CAD approach such as splines and NURBS do not work well for complex shapes - difficult to express in admissibility conditions (tangentially for complex shapes). This often results CAD failures and overestimating of the dimensionality of complex shapes [29]. In some cases, the mathematical definition can be manually manipulated. However, this is an extremely tedious and complex process of trial-and-error where individual vertices must be moved to achieve a smooth surface each time any modification is made [30]. While NURBS are most widely used in geometric modelling of ship hull form, they suffer drawbacks such as large quantity of control points and complications during surface fairing [1]. Furthermore, the resulting geometry is constrained in that designers are not able to adjust control points to achieve the desired shapes [31]. Therefore, there is a need for better geometry definition methods that allows design flexibility and can be representative enough for the selected optimiser and solvers to code and decode easily. One possible approach is by introducing a new spline-based design scheme, such as the partial shape preserving (PSP)-spine basis function [32] which can be used for preserving partial shape when blended with different existing designs.

While most modification methods are fast to execute and effective in achieving a smooth hull form, their capacity to explore the search space for optimum design is often limited. As such, the shape to be transformed is limited and optimal results are not guaranteed. A more robust, shape manipulation method that can link seamlessly to the optimiser would therefore be an ideal solution to alleviate this limitation.

B. Optimisation

The major challenge for most hull form optimisation is the development of a fully automated optimisation process [33], where key processes such as geometry representation, shape variation and performance evaluation can be integrated in a seamless manner. Numerical solution usually varies from user and can produce errors arising from discretization [33]. A common problem in existing optimisation algorithm applied to hull form design is the lack of a robust and effective way to produce meaningful and feasible solutions [34]. This in part is due to the search mechanisms used. However, a

larger issue lies in the lack of an effective and accurate mapping between the phenotype and genotype to accurately represent the problem. It is a crucial that a large solution set can be effectively represented.

Whilst it is more accurate, direct solver are impractical and time consuming to be applied in hull form optimisation. Advance solvers such as heuristic solvers in this case can help to improve the efficiency drastically but they cannot guarantee to obtain the optimal solution [11]. Approximation techniques or surrogate model can also be used to improve the overall computational cost without over-compromising accuracy of optimisation solution. In performing hull form optimisation, some commonly used approximation techniques include Kriging method [35], artificial neural networks [36] and polynomial response surface methodology (RSM) [37]. Another promising approximation technique in hull shape optimisation is Karhunen-Lo'eve expansion (KLE), which can be used effectively to reduce design space dimension while allowing high degree of geometry modification [38].

Furthermore, although ship design is a multi-disciplinary problem consisting of multiple design specifications, attention has been largely focused on solving it as a single objective problem [39]. Unfortunately, real-world engineering applications encompass multi-parameters and disciplines, where improvement in one specific aspect could potentially worsen another [10]. To advance SBD tools, one needs to also consider multi-disciplinary design specifications such as fluid structure interaction or hydro-elasticity.

C. Performance Evaluation

Investigating ship hydrodynamics is practically challenging. This is due to the fact that computing viscous resistance requires very fine grids to be modelled around the ship hull [1]. CFD based RANSE method are effective in accurately predicting the fluid dynamics, however, are computationally expensive when applied in hull form optimisation. Advance HPC and parallel computing can be utilise in this case to reduce the time for performance evaluation using CFD significantly. Some RANSE solvers in this case, allow the computation to start from an arbitrary approximation. It helps to save computation time by leveraging on existing flow field approximation for subsequent computation while initial computation are in progress [28].

Results obtained from potential flow or CFD may not be as accurate as compared to model testing, where unknown phenomenon can arise and missed out in the numeric calculation. However, recent advances in CFD had improved in accuracy tremendously and it definitely can provide some good indication during initial design stage and can be used to narrow down to few 'optimum' design before final model testing.

D. Other Relevant Factors

There are common skepticisms within ship design firms and shipyards on the use of optimisation and CFD procedures in the design process. This is partly due to the lack of expert knowledge on the concept behind the methods as well as sufficient training to use the tools. In some cases, this is directly link to the underlying

mechanisms of the methods, for example, in 'black-box' or model-free approaches where users have little or no understanding of how results are obtained. Results obtained from hull form design optimisation can also vary widely depending on the designer's experience and the algorithms used. In particular, for new or novel designs, unknown phenomenon may be neglected in the simulation resulting in erroneous designs. A potential solution to reduce the dependency and experience of designer here is to incorporate design knowledge extraction techniques such as data-mining techniques [37] or machine learning. When incorporated into hull form design framework, these techniques can help to acquire useful design knowledge about the hull form.

There are many commercial software that offers good SBD solution but they have usually rigid system configuration that does not allow users to customise to their specific needs. As such, inter-portability of models between different software can be an issue and even if they are compatible, much effort is still needed to clean-up and modify the transferred models. An opportunity here is the development of hull surface design tool that enable data to be exchanged between different systems using commonly supported representation with high degree of accuracy [29].

A SWOT analysis, summarising the aforementioned challenges and opportunities, is presented in Figure 2.

(S) Strengths	(W) Weaknesses
<ul style="list-style-type: none"> - Formal hull form optimisation procedure instead of ad-hoc design improvement - More efficient and cost effective as compared to model testing - Improved performance with optimised hull form design 	<ul style="list-style-type: none"> - Development of fully automated loop still lacking - Current numerical methods cannot guarantee optimal solution - Computational expensive especially for CFD calculations - Existing state-of-art lack robustness and practicability
(O) Opportunities	(T) Threats
<ul style="list-style-type: none"> - Advance computation methods such as approximation multi-disciplinary, data-mining, machine learning - Development of more automated and robust techniques to improve overall usability and efficiency of hull form design optimisation 	<ul style="list-style-type: none"> - Accuracy of result depends greatly on designer experience and input - Unknown phenomenon for novel hull design - 'Black-box' function

Figure 2. SWOT analysis

There is no doubt that both SBD and hull form optimisation has come a long way. Nonetheless, the above challenges would need to be adequately addressed before these techniques can be more widely accepted and implemented in the design of marine and offshore vessels.

IV. CASE STUDY OF SIMULATION-BASED HULL FORM DESIGN OPTIMISATION

To illustrate the application of SBD to hull form design optimisation, a case study of a 180 meters passenger ferry vessel was carried out using NAPA design solutions. The objective is to modify the shape of an existing bulbous bow design in order to improve the wave making resistance. Unlike conventional manual SBD

method where hull form optimisation are carried out individually and manually, purpose of this study is to demonstrate the practicability and efficiency of a fully automated hull form optimisation solution by integrating the different processes into one common environment. As such, this procedure was carried out using manager function of NAPA and the steps are given as follows:

1) The vessel was first modelled in NAPA hull modelling tool. For shape transformation, free-form deformation (FFD) technique was selected. The bulbous bow shape was modified simply by creating boundary box and extending the length by 3 metres toward forward direction (fig 3).

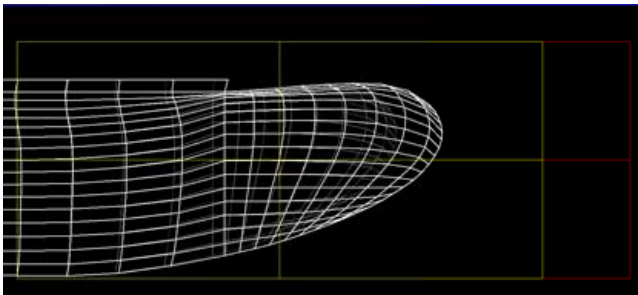


Figure 3. Free Form Deformation (FFD) of bulbous bow area

2) The hull surface mesh was created under NAPA panelisation manager, followed by evaluation of hull panelised model using potential flow method.

3) Selected MOGA for optimisation with setting of 10 population, 20 generations and encoding scheme linked to FFD. Total time taken for calculation was 30 minutes using standard intel Core i7-4702 CPU workstation. Optimum design was subsequently obtained demonstrating a slight reduction of 3% to forward end resistance coefficient.

4) The wave and pressure profile of hull model with modified bulbous bow are then exported to post-processor ParaView as shown in figure 4.

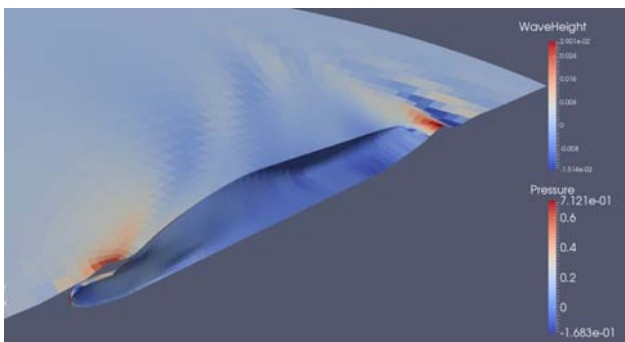


Figure 4. Wave profile of hull model with modified bow

The following findings have been observed from the above case study:

- This study has demonstrated the efficiency and effectiveness of an automated hull form optimisation process using FFD, MOGA and CFD techniques. Although the improvement is

marginal, it provides a good indication of how the hull should be modified to further improve the overall performance. This is useful to the designer and can be used as a starting point to further investigations using more detailed procedures such as RANSE.

- Hull geometry needs to be properly modelled and adequately smooth. Failing to do so will result in panelisation error as well as evaluation results. It was observed geometry modelling of ship hull was one of the most time consuming task.
- Free-form deformation was applied easily and it is capable of producing rather smooth curve without discontinuity in the geometry. However, it was noted the coordinate to move greatly depends on the experience of designer, which makes it difficult to locate the Pareto optimal.
- In order to increase the chance of obtaining the Pareto optimum set, more shape variations should be tested and used as inputs for the optimisation. However, this tends to be a 'trial and error' process and again depends greatly on the designer's experience for input. This thereby presents another area for improvement.

V. CONCLUSION

Due to stringent environmental regulations and volatile fuel prices, marine and offshore vessels are now expected to be more eco-friendly and fuel efficient. In this context, simulation-based hull form design optimisation is gaining increasing attention and importance. It has been demonstrated to be a very efficient and cost-effective tool as compared to ad-hoc manual design improvements and model testing. Key processes in simulation-based hull form design consist of geometry modelling and shape transformation, optimisation and performance evaluation. Main challenges in performing hull form optimisation include scepticism due to 'black-box' functions, high computational costs, automated optimisation loop, for example. Recent developments in advance computation techniques and more powerful computers have presented multiple opportunities such as multi-disciplinary optimisation, machine learning, approximation methods, and optimised hull form design based on variable speed or multi-draft. It has been demonstrated in the case study of bulbous bow optimisation that formal hull form optimisation can be applied successfully using combination of potential flow code method, Free-Form Deformation and MOGA techniques, to improve forward end resistance of ferry vessel. By addressing the challenges and further developing on the opportunities presented, it is envisaged that simulation-based hull form optimisation can become more accepted and widely applied in marine and offshore vessel design.

ACKNOWLEDGMENT

The authors would like to thank NAPA for granting the usage of their 'eco-design' packages and support provided for the case study applied in this paper.

REFERENCES

- [1] R. Sharma, Tae-wan Kim, Richard Lee Storch, Hans (J.J.) Hopman, Stein Ove Erikstad, "Challenges in computer applications for ship and floating structure design and analysis," *Comput. Aided Design*, vol. 44, 2012, pp. 166-185.
- [2] Emilio F. Campana, Daniele Peri, Yusuke Tahara, Frederick Stern, "Shape optimization in ship hydrodynamics using computational fluid dynamics," *Comput. Methods Appl. Mech. Engrg.* 196, 2006, pp. 634-651.
- [3] Shengzhong Li and Feng Zhao, "An innovative hull form design technique for low carbon shipping," *J. of Shipping and Ocean Eng.*, 2012.
- [4] Soonhung Han, Yeon-Seung Lee, Young Bok Choi, "Hydrodynamic hull form optimization using parametric models," *J. Mar. Sci. Technol.*, 2012.
- [5] H. Lackenby, "On the systematic geometrical variation of ship forms," *Trans. of the royal inst. of naval architects*, 1950.
- [6] Marcus Bole, "Interactive hull form transformations using curve network deformation," *Conf. Comput. and IT applicat. in Maritime Ind. (COMPIT)*, 2010.
- [7] Claus Abt, Stefan Harries, Justus Heimann, Henning Winter, "From redesign to optimal hull lines by means of parametric modeling," *Conf. Comput. and IT applicat. in Maritime Ind. (COMPIT)*, 2003.
- [8] Francisco Pérez, José A. Suárez, Juan A. Clemente, Antonio Souto, "Geometric modelling of bulbous bows with use of non-uniform rational B-spline surfaces," *J. Mar. Sci. Technol.*, 2007, pp. 83-94.
- [9] Daniele Peri, Emilio F. Campana, Manivannan Kandasamy, Seng Keat Ooi, Pablo Carrica, Frederick Stern, Phil Osborne, Neil Macdonald, Nic de Waal, "Potential flow based optimization of a high speed, foil-assisted, semi-planning catamaran for low wake," in 10th Int. Conf. on Fast Sea Transportation, Athens, Greece, 2009.
- [10] Yusuke Tahara, Daniele Peri, Emilio Fortunato Campana, Frederick Stern, "Computational fluid dynamics-based multiobjective optimization of a surface combatant using a global optimization method," *J. Mar. Sci. Technol.*, 2008, pp. 95-116.
- [11] Slawomir Koziel and Xin-She Yang, "Computational Optimization, Methods and Algorithms," Springer-Verlag Berlin Heidelberg, 2011.
- [12] Scott Perival, Dane Hendrix, Francis Noblesse, "Hydrodynamic optimization of ship hull forms," *Applied Ocean Research* 23, 2001, pp. 337-355.
- [13] Kazuo Suzuki, Hisashi Kai, Shigetoshi Kashiwabara, "Studies on the optimization of stern hull form based on a potential flow solver," *J. Mar. Sci. Technol.*, 2005, pp. 61-69.
- [14] V. Anantha Subramanian, G. Dhinesh, J.M. Deepti, "Resistance optimisation of high speed catamarans," *Canada's Arctic*, Vol. 1, No. 1, 2006, pp. 69-82.
- [15] D. Peri, M. Rossetti, E.F. Campana, "Design optimization of ship hulls via CFD techniques," *J. of Ship Research*, Vol. 45, 2001, pp. 140-149.
- [16] C. Cinquini, P. Venini, R. Nascimbene and A. Tiano, "Design of river sea ship by optimization," *Struct. Multidisc. Optim.* 22, 2001, pp. 240-247.
- [17] Ho-Hwan Chun, "Hull form parameterization technique with local and global optimization algorithm," *Proc. of The Int. Conf. on Marine Technol.*, Dhaka, Bangladesh, 2010.
- [18] Matteo Diez, Andrea Serani, Umberto Iemma, Emilio F. Campana, "A fish shoal algorithm for global derivative-free simulation-based ship design optimization," 17th Numerical Towing Tank Symp., 2014.
- [19] T. Ray, R.P. Gokarn, O.P. Sha, "A global optimization model for ship design," *Comput. in Ind.* 26, 1995, pp. 175-192.
- [20] Dunja Matulja, Roko Dejhalla, "Hydrodynamic optimisation of forepart of ship," *Proc. of XX Symp.*, Zagreb, Croatia, 2012.
- [21] Grzegorz Mazerski, "Optimisation of FPSO's main dimension using Genetic Algorithm," *Proc. of the ASME 31st Int. Conf. on Ocean, Offshore and Arctic Eng.*, Rio de Janeiro, Brazil, 2012.
- [22] Sheng Huang, Wanlong Ren, Chao Wang, and Chunyu Guo, "Application of an improved particle swarm optimization algorithm in hydrodynamic design," *ICSI 2013, Part I, LNCS 7928*, 2013, pp. 225-231.
- [23] Emilio Fortunato Campana, Giampaolo Liuzzi, Stefano Lucidi, Daniele Peri, Veronica Piccialli, Antonio Pinto, "New global optimization methods for ship design problems," *Optim. Eng.*, Vol.10, 2009, pp. 533-555.
- [24] Volker Betram, "Appropriate tools for flow analyses for fast ships," 4th Int. Conf. High-Performance Marine Vehicles (HIPER), Naples, 2008, pp. 1-9.
- [25] Frederick Stern, Jianming Yang, Zhaoyuan Wang, Hamid Sadat-Hosseini, Maysam Mousaviraad, Shanti Bhushan, Tao Xing, "Computational ship hydrodynamic: Nowadays and way forward," 29th Symp. on Naval Hydrodynamics, Gothenburg, Sweden, 2012.
- [26] Ruosi Zha, Haixuan Ye, Zhirong Shen, Decheng Wan, "Numerical study of viscoius wave making resistance of ship navigation in still water," *J. Marine Sci. Appl.*, vol. 13, 2014, pp. 158-166.
- [27] Horst Nowacki, "Hydrodynamic design of ship hull shapes by methods of computational fluid dynamics," *Progress in Ind. Math. at ECMI*, 1996, pp. 232-251.
- [28] Karsten Hochkirch, Benoit Mallol, "On the importance of full-scale CFD simulations for ships," *Conf. Comput. and IT applicat. in Maritime Ind. (COMPIT)*, 2013.
- [29] D. Bülent Danişman, Ömer Gören, Mustafa Insel, Mehmet Atlar, "An optimisation study for bow form of high speed catamarans," *Marine Technol.*, Vol. 38, No. 2, 2001.
- [30] Marcus Bole, Byung-Suk Lee, "Integrating parametric hull generation into early stage design," *Ship Technol. Research*, Vol 53, 2006, pp. 115-137.
- [31] Herbert J. Koelman, Bastiaan N. Veelo, "A technical note on the geometric representation of a ship hull form," *Comput. Aided Design*, vol. 45, 2013, pp. 1378-1381.
- [32] Qingde Li, Jie Tian, "Partial shape-preserving splines," *Comput. Aided Design*, vol. 43, 2011, pp. 394-409.
- [33] K.V. Kostas, A.I. Ginnis, C.G. Politis, P.D. Kaklis, "Ship-hull shape optimization with a T-spline based BEM-Isogeometric solver," *Comput. Methods Appl. Mech. Engrg.*, 2014.
- [34] Yusuke Tahara, Daniele Peri, Emilio Fortunato Campana, Frederick Stern, "Single and multiobjective design optimization of a fast multihull ship: numerical and experimental results," *J. Mar. Sci. Technol.*, vol. 16, 2011, pp. 412-433.
- [35] Daniele Peri and Federica Tinti, "A multistart gradient-based algorithm with surrogate model for global optimization" *Commun. in Appl. and Ind. Math.*, vol 3, no 1, 2012.
- [36] Kourosh Koushan, "Automatic hull form optimisation towards lower resistance and wash using artificial intelligence," *Int. Conf. on Fast Sea Transportation*, 2003.
- [37] Shinkyu Jeong and Hyunul Kim, "Development of an efficient hull form design exploration framework," *Math. Problems in Eng.*, vol. 2013.
- [38] Matteo Dieza, Emilio F. Campana, Frederick Stern, "Design-space dimensionality reduction in shape optimization by Karhunen-Loève expansion," *Comput. Methods Appl. Mech. Engrg.*, 2014.
- [39] Daniele Peri, Antonio Pinto, Emilio F. Campana, "Multi-objective optimisation of expensive objective functions with variable fidelity models," *Large-Scale Nonlinear Optim.*, vol. 83, 2006.

Author Index

- Abdallaa, Gaballa M., 203
Abdelhadi, Ahmed Saad, 383
Abdullin, Vildan V., 151
Abusaad, Samieh, 43
Ahmed, Hamed, 199
Al-Messabi, Naji, 399
Alexandru Codrean, 9
Alothman, Yaser, 299
Ang, Joo Hock, 253, 406
Ashari, Djoni, 49
Atkinson, Robert, 199
- Ball, Andrew D., 43, 49, 136, 145, 203, 209
Bansal, Vinit, 155
Bao, Jie, 167
Bhanot, Surekha, 155
Brethee, Khaldoon F., 136
- Cang, Shuang, 77, 311
Cao, Li, 295
Cao, Yi, 27, 341
Carvalho, Cinthya R., 235
Chalupa, Jan, 173, 317
Chang, Wenlong, 124, 389
Chatwin, Chris, 365
Chen, Leo Yi, 260
Chen, Wei-neng, 260
Chen, Yi, 285
Cheng, Kai, 2
- Dragomir, Toma-Leonida, 9
Dunnigan, Matthew, 130
- Ehinmowo, Adegboyega, 27
Elmrabit, Nebrase, 108
Ertiame, A.M.S., 347
Eze, Chinedu, 365
Eze, Elias, 90
- F., Zouhar, 323
Fyson, John, 191
- Gao, Bo, 53
Gao, Jingwei, 136
Glover, Ian, 199
Goh, Cindy, 399, 406
Gomm, J. Barry, 383
Gregor, Milan, 254
Grepl, Robert, 173, 317, 323
Grznár, Patrik, 254
Gu, Dongbing, 216, 222, 299
Gu, Fengshou, 43, 49, 136, 145, 203, 209
- Gui, Jianjun, 216
Gupta, Manoj, 279
- Hand, Ronan, 124
Hao, Guangbo, 124
Harrison, David, 191
Hasan, Mohammad S., 84
Haxha, Shyqyri, 96
Herčko, Jozef, 254
Herpe, Xavier, 130
Hu, Huosheng, 216, 222
Hu, Niaoqing, 43
Huang, Jianfeng, 285
Huo, Dehong, 279
Hussain, Mushahid, 377
- Igual, Javier Zamorano, 267
- Jaber, Adel, 199
Jasim, Wesam, 299
Jiang, Peiran, 120
Joza, Ana, 102
Judd, Martin, 199
- Kang, Jinsheng, 194
Kanwal, Kapil, 96
KhademOlama, Ehsan, 179
Khalil, Ashraf, 3
Khan, Umar, 199
Kong, Xianwen, 124, 130
- Lane, Mark, 49
Lang, Ziqiang, 33
Lawal, Sulaiman A., 329
Lazaridis, Pavlos, 199
Leithead, William Edward, 1, 167
Li, Guoxing, 145
Li, Siyan, 59
Li, Yun, 53, 260, 285, 371, 399, 406
Li, Zhong, 359
Liu, Enjie, 90
Liu, Pengcheng, 311
Liu, Ying, 353
Lou, Huanzhi, 246
Lu, Zhongyu, 273
Luo, Wuqiao, 53
Luo, Xichun, 120, 124, 291, 389
Luo, Yuanxin, 285
- M., Bastl, 323
Ma, Sicong, 59
Ma, Xiandong, 37
Mariani, Alessandro, 185
Martin, Richard, 266
- Matejasko, Michal, 323
Mather, Peter, 199
Medeiros, Fátima N. S. de, 229
Mihajlovic, Zivorad, 102
Milosavljevic, Vladimir, 102
Mir, Amir, 291
Mohanta, Harekrishna, 155
Moutinho, Luiz, 371
Mustafa, Safi, 377
- Naheed, Naila, 114
- Olama, Ehsan Khadem, 305
Opong, Kwaku, 371
- Pang, Yang, 371
Pereira, Nicolás S., 235
- Qian, Peng, 37
Qin, Shengfeng, 240
- Rajaratnam, Kumaran, 383
Rajs, Vladimir, 102
Rehab, Ibrahim, 203, 209
Righettini, Paolo, 179, 305
Rocher, Paulino, 252
Rubio, Luis, 389
- Saeed, Bakhtiar, 199
Safdar, Ghazanfar, 96
Saldivar, Alfredo Alan Flores, 260
Samuel, Raphael, 341
Shaeboubm, Abdulkarim, 43
Shah, Munam Ali, 114, 377
Shahid, Muhammad Bilal, 377
Sheng, Ning, 15, 335
Shnayder, Dmitry A., 151
Shu, Yu, 359
Siddiq, Amir, 291
Silva, Rodrigo D. C., 229
Singh, Parikshit, 155
Southee, Darren, 191
Sova, Václav, 173, 317
Stefan, Octavian, 9
Stevenson, Peter, 185
Stockley, Thomas, 185
Strada, Roberto, 179, 305
Sun, Wenlei, 194, 295
- Tan, Yuanhua, 295
Tanwilaisiri, Anan, 191
Tao, Qing, 194, 295
Teng, Xiangyu, 279
Thé, George A. P., 229, 235
Thanapalan, Kary, 185

Thiengburanathum, Pree, [77](#)
Tian, Huaiwen, [240](#)
Tian, Xiang, [203](#), [209](#)
Tian, Zhong, [53](#)
Tong, Ling, [53](#)

Upton, David, [199](#)

Valilou, Shirin, [305](#)
Valilou, Sirin, [179](#)
Vieira, Maria Fatima Queiroz, [199](#)

Walker, Ross, [130](#)
Wang, Aiping, [15](#)
Wang, Dianwei, [353](#)
Wang, Hong, [15](#), [21](#), [335](#)
Wang, Houjun, [53](#)
Wang, Jihong, [3](#)
Wang, Jing, [353](#)
Wang, Liquan, [120](#)
Wang, Mengling, [167](#)
Wang, Tie, [145](#), [203](#)
Wang, Xuan, [65](#)
Wang, Yifei, [37](#)
Wang, Zhonglai, [285](#)

Wang, Zhuo, [21](#)
Wei, Nasha, [145](#)
Williams, Jonathan, [185](#)
Wong, Wai Leong Eugene, [279](#)
Wu, Baozhong, [393](#)

Xu, Qiang, [267](#), [273](#)
Xu, Yanmeng, [191](#)
Xu, Yuandong, [145](#)
Xu, Zhenying, [393](#)
Xu, Zhijie, [353](#)

Yan, Yuehui, [393](#)
Yang, Erfu, [285](#)
Yang, Hongji, [59](#), [65](#), [71](#)
Yang, Lili, [108](#)
Yang, Longjie, [145](#)
Yang, Shuang Hua, [108](#)
Yang, Tai, [359](#), [365](#)
Yang, Wen-Xian, [33](#)
Yang, Xin, [273](#)
Yin, Xin, [15](#)
Yu, Dingli, [161](#), [347](#), [383](#)
Yu, Dingwen, [161](#), [347](#)
Yu, Feng, [161](#), [347](#)

Yu, Hong, [365](#)
Yu, Hongnian, [77](#), [84](#), [311](#)
Yu, Hui, [246](#)
Yue, Dong, [365](#)
Yue, Hong, [167](#)

Zeng, Baoqing, [53](#)
Zhan, Zhi-hui, [260](#)
Zhang, Jie, [329](#)
Zhang, Jun, [260](#)
Zhang, Long, [33](#)
Zhang, Meng, [393](#)
Zhang, Qichun, [21](#)
Zhang, Ruirong, [191](#)
Zhang, Sijing, [90](#), [114](#), [377](#)
Zhang, Ye, [246](#)
Zhang, Yong, [199](#)
Zhao, Cheng, [222](#)
Zhong, Wenbin, [389](#)
Zhou, Jianxing, [194](#)
Zhu, Lanxiang, [161](#), [347](#)
Zivanov, Milos, [102](#)
Zou, Feng, [393](#)
Zou, Lin, [71](#)